САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

На правах рукописи

Сенов Александр Алексеевич

Методы оптимизации и оценивания параметров в многомерных задачах с произвольными помехами

Специальность 01.01.09 —
«Дискретная математика и математическая кибернетика»

Диссертация на соискание учёной степени
кандидата физико-математических наук

Научный руководитель:
доктор физико-математических наук, профессор
Граничин Олег Николаевич

Санкт-Петербург — 2020

Оглавление

Стр.

Стр.

Введение

Решение многих задач адаптивного управления, распознавания образов, моделирования, обработки сигналов сводится к решению соответствующих задач нелинейной оптимизации и оценивания параметров [1—4]. Рост объема и источников данных, развитие вычислительной техники и повышение требований к качеству моделей создает потребность в разработке методов, применимых в пространствах высокой размерности. Примером могут служить задачи распознавания образов: существует прямая связь между числом параметров модели и ее обобщающей способностью, а сложная природа входных объектов, таких как изображения и звук, приводит к тому, что размерность пространства параметров может исчисляться миллионами [5—7].

Итеративные методы выпуклой оптимизации нашли широкое применение в задачах адаптивного управления, распознавания образов, моделирования, анализа данных, обработки сигналов [2]. Первые формулировки итеративных методов оптимизации можно отнести к работам И. Ньютона, а становление выпуклой оптимизации как самостоятельной дисциплины и исследование свойств сходимости методов относятся к середине XX-го века. Метод градиентного спуска и метод Ньютона–Рафсона являются, вероятно, наиболее известными методами оптимизации. В надлежащих условиях метод Ньютона-Рафсона сходится к точке минимума функции с квадратичной скоростью, достаточной для большинства практических задач. Однако необходимость расчета и обращения матрицы вторых производных делает его неприменимым в случае рассматриваемых задач высокой размерности. В свою очередь, обладающий низкой ресурсоемкостью метод градиентного спуска имеет и низкую скорость сходимости. Этот “зазор” между методами градиентного спуска и Ньютона-Рафсона активно заполняется с середины XX-го века разработкой таких методов оптимизации, как: методы сопряженных градиентов [8—13], квазиньютоновские методы [14—19], методы градиентного спуска с памятью [20—22] и др.

Последовательная подпространственная оптимизация (ППО) — еще одно направление, предложенное на исходе XX-го века в работе Р. Конна, Н. Гоулда, А. Сартэнаэра и Ф. Тоинта [23] и получившее активное развитие в работах Г. Наркисса, М. Жибулевского, Ю. Юань, Э. Шузену и соавторов [24—26]. Основная идея подхода заключается в последовательном формировании подпространства

существенно меньшей размерности чем оригинальное и последующей оптимизации целевой функции вдоль выбранного подпространства. Перевод задачи в подпространство меньшей размерности позволяет сократить использование вычислительных ресурсов, что особенно актуально в задачах высокой размерности. Методы последовательной подпространственной оптимизации успешно применяются на практике, в том числе в задачах распознавания образов [27—29] и анализа изображений [26; 30]. Одним из существенных недостатков направления являются слабо исследованные теоретические свойства: гарантии сходимости известны лишь для некоторых из методов; слабо изучены общие теоретические свойства, такие как влияние выбора подпространств и качества решения подпространственной задачи оптимизации на сходимость методов, нижние границы сходимости. *Таким образом, исследование свойств и синтез методов последовательной подпространственной оптимизации представляются актуальными для задач оптимизации в пространствах высокой размерности*.

При наличии неопределенностей, вместо решения оптимизационной задачи могут быть использованы методы оценивания. Основы теории оценивания в условиях центрированных аддитивных помех были заложены Н.Винером, А.Н.Колмогоровым, Р.Калманом и Р.Бьюси в середине XX века. Дальнейшее развитие методов оценивания в направлении почти произвольных помех было произведено В.Н.Фоминым, А.Л.Фрадковым, В.А.Якубовичем. Стоит отметить работы А.Б.Цыбакова, А.В.Гольденшлюгера, Д.Спала, Б.Т.Поляка и О.Н.Граничина по рандомизированным методам стохастической аппроксимации и оцениванию параметров линейных моделей при произвольных помехах. Однако, в случае малого числа наблюдений и наличия систематических помех, оценка, полученная путем применения методов оптимизации и точечных методов оценивания, может привести к неудовлетворительным результатам. В подобной ситуации зачастую используется альтернативный подход, основанный на оценке *доверительных множеств*, содержащих истинное значение параметра с заданной вероятностью. Особый интерес в указанных условиях представляет задача построения *точного* доверительного множества, которое содержит истинное значение параметра точно с заданной вероятностью вне зависимости от числа наблюдений. До недавнего времени были известны способы определения лишь асимптотических доверительных множеств для параметра, применимые при достаточно сильных предположениях о распределении помех [1]. Значительным недостатком этих методов является то, что они гарантируют результат лишь при

количестве измерений стремящимся к бесконечности, в то время как для небольшого количества измерений результаты могут оказаться неудовлетворительными.

Распространенным подходом к решению задач оценивания в условиях малого числа измерений и высокой роли неопределенности является подход *рандомизации*, который заключается в добавлении в алгоритм случайных, но контролируемых экспериментатором возмущений. Рандомизация успешно используются в методах оценивания параметров при почти произвольных помехах [31—36], а также в алгоритмах построения доверительного множества для параметров системы [37—39]. Так, в работе М.Кампи и Э.Вейера [40] был предложен подход исключения областей знакодоминирующих корреляций (LSCR, leave-out sign-dominant correlation regions) для поиска точных доверительных интервалов линейной системы в одномерном случае. В последствии на основе этого подхода в работе [41] теми же авторами был предложен метод Знако-Возмущенных Сумм (SPS, Sign-Perturbated Sums, далее ЗВС) для определения точного доверительного множества параметров линейной модели при центрированных и симметрично распределенных аддитивных помехах. Одним из основных ограничений методов LSCR и ЗВС является условие симметричности распределения помех относительно нуля. Это затрудняет его применение, так как во многих практических задачах помехи могут быть не только смещенными, но и иметь неслучайную природу. Для случая неизвестных почти произвольных помех, но контролируемых входов линейного обьекта в работе [42] была предложена модификация метода LSCR для одномерного случая. Существенным недостатком методов заключается его применимость лишь в одномерном случае. *Таким образом, задача определения и характеристики точного доверительного множества параметров линейной модели при почти произвольных помехах остается открытой.*

Целью диссертационной работы состоит из двух частей: (1) разработка и исследование свойств методов последовательной подпространственной оптимизации для задачи оптимизации строго выпуклой дифференцируемой функции и (2) разработка методов оценивания точных доверительных множеств линейной системы при малом числе наблюдений и слабых ограничениях на природу помех.

Для достижения поставленной цели необходимо было решить следующие **задачи**:

– исследовать общие свойства сходимости методов последовательной подпространственной оптимизации;

– разработать эффективные методы последовательной подпространственной оптимизации с ограничением на размер используемой памяти, применимые в пространствах больших размерностей;
– исследовать возможность построения доверительного множества для параметров линейной модели в условии произвольных внешних помех.

Научную новизну работы составляют следующие основные **результаты, выносимые на защиту**:

1. установлены критерии сублинейной, линейной и суперлинейной скоростей сходимости методов последовательной подпространственной оптимизации с квадратичным суррогатом для случаев квадратичной и строго выпуклой целевых функций;
2. разработан метод последовательной подпространственной оптимизации, сходящийся за конечное число итераций в квадратичном случае и обладающий линейной скоростью сходимости;
3. разработан метод определения точного доверительного множества параметров линейной модели в условии не зависимых с входами, а в остальном произвольных аддитивных помех в наблюдениях, для одномерного случая получено аналитическое выражение границ доверительного интервала и условия их состоятельности.

Теоретическая ценность и практическая значимость Подход последовательной подпространственной оптимизации представляет собой активно развивающееся в последние годы направление математического программирования, направленное на решение задач в пространствах высоких размерностей. Подход демонстрирует актуальность в практических задачах, таких как обработка изображений и распознавание образов. Развитие этого направления вносит вклад в теорию (область нелинейной оптимизации) и в практику (упомянутые задачи анализа изображений и распознавания образов, а также анализ данных, адаптивное управление, обработка сигналов). Основной теоретический вклад заключается в полученных условиях сублинейной, линейной и сверхлинейной скоростей сходимости и характеристики скорости сходимости через выбираемые подпространства и качество решения подпространственной задачи оптимизации, а также двух предложенных методах последовательной подпространственной оптимизации с установленной скоростью сходимости.

До недавнего времени отсутствовали методы получения точных доверительных множеств при независимых, а в остальном произвольных помехах в

многомерном случае. К основному теоретическому вкладу относятся: разработка метода оценивания точного доверительного множества параметра линейной модели при условии независимых с входами, а в остальном произвольных помех, а также доказательство несмещенности и сходимости границ доверительного множества к истинному значению параметра в одномерном случае. На практике это дает возможность определить точное доверительное множество параметра при отсутствии априорной информации о распределении помех.

Апробация работы. Основные результаты диссертации докладывались на семинарах кафедры математического моделирования и энергетических систем факультета прикладной математики — процессов управления Санкт-Петербургского Государственного Университета и лаборатории “Управление сложными системами” Института проблем машиноведения Российской Академии Наук, шестой традиционной всероссийской молодежной летней школе “Управление, информация и оптимизация” (г. Солнечногорск, Московская обл., 14-20 июня, 2015) и на следующих международных конференциях: American Control Conference (Портлэнд, Орегон, США, 2014), VII-ом всероссийском совещании по проблемам управления, ВСПУ-2014 (Москва, Россия, 2014), 16th IFAC Workshop “Control Applications of Optimization”, CAO-2015 (Гармиш-Партенкирхен, Германия, 2015), International Conference on Learning and Intelligent Optimization, LION-2017 (Нижний Новгород, Россия, 2017), 20th World Congress of the International Federation of Automatic Control, IFAC World Congress 2017 (Тулуза, Франция, 2017), The 3rd International Conference on Machine Learning, Optimization and Big Data, MOD 2017 (Вольтерра, Италия, 2017).

Результаты диссертации были частично использованы в работе по грантам РФФИ 13–07–00250, 17–51–53053 и проекту РНФ 16–19–00057.

По материалам работы было получено свидетельство об официальной регистрации программы для ЭВМ ОРТА-2Б.ГРС [43].

Публикации результатов. Результаты, полученные в ходе работы над диссертацией нашли отражение в 14 научных работах [44—57]. Шесть работ [46; 48; 51—54] опубликованы в изданиях, индексируемых в международных наукометрических базах Scopus/Web of Science. Работы [44; 46; 47; 51; 53] написаны в соавторстве. В работе [51] А.Д. Кнышу принадлежит постановка задачи, А.А. Боярову принадлежит разработка и обучение алгоритма классификации, а А.А Сенову — разработка алгоритмов предобработки и сегментации изображений. В работах [46; 47; 53] О.Н. Граничину принадлежит общая постановка задачи,

а А.А. Сенову — реализация описываемых методов, формулировки и доказательства теоретических результатов, программная реализация и апробация.

В **первой главе** приведены необходимые результаты из теории оценивания и оптимизации. В первом параграфе проиллюстрирована связь между задачей оценивания точного доверительного множества параметра модели и задачей выпуклой оптимизации в пространствах больших размерностей на примере задачи распознавания образов. Во втором параграфе поставлена задача оптимизации строговыпуклой функции, изложены необходимые сведения из выпуклого анализа, рассмотрены квазиньютоновские методы, методы сопряженных градиентов, методы последовательной подпространственной оптимизации. В третьем параграфе изложены методы построения доверительного множества при условии нормально распределенных помех, при условии симметрично распределенных помех, поставлена задача определения доверительного множества параметра линейной модели при почти произвольных помехах.

Во **второй главе** изложены основные результаты работы. В первом параграфе описана общая схема методов последовательной подпространственной оптимизации, обоснованы нижняя и верхние границы сходимости методов ППО, продемонстрирована зависимость скорости сходимости от выбираемых подпространств и качества решения подпространственной задачи для квадратичного и строго выпуклого случаев. Во втором параграфе приведены методы оценки квазиньютоновского направления, способ построения подпространств. В третьем параграфе на основе полученных теоретических результатов сформулированы два метода последовательной подпространственной оптимизации, доказана линейная скорость их сходимости. В четвертом параграфе приведен метод модифицированных знако–возмущенных сумм, доказано, что полученное с помощью него множество является точным доверительным множеством параметра линейной модели при условии независимых, а в остальном произвольных аддитивных помех. В пятом параграфе для одномерного случая получено аналитическое выражение границ доверительного интервала метода модифицированных знако–возмущенных сумм, доказана несмещенность и сходимость его границ к истинному значению параметра.

Третья глава посвящена иллюстрации работы предложенных методов. В первом параграфе свойства доверительного множества, полученного методом модифицированных знако–возмущенных сумм, проиллюстрированы несколькими примерами на модельных данных. Во втором параграфе приведены результаты

сравнения предложенного метода последовательной подпространственной оптимизации с аналогами как на модельных, так и на реальных данных.

Объем и структура работы. Диссертация состоит из введения, трёх глав и заключения. Полный объём диссертации составляет 85 страниц, включая 8 рисунков и 1 таблицу. Список литературы содержит 115 наименований. Рисунки и таблицы пронумерованы по главам.

Глава 1. Оптимизация в пространствах высоких размерностей и оценивание в условиях неопределённостей

В этой главе рассмотрены две задачи: задача выпуклой оптимизации в пространствах высокой размерности (Раздел 1.2) и задача оценивания доверительных множеств в условиях неопределённостей и конечного числа измерений (Раздел 1.3). Взаимосвязь между двумя этими задачами проиллюстрирована на примере проблемы распознавания образов (Раздел 1.1), в рамках которой возникает необходимость решения соответствующих задач оптимизации и оценивания доверительных множеств.

1.1 Оптимизация и оценивание в задачах распознавания образов

Рассмотрим следующую постановку задачи распознавания образов через минимизацию функционала эмпирического риска (подробно задача распознавания образов изложена во 2-ой главе книги [4]):

$$f(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} \ell\left(y_i, g(\boldsymbol{\varphi}_i, \mathbf{x})\right) \to \min_{\mathbf{x}\in\mathbb{R}^{n_x}}, \tag{1.1}$$

где N — число наблюдений, $y_i \in \mathbb{R}$ — наблюдаемые выходы модели, $\boldsymbol{\varphi}_i \in \mathbb{R}^{n_\varphi}$ — наблюдаемые входы модели, $g : \mathbb{R}^{n_\varphi} \times \mathbb{R}^{n_x} \to \mathbb{R}$ — *решающая функция*, параметризованная вектором $\mathbf{x} \in \mathbb{R}^{n_x}$, а ℓ — *функция потерь*. При этом, наблюдения формируются по модели $y_i = g(\boldsymbol{\varphi}_i, \mathbf{x}_\star) + \varepsilon_i$, где $\mathbf{x}_\star \in \mathbb{R}^{n_x}$ — искомое значение параметра, а $\varepsilon_i \in \mathbb{R}$ — аддитивные помехи. Отметим, что подобная формулировка задачи распознавания образов не является ни общей ни единственной, подробно аспекты минимизации функционала эмпирического риска изложены в работах [4; 36; 58—60]. Однако этой постановки достаточно для демонстрации следующего тезиса: задача оптимизации и задача оценивания доверительных множеств дополняют друг друга в том смысле, что являются частными случаями одной и той же практической задачи при разных условиях. В последующих подразделах этот тезис проиллюстрирован двумя примерами: случаем высокой размерности про-

странства параметров n_x и случаем малого числа измерений N при произвольных помехах.

1.1.1 Случай большого числа параметров

Пусть функция потерь ℓ и решающая функция g строго выпуклы по второму аргументу $\mathbf{x} \in \mathbb{R}^{n_x}$. Тогда функция $f : \mathbb{R}^{n_x} \to \mathbb{R}$ также сильно выпукла, и для её минимизации могут быть использованы подходы, описанные в Разделе 1.2. При определенных видах функции g, функции ℓ и помехах ε_i, точка минимума функционала среднего риска $\operatorname{argmin} f(\mathbf{x})$ в некотором смысле сходится к истинному значению параметра $\mathbf{x}_\star$ при $N \to \infty$. Примерами строго выпуклых функций потерь могут служить: квадратичная $\ell(y, g(\boldsymbol{\varphi}, \mathbf{x})) = (y - g(\boldsymbol{\varphi}, \mathbf{x}))^2$, абсолютная с ℓ_2-регуляризацией $\ell(y, g(\boldsymbol{\varphi}, \mathbf{x})) = |y - g(\boldsymbol{\varphi}, \mathbf{x})| + \lambda\|\mathbf{x} - \mathbf{x}_0\|^2$, логистическая функция потерь $\ell(y, g(\boldsymbol{\varphi}, \mathbf{x})) = -y \log g(\boldsymbol{\varphi}, \mathbf{x}) - (1 - y)(1 - \log g(\boldsymbol{\varphi}, \mathbf{x}))$. В свою очередь, в качестве решающих функций распространены линейная по $\mathbf{x}$: $g(\boldsymbol{\varphi}, \mathbf{x}) = \sum\limits_{j=1}^{n_x} \mathbf{x}_j g_j(\boldsymbol{\varphi})$, и логистическая: $g(\boldsymbol{\varphi}, \mathbf{x}) = (1 + \exp(-\sum\limits_{j=1}^{n_x} \mathbf{x}_j g_j(\boldsymbol{\varphi})))^{-1}$ функции. Систематизирование изложение функций потерь и решающих функций приведено в работах [4; 5; 60].

Ключевое свойство решающих функций — моделей машинного обучения — это обобщающая способность, характеризующая сложность взаимосвязей, которые соответствующая модель может описать (подробнее понятие обобщающей способности изложено в [61]). Один из основных способов повышения обобщающей способности, помимо выбора вида решающей функции, заключается в увеличении числа параметров — размерности пространства n_x. Таким образом, задача минимизации сильно выпуклых функций высокой размерности представляется актуальной в контексте задачи распознавания образов.

В случае квадратичной функции потерь с линейной по $\mathbf{x}$ решающей функцией уравнение (1.1) принимает следующий вид:

$$f_{MHK}(\mathbf{x}) = \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{n_x} \mathbf{x}_j g_j(\boldsymbol{\varphi}_i) \right)^2 \to \min_{\mathbf{x}}, \tag{1.2}$$

и точка минимума может быть оценена методом наименьших квадратов: $\mathbf{x}_{MHK} = \left(\boldsymbol{\Phi}^\top \boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^\top \mathbf{y}$, где $\boldsymbol{\Phi}_{i,j} = g_j(\boldsymbol{\varphi}_i)$ и $\mathbf{y} = (y_1, \ldots, y_N)^\top$. При условии отсутствия

помех и невырожденности матрицы $\mathbf{\Phi}^\top \mathbf{\Phi}$, решение задачи (1.2) существует, единственно и совпадает с истинным значением параметра: $\hat{\mathbf{x}}_{MHK} = \operatorname{argmin} f_{MHK} = \mathbf{x}_\star$. Если же помехи ε_i аддитивны, центрированы, имеют одинаковую конечную дисперсию, независимы друг с другом и с входами $\boldsymbol{\varphi}_j$, то соответствующая оценка состоятельна: $\hat{\mathbf{x}}_{MHK} \xrightarrow[N\to\infty]{P} \mathbf{x}_\star$. Таким образом, при выполнении определенных условий на природу помехи, структуру модели и достаточном количестве наблюдений минимизация функционала эмпирического риска может привести к удовлетворительному решению.

1.1.2 Случай неопределённостей и малого числа измерений

В предыдущем разделе продемонстрировано, что при большом числе измерений и определённых ограничениях на природу помех, решение задачи минимизации функционала эмпирического риска может привести к удовлетворительному решению. Однако в случае малого числа измерений и высокой неопределённости, выражающейся в неизвестной природе помех, потенциально неограниченных и носящих систематический характер, минимизация функционала эмпирического риска может приводить к сколь угодно плохим оценкам и в этом смысле малополезна. Примером источников помех могут служить: погрешности измерения выходов, неадекватность выбранной модели, а также умышленно внедренные оппонентом системы помехи [3]. В подобной ситуации альтернативой точечным оценкам служат оценки *доверительного множества*, содержащего истинное значение параметра с заданной вероятностью. Отметим, что асимптотические доверительные множества, содержащие значение параметра с заданной вероятностью лишь при количестве наблюдений стремящемся к бесконечности, малополезны при малом числе измерений. Поэтому особенно актуальна задача построения *точного* доверительного множества, содержащего истинное значение параметра точно с заданной вероятностью вне зависимости от числа наблюдений. Формально задача ставится следующим образом: для конечного набора наблюдений $\{\boldsymbol{\varphi}_i, y_i\}_{i=1}^N$ и параметра $\alpha \in [0,1]$ построить множество $\mathcal{X}_\alpha$, содержащее истинное значение параметра с заданной вероятностью α:

$$P(\mathbf{x}_\star \in \mathcal{X}_\alpha) = \alpha.$$

без значительных ограничений на тип распределения помех, за исключением независимости друг с другом и с входами $\{\boldsymbol{\varphi}_i\}_1^N$.

1.2 Оптимизация в пространствах высоких размерностей

Рассмотрим задачу безусловной оптимизации

$$f(\mathbf{x}) \to \min_{\mathbf{x}\in\mathbb{R}^n}, \tag{1.3}$$

где f принадлежит множеству $\mathcal{F}_{\mu,L}$ — μ-сильно выпуклых дважды дифференцируемых функций с липшицевым градиентом:

$$\mu\|\mathbf{x}-\mathbf{y}\|_2^2 \leqslant (\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^\top(\mathbf{x}-\mathbf{y}) \leqslant L\|\mathbf{x}-\mathbf{y}\|_2^2. \tag{1.4}$$

Понятие "высокой" размерности определим следующим образом: рассматриваемые алгоритмы решения задачи (1.3) должны принадлежать классу сложности $\mathcal{O}(n)$ по используемой памяти. Подобное ограничение, например, делает невозможным расчет и хранение матрицы Гессе, а также её приближений полного ранга.

Отметим, что для функций класса $\mathcal{F}_{\mu,L}$ первое неравенство из уравнения (1.4) может быть уточнено.

***Утверждение* 1.** *Пусть* $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^d)$, $\mathbf{x},\mathbf{y}\in\mathbb{R}^n$. *Тогда*

$$(\mathbf{x}-\mathbf{y})^\top(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})) \geqslant \frac{\mu L}{\mu+L}\|\mathbf{x}-\mathbf{y}\|^2 + \frac{1}{\mu+L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2.$$

Доказательство. Смотри доказательство Теоремы 2.1.12 в [62]. □

Известно, что решение задачи (1.3) существует и единственно $\forall\ f \in \mathcal{F}_{\mu,L}$: $\mathbf{x}_\star = \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})$. В случаях, когда точка минимума $\mathbf{x}_\star$ не может быть выражена аналитически, либо вычисление аналитического выражения трудоемко, может быть использован широкий ассортимент итеративных методов оптимизации (смотри, например, [2; 63—65]), которые строят рекуррентную последовательность $\mathbf{x}_0, \mathbf{x}_1, \ldots$ оценок точки минимума $\mathbf{x}_\star$.

В случае дважды дифференцируемой функции f для решения задачи может быть использован метод Ньютона-Рафсона, определяемый рекуррентным соотношением $\mathbf{x}_{t+1} = \mathbf{x}_t - \left[\nabla^2 f(\mathbf{x}_t)\right]^{-1}\nabla f(\mathbf{x}_t)$ (подробно метод и история его появления

разобраны в [66; 67]) и обладающий квадратичной скоростью сходимости при дополнительном условии липшицевости гессиана (понятие скорости сходимости процесса оптимизации рассматривается в разделе 1.2.1). Метод обладает существенным недостатком: необходимостью вычисления и хранения матрицы вторых производных, что делает его неприменимым в пространствах высокой размерности. Другим распространенным методом решения задачи 1.3 является метод градиентного спуска (общепринято приписывать авторство работе Коши [68], историческая справка вместе с подробным изложением приведены в работе [69]), использующий вместо обратной матрицы Гессе определяемую экспериментатором последовательность размеров шага α_t: $\mathbf{x}_{t+1} = \mathbf{x}_t + \alpha_t \nabla f(\mathbf{x}_t)$. Несмотря на простоту и вычислительную эффективность, недостатком метода является его низкая скорость сходимости в теоретическом смысле и на практике, а также чувствительность к выбору размера шага. В контексте поставленной задачи 1.3 представляют интерес методы, находящиеся в некотором смысле между методами Ньютона-Рафсона и градиетного спуска: превосходящие последний по скорости сходимости и одновременно применимые при больших значениях размерности пространства n.

1.2.1 Оценка качества алгоритмов оптимизации

Рассмотрим основные подходы к оценке качества алгоритмов оптимизации. Распространенным способом измерения скорости работы алгоритмов является оценка числа арифметических операций, необходимых для завершения алгоритма. В силу потенциальной неограниченности необходимого числа итераций, а так же того, что сложность вычисления функции и ее производных кардинально различается в зависимости от вида функции и может значительно превосходить сложность вычислений самого алгоритма, оценка числа вычислительных операций редко используется для измерения качества итеративных методов оптимизации [62; 70]. Далее перечислим наиболее распространенные методики оценивания скорости их работы.

– *Оракульная сложность* — число обращений к *оракулу* (абстракции, по запросу предоставляющей значения функции и ее производных в данной точке), необходимое для достижения заданного значения ошибки по

функции, градиенту, или аргументу. Например то, что алгоритм $\mathcal{A}$ имеет сложность $\mathcal{O}\left(N_{\mathcal{A}}(\varepsilon)\right)$ означает, что для достижения условия $f(\hat{\mathbf{x}}_\varepsilon) - f_\star < \varepsilon$ ему необходимо сделать порядка $\mathcal{O}\left(N_{\mathcal{A}}(\varepsilon)\right)$ вызовов оракула.

– *Q-сходимость* — характеристика скорости сходимости, основывающаяся на рекуррентном выражении последовательностей $f(\mathbf{x}_t) - f(\mathbf{x}_\star)$ и $\|\mathbf{x}_t - \mathbf{x}_\star\|$. Последовательность $\|\mathbf{x}_t - \mathbf{x}_\star\| \underset{t}{\to} 0$, сходится
  – *Q-сублинейно*, если $\lim \frac{\|\mathbf{x}_{t+1}-\mathbf{x}_\star\|}{\|\mathbf{x}_t-\mathbf{x}_\star\|} = 1$;
  – *Q-линейно*, если $\exists r \in (0,1) : \lim \frac{\|\mathbf{x}_{t+1}-\mathbf{x}_\star\|}{\|\mathbf{x}_t-\mathbf{x}_\star\|} \leqslant r$;
  – *Q-сверхлинейно*, если $\lim \frac{\|\mathbf{x}_{t+1}-\mathbf{x}_\star\|}{\|\mathbf{x}_t-\mathbf{x}_\star\|} = 0$;
  – *Q-квадратично*, если $\exists r > 0 : \lim \frac{\|\mathbf{x}_{t+1}-\mathbf{x}_\star\|^2}{\|\mathbf{x}_t-\mathbf{x}_\star\|} \leqslant r$.
– *R-сходимость* (от англ. “root” — корень) — более слабый аналог Q-сходимости, характеризующий общую скорость сходимости, вместо скорости сходимости на каждой итерации. Последовательность $\|\mathbf{x}_t - \mathbf{x}_\star\| \underset{t}{\to} 0$, сходится *R-линейно*, если $\exists\ r_t$: $\|\mathbf{x}_t - \mathbf{x}_\star\| \leqslant r_t$ и r_t сходится к нулю Q-линейно. Аналогичным образом определяются и другие виды R-сходимости.

Далее, если не указано обратное, под скоростью сходимости будет пониматься Q-сходимость. Так, например, при определённых условиях метод Ньютона имеет квадратичную сходимость, а метод градиентного спуска — линейную (смотри теорему 3, §4 и теорему 1, §5 1–ой главы в [65], а также теоремы 1.2.4, 1.2.5 и 2.1.15 в [62]).

1.2.2 Квазиньютоновские методы

Квазиньютоновские методы пошагово аппроксимируют матрицу Гессе и для многих задач демонстрируют более высокую чем градиентный спуск скорость сходимости [70]. Шаг квазиньютоновского алгоритма имеет вид $\mathbf{x}_{t+1} = \mathbf{x}_t - \alpha_t \mathbf{H}_t \nabla f\left(\mathbf{x}_t\right)$, где $\mathbf{H}_t$ — это аппроксимация матрицы Гессе, при этом $\mathbf{H}_{t+1}$ зачастую вычисляется путем добавления к предыдущей оценке $\mathbf{H}_t$ матрицы ранга 1 или 2. К квазиньютоновским методом относятся: метод Давидона–Флетчера–Паувелла [14; 18], метод симметричного обновления ранга 1 [19], метод Бройдена–Флетчера–Гольдфарба–Шэнно [15; 18] (BFGS) и др. Эти методы разде-

ляют недостаток метода Ньютона: требуется расчет и хранение матрицы матрицы Гессе или ее обратной в памяти. Квазиньютоновские методы с ограничением использования памяти, например, метод L-BFGS [17]), обходят эту трудность напрямую аппроксимируя произведение $\left[\nabla^2 f(\mathbf{x})\right]^{-1} \nabla f(\mathbf{x})$, не восстанавливая сам гессиан.

Большинство квазиньютоновских методов основываются на *уравнении хорд* (так же известных, как *уравнения секущих*) — распространенном инструменте поиска корней уравнений и построения методов оптимизации. Рассмотрим дифференцируемую функцию $f : \mathbb{R}^n \to \mathbb{R}^m$, уравнение хорд для нее имеет следующий вид:

$$\nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \approx f(\mathbf{y}) - f(\mathbf{x}), \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^n,$$

что следует из разложения функции f в ряд Тейлора до второго элемента в точке $\mathbf{x}$: $f(\mathbf{y}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + o(\|\mathbf{y} - \mathbf{x}\|)$.

При построении квазиньютоновских методов оптимизации уравнения хорд играют ключевую роль. Рассмотрим дважды дифференцируемую функцию $f : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, тогда уравнение хорд для градиента ∇f будет иметь вид:

$$\nabla^2 f(\mathbf{x})(\mathbf{y} - \mathbf{x}) \approx \nabla f(\mathbf{y}) - \nabla f(\mathbf{x}). \tag{1.5}$$

Заменив $\nabla^2 f(\mathbf{x})$ на неизвестную матрицу $\mathbf{B}_t$ и заменив $\mathbf{x}$, $\mathbf{y}$ на последовательные оценки алгоритма, мы получим систему линейных уравнений относительно матрицы $\mathbf{B}_t$:

$$\mathbf{B}_t(\mathbf{x}_j - \mathbf{x}_{j-1}) = \nabla f(\mathbf{x}_j) - \nabla f(\mathbf{x}_{j-1}), \quad j = 0, \ldots, t. \tag{1.6}$$

Впервые уравнения хорд были использованы для получения оценки на матрицу вторых производных в середине 1950-ых (соответствующая работа была опубликована лишь в 1991-ом году [14]). На данный момент предложено множество методов, строящих оценки матрицы Гессе, исходя из уравнений хорд: метод Давидона–Флетчера–Паувелла [14; 18], SR-1 [19], метод Бройдена–Флетчера–Гольдфарба–Шэнно [15; 18] (BFGS), метод BFGS с усеченной историей [17] и другие. Отметим, что уравнения хорд используются не только в квазиньютоновских методах оптимизации (смотри, например, метод Барзилая-Борвейна [71; 72]). Кроме того, известны модификации уравнений хорд, приводящие к более качественным оценкам матрицы Гессе (смотри, например [73; 74]).

Заметим, что погрешность в уравнении (1.5) допускает точную оценку при Липшицевости градиента.

***Утверждение* 2.** *Пусть f — дважды дифференцируемая функция с L-липшицевым Гессианом. Тогда $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ верно:*

$$\|\nabla^2 f(\mathbf{x})(\mathbf{x}-\mathbf{y}) - (\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))\| \leqslant \frac{L}{2}\|\mathbf{x}-\mathbf{y}\|^2. \tag{1.7}$$

Доказательство. Следует из формулы конечных приращений и L-липшицевости матрицы Гессе. Смотри, например, Главу 3, Теорему 3.5, в [70]. □

Заметим, что система уравнений (1.6) может быть записана в матричном виде и дополнена требованием симметричности матрицы $\mathbf{B}_t$. Один из способов решения линейных матричных уравнений при условии симметричности искомой матрицы предложен в работе [75].

***Теорема* 1** (Don, 1987)**.** *Рассмотрим известные матрицы $\mathbf{A} \in \mathbb{R}^{n\times m}$, $\mathbf{B} \in \mathbb{R}^{n\times m}$ и неизвестную матрицу $\mathbf{X} \in \mathbb{R}^{m\times m}$. Тогда система линейных уравнений относительно симметричной матрицы*

$$\begin{cases} \mathbf{A}\mathbf{X} = \mathbf{B} \\ \mathbf{X} = \mathbf{X}^\top \end{cases} \tag{1.8}$$

имеет решение относительно $\mathbf{X}$ тогда и только когда $\exists \mathbf{A}^\sim \; : \; \mathbf{A}\mathbf{A}^\sim\mathbf{B} = \mathbf{B}$ и $\mathbf{A}\mathbf{B}^\top = \mathbf{B}\mathbf{A}^\top$. Все решения системы (1.8) описываются формулой:

$$\begin{aligned} \mathbf{X}_* = {} & \mathbf{A}^\sim\mathbf{B} + (\mathbf{I} - \mathbf{A}^\sim\mathbf{A})\left(\mathbf{A}^\sim\mathbf{B}\right)^\top \\ & + (\mathbf{I} - \mathbf{A}\mathbf{A}^\sim)\,\Theta\,(\mathbf{I} - \mathbf{A}\mathbf{A}^\sim)\,, \end{aligned} \tag{1.9}$$

где $\Theta \in \mathbb{R}^{n\times n}$ — случайная матрица. Более того, решение с минимальной нормой достигается при $\Theta = 0$.

Доказательство. Смотри доказательство Теоремы 2 в [75]. □

1.2.3 Метод сопряженных градиентов

Метод сопряженных градиентов, предложенный для решения систем линейных уравнений в работе [8] и распространенный на решение квадратичных задач

оптимизации работой [12], использует в качестве направления движения линейную комбинацию предыдущего направления и текущего значения градиента:

$$\begin{aligned} \mathbf{d}_t &= -\nabla f\left(\mathbf{x}_t\right) + \beta_t \mathbf{d}_{t-1}, \\ \mathbf{x}_{t+1} &= \mathbf{x}_t + \alpha_t \mathbf{d}_t, \end{aligned} \tag{1.10}$$

где размер шага α_t рассчитывается с помощью линейного поиска вдоль направления $\mathbf{d}_t$, а $\mathbf{d}_{-1}$ полагается равным нулю. Одной из наиболее распространенных формул расчета коэффициента β_t является формула Полака–Рибьере–Поляка, предложенная независимо в работах [10] и [11]:

$$\beta_t^{PR} = \frac{\nabla f\left(\mathbf{x}_t\right)^\top \left(\nabla f\left(\mathbf{x}_t\right) - \nabla f\left(\mathbf{x}_{t-1}\right)\right)}{\nabla f\left(\mathbf{x}_{t-1}\right)^\top \nabla f\left(\mathbf{x}_{t-1}\right)}. \tag{1.11}$$

Отметим, что формула (1.11) — не единственный вариант вычисления коэффициентов β_t. Изначально для расчета использовалась формула Флетчера–Ривза [9], и в последствии было предложено множество альтернативных вариантов расчета коэффициента (подробный список приведен в работе [76], разделы 3 и 4). Несмотря на обилие альтернатив, формула Полака–Рибьере–Поляка является стандартным способом расчета коэффициента β_t (смотри раздел 5.2 в [70]). Далее для расчета коэффициента β_t будет использована формула (1.11), если не указано обратное.

Приведем несколько важных свойств метода сопряженных градиентов

***Утверждение* 3.** *Пусть $f \in \mathcal{F}_{\mu,L} : \mathbb{R}^n \to \mathbb{R}$ — квадратичная функция с матрицей Гессе $\nabla^2 f \equiv \mathbf{A}$. Тогда, метод сопряженных градиентов сходится не более чем за n шагов с линейной скоростью:*

$$(\mathbf{x}_{t+1} - \mathbf{x}_\star)\mathbf{A}(\mathbf{x}_{t+1} - \mathbf{x}_\star) \leqslant \left(\frac{\frac{L}{\mu} - 1}{\frac{L}{\mu} + 1}\right)(\mathbf{x}_t - \mathbf{x}_\star)\mathbf{A}(\mathbf{x}_t - \mathbf{x}_\star),$$
$$\|\mathbf{x}_t - \mathbf{x}_\star\| = 0, \quad t \geqslant n.$$

При этом,

- *$\mathbf{x}_{t+1}$ является точкой минимума функции f на множестве $\mathbf{x}_0 + \operatorname{span}\{\nabla f\left(\mathbf{x}_0\right), \ldots, \nabla f\left(\mathbf{x}_t\right)\}$;*
- *направления $\{\mathbf{d}_j\}$ являются сопряжёнными относительно матрицы $\mathbf{A}$: $\mathbf{d}_i \mathbf{A} \mathbf{d}_j = 0$, $\forall\, 0 \leqslant i < j < n$.*

Доказательство. Смотри доказательство Теорем 5.4 и 5.5 в [70]. □

Последнее свойство сопряженности направлений $\{\mathbf{d}_j\}$ является ключевым, что и отражено в названии метода. Если функция f не может быть описана квадратичным полиномом, это свойство нарушается, а вместе с ним и остальные составляющие Утверждения 3. Аналогичная проблема возникает и в случае неизбежных на практике ошибок вычисления: в силу конструкции метода ошибки имеют свойство накапливаться от итерации к итерации, нарушая тем самым свойство сопряженности.

Распространенным способом борьбы с нарушением свойства сопряженности является *техника рестартов*: перезапуск алгоритма путём приравнивания очередного направления к направлению градиента каждые n итераций [9]. Известно, что для достаточно широкого класса функций использование рестартов приводит к n-квадратичной сходимости (смотри Теорему 1 в [13]), что подтверждается практическими результатами. Данный подход обладает несколькими недостатками: рестарт в направлении $-\nabla f(\mathbf{x}_t)$ не учитывает накопившуюся информацию о кривизне функции, он приводит к меньшему *немедленному* уменьшению функции и, что наиболее важно в контексте поставленной задачи, техника рестартов неприменима в пространствах больших размерностей, так как размерность пространства n превосходит желаемое количество итераций.

Метод рестартов Била-Пауэлла [77] — альтернативный подход, использующий модифицированную формулу направления $\mathbf{d}_t$:

$$\mathbf{d}_t = -\nabla f(\mathbf{x}_t) + \beta_t \mathbf{d}_{t-1} + \gamma_t \mathbf{d}_k. \tag{1.12}$$

Третий член $\gamma_t \mathbf{d}_k$ здесь играет роль рестарта в том смысле, что свойство сопряженности сохраняется лишь для направлений после шага k: $\{\mathbf{d}_j\}_{k+1 \leqslant j \leqslant t}$. Рестарт — обновление индекса k до $t-1$ — выполняется при выполнении одного из неравенств:

$$\begin{aligned} |\nabla f(\mathbf{x}_t)^\top \nabla f(\mathbf{x}_{t-1})| &\geqslant c_1 \|\nabla f(\mathbf{x}_t)\|^2, \\ c_2 \|\nabla f(\mathbf{x}_t)\|^2 \leqslant -\mathbf{d}_t^\top \nabla f(\mathbf{x}_t) &\leqslant c_3 \|\nabla f(\mathbf{x}_t)\|^2. \end{aligned} \tag{1.13}$$

Первое неравенство означает нарушение свойства ортогональности последующих градиентов, а второе — отклонение направления $\mathbf{d}_t$ от направления наискорейшего убывания функции f. В качестве коэффициентов c_1, c_2, c_3 обычно берутся значения 0.2, 0.8 и 1.2 соответственно. Отметим, что для метода Била-Пауэлла не гарантируется сходимость, но он демонстрирует хорошие результаты на практике, в том числе в многомерных задачах [78].

1.2.4 Последовательная подпространственная оптимизация

В работе [23] был предложен подход *последовательной подпространственной оптимизации* (iterated–subspace minimization, далее — ППО). Схожий подход был независимо предложен в работе [24]. Идея подхода заключается в последовательном применении двух операций:

1. построение подпространства $\mathcal{D}_t \subset \mathbb{R}^n$ малой размерности: $|\mathcal{D}_t| = m_t << b$;
2. оптимизация целевой функции вдоль выбранного подпространства: $\mathbf{x}_{t+1} = \underset{\mathbf{d}\in\mathcal{D}_t}{\operatorname{argmin}} f(\mathbf{x}_t + \mathbf{d})$.

В работе [23] авторы используют сопряжённые направления как образующие подпространств $\mathcal{D}_t$ и усечённый метод Ньютона для минимизации вдоль подпространства — задачи *подпространственной оптимизации*. За счет использования направления антиградиента в качестве одного из образующих гарантируется сходимость для класса дважды-дифференцируемых ограниченных снизу функций. В работе [24] в качестве образующих подпространства $\mathcal{D}_t$ используются: значения градиентов $\nabla f(\mathbf{x}_i)$, предшествующие направления $\mathbf{x}_i - \mathbf{x}_{i-1}$, а так же так называемые направления Немировского [79]. Задачу подпространственной оптимизации предлагается решать методом Ньютона. Для класса гладких выпуклых функций с липшицевым градиентом гарантируется сублинейная скорость сходимости ($\mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}}\right)$ в терминологии оракульной сложности).

В работах [23; 24] подход последовательной подпространственной оптимизации был впервые сформулирован в явном виде. Однако схожие идеи использовались в различных работах по оптимизации и ранее: в случае квадратичной целевой функции метод сопряженных градиентов неявным образом находит минимум функции вдоль соответствующих подпространств Крылова [80], в [81] был предложен метод оптимизации в постепенно наращиваемом “существенном” подпространстве, в работе [79] был предложен метод последовательной оптимизации вдоль подпространств, образованных текущим значением градиента и направлениями Немировского, в работах [20; 21] рассматриваются модификации метода градиентного спуска, дополнительно минимизирующие целевую функцию вдоль предшествующих направлений. Многие методы оптимизации могут быть интерпретированы в контексте подхода последовательной подпространственной оптимизации, если минимизацию целевой функции во 2-ом шаге

схемы ППО (решение подпространственной задачи минимизации), заменить на минимизацию суррогата: градиентный спуск может быть расмотрен как метод ППО с подпространством образованным текущим значением градиента $\mathcal{D}_t = \operatorname{span}\{\nabla f(\mathbf{x}_t)\}$ и суррогатной моделью $q_t(z) = z + \frac{1}{2\lambda_t} z^2$; метод сопряженных градиентов в случае квадратичной целевой функции соответствует методу ППО с подпространствами вида $\mathcal{D}_t = \operatorname{span}\{\mathbf{d}_t, \nabla f(\mathbf{x}_t)\}$ и точным решением задачи подпространственной оптимизации; многие квазиньютоновские методы, такие как SR1 [19], PSB [82] и методы из семейства Бройдена [70] используют подпространство $\mathcal{D}_t = \operatorname{span}\{\nabla f(\mathbf{x}_0), \ldots, \nabla f(\mathbf{x}_t)\}$ (смотри Теорему 2.1 в [83]). Тем не менее, именно в работах [23; 24] последовательная подпространственная оптимизация впервые была сформулирована явно и в общем виде.

Отметим работы Yuxian Yuan и соавторов, посвященные *подпространственным методам оптимизации*. Так, в работе [84] авторы предложили метод последовательной минимизации функции вдоль текущего направления градиента и предыдущего шага, обобщающий метод сопряженных градиентов, и обосновали его сходимость при достаточно общих условиях и использовании условий Вольфе [85; 86] для выбора размера шага. В работе [87] формулируется общая модель подпространственных методов оптимизации. В качестве образующих подпространства $\mathcal{D}_t$ берутся текущие значение градиента и усеченная история предшествующих направлений $\{\nabla f(\mathbf{x}_t), \mathbf{s}_{t-1}, \ldots, \mathbf{s}_{t-m}\}$. Для минимизации вдоль подпространства $\mathcal{D}_t$ используется квадратичный суррогат $q_t(\mathbf{d}) = \nabla f(\mathbf{x}_t)^\top \mathbf{d} + \frac{1}{2}\mathbf{d}^\top \mathbf{B}_t \mathbf{d}$, где аппроксимация Гессиана $\mathbf{B}_t$ оценивается исходя из уравнений хорд. Также рассматривается вариация для случая задачи с ограничениями. В [83] на основе подпространственного похода предлагается квазиньютоновский метод с использованием доверительных множеств. В работе [88] предложенная модель распространяется на задачу решения системы нелинейных уравнений, а в работе [25] упомянутые результаты излагаются в обобщённом и систематизированном виде.

Методы последовательной подпространственной оптимизации активно применяются на практике. В работе [26] в контексте проблемы восстановления исходного изображения предлагается метод ППО, где задача подпространственной оптимизации решается за счет минимизации квадратичной мажоранты (суррогата) целевой функции, специфичного для задачи восстановления исходного изображения. В [27] метод ППО применяется для решения частной задачи распознавания образов — поиска параметров машины опорных векторов [4].

В работе [30] рассматриваются аспекты применения методов оптимизации к задачам обработки изображений, экспериментально демонстрируется превосходство конкретного метода последовательной подпространственной оптимизации над аналогами в задачах удаления размытия и восстановления изображения из проекций.

Подход последовательной подпространственной оптимизации активно используется при конструировании других методов. Так, в работе [28] подход последовательной подпространственной оптимизации адаптирован для поиска аппроксимации постериорного распределения, минимизирующего дивергенцию Кульбака-Лейблера. В работе [29] в контексте решения задачи стохастической оптимизации, авторы предлагают ускорять выбранный метод стохастической оптимизации регулярными итерациями метода ППО с подпространством, сформированным из предыдущих значений градиентов, направлений Немировского и векторов между текущей оценкой и т.н. *якорными точками* — зафиксированными оценками в прошлом. В работе демонстрируется эффективность предложенной модификации на таких задачах распознавания образов (минимизации функционала эмпирического риска, смотри главу 2, §7, в [4]), как задачи классификации изображений MNIST [89] и CIFAR-10 [90] методами глубокого обучения.

Подход проекции–аппроксимации–восстановления в оптимизации является частным случаем последовательной подпространственной оптимизации. Основное отличие заключается в способе построения суррогата q_t: в упомянутых выше работах [22; 24; 25; 71; 83; 87] суррогат конструируется аналитически, в то время как в подходе проекции–аппроксимации–восстановления коэффициенты суррогата аппроксимируются исходя из решения регрессионной задачи. Рассматривается параметризованный суррогат $q_t(\mathbf{z}) = q(\mathbf{z}|\theta_t)$ и параметр θ оценивается исходя из задачи минимизации

$$\sum_{i=t}^{t-m+1} \left(f(\mathbf{x}_i) - q(\mathbf{D}_t\mathbf{x}_i|\theta)\right)^2 \to \min_{\theta}.$$

В работах [52; 53] подход проекции–аппроксимации–восстановления используется для ускорения градиентного спуска: после каждых m итераций градиентного спуска на основе последних m значений $\{\mathbf{x}_i\}_{t-m+1}^{t}$ и $\{f(\mathbf{x}_i)\}_{t-m+1}^{t}$ строится суррогат $q(\cdot|\theta_t)$, и точка его минимума используется как очередная оценка минимума $\mathbf{x}_{t+1}$. В [54] приводится четыре алгоритма, основанные на последовательном применении следующих операций:

1. построение матрицы проекции $\mathbf{D}_t \in \mathbb{R}^{n\times m}$;
2. аппроксимация функции f суррогатом $q(\cdot|\theta_t)$ на в точках $\{\mathbf{x}_i, f(\mathbf{x}_i)\}_{t-m+1}^{t}$;
3. восстановление минимума функции f (очередной оценки) из полученного суррогата q_t.

Для построения матрицы $\mathbf{D}_t$ в работе используются как предыдущие значения градиента $\{\nabla f(\mathbf{x}_i)\}_{i=0}^{t}$, так и вектора случайной природы. Очередная оценка $\mathbf{x}_{t+1}$ получается напрямую из точки минимума $\operatorname{argmin} q_t$, где в качестве суррогата q_t используется квадратичная функция.

1.3 Оценивание доверительных множеств в условиях неопределённостей и конечного числа наблюдений

Рассмотрим задачу оценки параметра линейной модели входа–выхода по наблюдениям:

$$y_i = \boldsymbol{\varphi}_i^T \mathbf{x}_\star + \varepsilon_i, \quad i = 1..N, \tag{1.14}$$

где N — число измерений $y_i \in \mathbb{R}$ — наблюдаемые выходы системы, $\boldsymbol{\varphi}_i \in \mathbb{R}^n$ — наблюдаемые входы системы, $\mathbf{x}_\star \in \mathbb{R}^n$ — искомый вектор параметров, а $\varepsilon_i \in \mathbb{R}$ — неизвестные аддитивные помехи.

При большом числе наблюдений и благоприятных условиях на помехи ε_i, удовлетворительнуя оценка $\hat{\mathbf{x}}$ истинного значения параметра $\mathbf{x}_\star$ может быть получена методами регрессионного анализа. Многие из регрессионных методов, в свою очередь, формулируются в виде оптимизационной задачи. Однако в случае малого числа измерений, низкого соотношения сигнал/шум и при неизвестном распределении помех полученные оценки $\hat{\mathbf{x}}$ могут значительно отличаться от истинного значения параметра и в этом смысле ненадёжны. Альтернативный подход — построение *доверительного множества*, содержащего истинное значение параметра с заданной вероятностью. Особый интерес представляет задача построения *точного* доверительного множества, которое содержит истинное значение параметра точно с заданной вероятностью вне зависимости от числа наблюдений — в противовес асимптотическому доверительному множеству, которое содержит истинное значение параметра с заданной вероятностью лишь при

числе наблюдений стремящемся к бесконечности. Отметим, что источником случайности могут служить как помехи, так и входы системы (1.14).

Рассмотрим основные виды помех и соответствующие подходы к определению доверительных множеств параметра системы. Для случая независимо и одинаково нормально распределенных помех известны выражения для точных доверительных множеств параметра $\mathbf{x}$ [91; 92]. В случае большого числа измерений и распределения помех отличном от нормального аналогичные процедуры приведут к *асимптотическому* доверительному множеству. Аналогичными недостатками обладают и методы построения доверительных множеств, основанные на бутстрэппинге [93—95]. В случае равномерной ограниченности помех: $\exists C > 0$: $|\varepsilon| \leqslant C$, — возможно построение множества *точно* содержащего истинное значение параметра (т.н. *set membership approach*) [96; 97], которое можно рассмотреть как точное доверительное множество, содержащее истинное значение параметра с вероятностью 1. Для случая независимых одинаково распределенных относительно нуля помех существует несколько методов построения точных доверительных множеств: метод исключения областей знакодоминирующих корреляций (LSCR, leave-out sign-dominant correlation regions [40; 98]) для одномерного случая и его обобщение на многомерный случай — метод знаковозмущенных сумм (SPS, sign-perturbed sums [41; 99; 100]). Отдельно стоит отметить работу [42], где была предложена модификация метода LSCR для случая неизвестных помех, но контролируемых входов линейного объекта управления.

1.3.1 Случай нормально распределенных помех

В предположении нормальности, независимости и одинакового распределения помех $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ с неизвестной дисперсией σ^2 точное доверительное множеств параметра $\mathbf{x}_\star$ может быть построено с использованием распределением Фишера. Пусть $\alpha \in [0,1]$, тогда $P\left(\mathbf{x}_\star \in \mathcal{X}_\alpha^{\mathcal{N}}\right) = \alpha$, где доверительное множество $\mathcal{X}_\alpha^{\mathcal{N}}$ определяется по следующей формуле (смотри Главу 5 в [92]):

$$\mathcal{X}_\alpha^{\mathcal{N}} = \left\{ \mathbf{x} \in \mathbb{R}^N : \frac{1}{N-n}(\mathbf{x} - \hat{\mathbf{x}}_{MHK})^T \boldsymbol{\Xi}_{\mathbf{x}}^{-1} (\mathbf{x} - \hat{\mathbf{x}}_{MHK}) \leqslant F_\alpha(n, N-n) \right\}, \tag{1.15}$$

где $F_\alpha(n, N-n)$ — α-квантиль распределения Фишера с n и $N-n$ степенями свободы, $\hat{\mathbf{x}}_{MHK} = \left(\mathbf{\Phi}^\top \mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^\top \mathbf{y}$ — оценка параметра $\mathbf{x}_\star$ методом наименьших квадратов, $\mathbf{\Phi} = [\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_N]^\top$ — матрица входов системы, $\mathbf{y} = (y_1, \ldots, y_N)$ — вектор выходов системы, $\hat{\mathbf{\Xi}}_{\mathbf{x}} = \hat{\sigma}^2 \left(\Phi^\top \Phi\right)^{-1}$ — ковариационная матрица $\hat{\mathbf{x}}_{MHK}$, $\hat{\sigma}^2 = \frac{1}{N-n} \|\mathbf{y} - \mathbf{\Phi} \hat{\mathbf{x}}_{MHK}\|^2$ — несмещенная оценка σ^2.

1.3.2 Случай симметрично распределенных помех

Введем определение *симметричной* многомерной случайной величины.

***Определение* 1.** *Пусть $(\Omega, \mathcal{F}, P)$ — вероятностное пространство, тогда многомерная случайная величина $\xi : \Omega \to \mathbb{R}^n$ называется симметричной, если*

$$\forall\, A \in \mathcal{F} : \; P(\xi \in A) = P(-\xi \in A). \tag{1.16}$$

Предположим, что помехи и входы системы $\{\varepsilon_1, \ldots, \varepsilon_N, \boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_N\}$ взаимно независимы, а так же помехи симметрично распределены относительно нуля. В таком случае, для построения точного доверительного множества параметра $\mathbf{x}_\star$ может быть использован метод *знако-возмущённых сумм* (ЗВС), предложенный в работе [41]:

$$\begin{gathered}
\mathcal{X}^{\text{ЗВС}}_{M,q} = \left\{ \mathbf{x} : \left| \{ k \in \{1, \ldots, M-1\} : \|\tilde{S}_0(\mathbf{x})\| < \|\tilde{S}_k(\mathbf{x})\| \} \right| \geqslant q \right\}, \\
\tilde{S}_0(\mathbf{x}) = \sum_{i=1}^{N} \boldsymbol{\varphi}_i (y_i - \boldsymbol{\varphi}_i^T \mathbf{x}), \qquad \tilde{S}_k(\mathbf{x}) = \sum_{i=1}^{N} a_{k,i} \boldsymbol{\varphi} (y_i - \boldsymbol{\varphi}_i^T \mathbf{x}), \\
a_{k,i} = \begin{cases} 1 & \text{с вероятностью } \frac{1}{2} \\ -1 & \text{с вероятностью } \frac{1}{2} \end{cases}, \qquad k = 1, \ldots, M-1.
\end{gathered} \tag{1.17}$$

Справедлив следующий результат.

***Утверждение* 4** (Csaji, Campi, Weyer [41])**.** *В сделанных выше предположениях, множество $\mathcal{X}^{\text{ЗВС}}_{M,q}$ является точным доверительным множеством для параметра $\mathbf{x}_\star$:*

$$P\left(\mathbf{x}_\star \in \mathcal{X}^{\text{ЗВС}}_{M,q}\right) = 1 - \frac{q}{M}.$$

Доказательство. Смотри доказательство Теоремы 1 в [41]. □

В свою очередь, Утверждение 4 существенным образом базируется на следующих утверждениях.

***Утверждение* 5** (Csaji, Campi, Weyer [41]). *Пусть* ξ *— симметричная многомерная случайная величина, а* α *— случайный знак (смотри формулу* (1.17)*). Тогда случайные величины* ξ *и* $\alpha\xi$ *— независимы*

Доказательство. Смотри доказательство Леммы 1 в [41]. □

Для изложения следующего утверждения введем понятие *равномерно упорядоченных* случайных величин.

***Определение* 2.** *Случайные величины* $\{\xi_i\}_1^n$ *называются равномерно упорядоченными, если для любой перестановки* π *множества* $\{1,\ldots,n\}$ *соответствующий порядок* $\{\xi_i\}_1^n$ *равновероятен:*

$$P(\xi_{\pi(1)} < \ldots < \xi_{\pi(n)}) = \frac{1}{n!}.$$

***Утверждение* 6** (Csaji, Campi, Weyer [41]). *Пусть* $\{\xi_i\}_1^n$ *— независимые одинаково и симметрично распределённые непрерывные случайные величины. Тогда они равномерно упорядочены.*

Доказательство. Смотри доказательство Леммы 4 в [41]. □

Отметим, что для множества $\mathcal{X}_{M,q}^{3BC}$ справедливо следующее свойство: $\hat{\mathbf{x}}_{MHK} \in \mathcal{X}_{M,q}^{3BC}$.

1.3.3 Случай независимых, а в остальном произвольных помех

Рассмотрим систему (1.14) и предположим, что входы $\{\boldsymbol{\varphi}_i\}_{i=1}^N$ имеют вероятностную природу, помехи $\{\varepsilon_i\}_{i=1}^N$ удовлетворяют следующим предположениям, а в остальном произвольны:

- помехи и входы системы $\{\varepsilon_1,\ldots,\varepsilon_N,\boldsymbol{\varphi}_1,\ldots,\boldsymbol{\varphi}_N\}$ взаимно независимы;
- входы системы $\boldsymbol{\varphi}_1,\ldots,\boldsymbol{\varphi}_N$ симметрично и непрерывно распределены вокруг известного математического ожидания $m_\varphi \in \mathbb{R}^n$.

Поставим задачу построения точного доверительного множества для параметра $\mathbf{x}_\star$: для наблюдений $\{\boldsymbol{\varphi}_i, y_i\}_1^N$, удовлетворяющих поставленным выше условиям, и заданного $\alpha \in [0,1]$ построить множество $\mathcal{X}_\alpha \in \mathbb{R}^n$ такое, что:

$$P(\mathbf{x}_\star \in \mathcal{X}_\alpha) = \alpha. \tag{1.18}$$

Отметим, что некоторые из изложенных в диссертации оригинальных результатов получили развитие в работах других авторов. Так, изложенный в Разделе 2.4 метод модифицированных знако–возмущенных сумм получения точного доверительного множества параметра линейной модели при почти произвольных помехах (1.18) был обобщен в диссертационной работе Волковой М. В. [101] на нелинейный случай и применен к задаче определения доверительного интервала инкубационного времени разрушения материалов.

Глава 2. Методы последовательной подпространственной оптимизации и модифицированных знако–возмущенных сумм

2.1 Свойства методов ППО

2.1.1 Общая схема методов ППО

Методы последовательной подпространственной оптимизации сводятся к последовательному применению двух операций:

1. Формирование подпространства $\mathcal{D}_t \subset \mathbb{R}^n$, размерность которого значительно меньше размерности исходного пространства $|\mathcal{D}_t| = m_t \ll n$.
2. Аппроксимация минимума целевой функции вдоль этого подпространства относительно текущего значения $\mathbf{x}_t$:

$$\mathbf{x}_{t+1} \approx \underset{\mathbf{x} \in \{\mathbf{x}_t + \mathbf{d} \,:\, \mathbf{d} \in \mathcal{D}_t\}}{\operatorname{argmin}} f(\mathbf{x}). \tag{2.1}$$

При этом, как подпространство $\mathcal{D}_t$, так и метод поиска минимума f вдоль подпространства могут варьироваться. Распространены такие образующие пространства $\mathcal{D}_t$, как: текущее и предыдущие значения градиента $\nabla f(\mathbf{x}_k)$, изменения градиента $\mathbf{y}_k = \nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}_{k-1})$ и аргумента $\mathbf{s}_k = \mathbf{x}_k - \mathbf{x}_{k-1}$, а так же их комбинации. В качестве метода решения (2.1) зачастую используется суррогатный подход, при котором вместо функции f минимизируется более простая функция-суррогат q_t, в некотором смысле близкая функции f в окрестности $\mathbf{x}_t$. Одним из наиболее распространенных видов суррогата является квадратичная функция $q_t(\mathbf{d}) = \nabla f(\mathbf{x}_t)^\top \mathbf{d} + \mathbf{d}^\top \mathbf{B}_t \mathbf{d}$, где $\mathbf{B}_t$ — аппроксимации матрицы Гессе. Стоит отметить, что размер шага вдоль найденного направления $\mathbf{d}_t$ может вычисляться линейным поиском.

Приведем общую схему методов последовательной подпространственной оптимизации.

0. Задается целевая функция f, начальная точка $\mathbf{x}_0$, $t \leftarrow 0$.
1. Строится новое подпространство $\mathcal{D}_t$.
2. Строится аппроксимация функции f в окрестности точки $\mathbf{x}_t$ по подпространству $\mathcal{D}_t$: $q_t(\mathbf{d}) \sim f(\mathbf{x}_t + \mathbf{d})$, $\mathbf{d} \in \mathcal{D}_t$.

3. Решается задача подпространственной минимизации:

$$\mathbf{d}_t = \underset{\mathbf{d}\in\mathcal{D}_t}{\operatorname{argmin}}\, q_t(\mathbf{d}). \tag{2.2}$$

4. Вдоль полученного направления выполняется линейный поиск: $\alpha_t = \underset{\alpha}{\operatorname{argmin}}\, f(\mathbf{x}_t + \alpha\mathbf{d}_t)$.
5. Рассчитывается очередная оценка: $\mathbf{x}_{t+1} = \mathbf{x}_t + \alpha_t\mathbf{d}_t$.
6. $t \leftarrow t + 1$. Переход к шагу 1.

Данная схема совпадает с представленной в работе [87] (смотри Алгоритм 3.1), за исключением наличия ограничения на норму размера шага, монотонности убывания целевой функции, а так же отсутствия шага линейного поиска. Отметим, что линейный поиск на шаге (4) может выполняться различными способами: исходя из условий Вольфе [85], численной минимизации или на основе детерминированного правила (например, $\alpha_t \equiv 1$). Исходя из приведенной схемы, определяющим для методов ППО является выбор:

a. способа построения подпространств $\{\mathcal{D}_t\}_{t\geqslant 0}$,
b. способа построения локального аппроксиматора q_t.

Стоит отметить, что зачастую аппроксиматор q_t и подпространство $\mathcal{D}_t$ не задаются явно и шаг $\mathbf{d}_t$ выписывается из соображений отличных от минимизации суррогата вдоль некоторого подпространства.

На практике удобно оперировать не подпространствами $\mathcal{D}_t$ а наборами образующих их векторов. Обозначим $\mathbf{D}_t = \left[\mathbf{d}_1^{(t)}, \ldots, \mathbf{d}_{m_t}^{(t)}\right] \in \mathbb{R}^{n\times m_t}$ — матрица, столбцы которой образуют множество $\mathcal{D}_t$. Тогда, схема ППО переписывается следующим образом:

0. Задается целевая функция f, начальная точка $\mathbf{x}_0$ и матрица $\mathbf{D}_0$; $t \leftarrow 0$.
1. Строится очередная матрица $\mathbf{D}_t$, задающая подпространство для оптимизации $\mathcal{D}_t$.
2. Строится аппроксимация функции f по подпространству $\mathcal{D}_t$ в окрестности точки $\mathbf{x}_t$: $q_t(\mathbf{z}) \sim f(\mathbf{x}_t + \mathbf{D}_t\mathbf{z})$, $\mathbf{z} \in \mathbb{R}^{m_t}$.
3. Решается задача подпространственной минимизации:

$$\mathbf{z}_t = \underset{\mathbf{z}\in\mathbb{R}^{m_t}}{\operatorname{argmin}}\, q_t(\mathbf{z}). \tag{2.3}$$

4. Вдоль полученного направления выполняется линейный поиск: $\alpha_t = \underset{\alpha}{\operatorname{argmin}}\, f(\mathbf{x}_t + \alpha\mathbf{d}_t)$.
5. Вычисляется очередная оценка: $\mathbf{x}_{t+1} = \mathbf{x}_t + \alpha_t\mathbf{d}_t = \mathbf{x}_t + \alpha_t\mathbf{D}_t\mathbf{z}_t$.

6. $t \leftarrow t + 1$. Переход к шагу 1.

Аналогичная схема рассматривается в работе [23]. В более строгом виде эта схема представлена в Алгоритме 1. Стоит отметить, что функции subspaceUpdate, и subspaceSearch могут хранить состояние: предыдущие значения точек, градиентов, их линейные комбинации и пр.

Algorithm 1 SSO_scheme($\mathbf{x}_0$, subspaceUpdate, subspaceSearch, lineSearch)

$\mathbf{D}_t \leftarrow$ subspaceUpdate($\mathbf{D}_{t-1}$, $\nabla f(\mathbf{x}_t)$)
$\mathbf{z}_t \leftarrow$ subspaceSearch(f, $\mathbf{D}_t$, $\mathbf{x}_t$)
$\mathbf{d}_t \leftarrow \mathbf{D}_t\mathbf{z}_t$
$\alpha_t \leftarrow$ lineSearch(f, $\mathbf{x}_t$, $\mathbf{d}_t$)
$\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t + \alpha_t\mathbf{d}_t$
$t \leftarrow t + 1$; go to 1

Заметим, что две приведённые схемы методов ППО (схемы с неявной и с явной проекцией) эквивалентны.

***Замечание* 1.** *Допустим* $\mathcal{D}_t = \operatorname{span}\left\{\mathbf{d}_t^{(1)}, \ldots, \mathbf{d}_t^{(m)}\right\}$, *а* $\mathbf{D}_t = \left[\mathbf{d}_t^{(1)}, \ldots, \mathbf{d}_t^{(m)}\right]$ *— матрица, чьи столбцы образуют* $\mathcal{D}_t$. *Тогда задача минимизации суррогата* $\operatorname{argmin}_{\mathbf{d}\in\mathcal{D}_t} q_t(\mathbf{d})$ *эквивалентна* $\operatorname{argmin}_{\mathbf{z}\in\mathbb{R}^m} q_t(\mathbf{D}_t\mathbf{z})$.

Как и в схеме с неявной проекцией, на шаге (4) могут использоваться различные процедуры линейного поиска. Стоит отметить, что использование матрицы $\mathbf{D}_t$ вместо подпространства $\mathcal{D}_t$ не только упрощает практическую реализацию алгоритмов, но позволяет получить некоторые свойства методов ППО и также может уменьшить вычислительную сложную алгоритма. Действительно, задача подпространственной оптимизации в явной постановке (2.3) интуитивно проще задачи в неявной постановке (2.2) в силу меньшей размерности аргумента и постановки задачи: вместо задачи n-мерной оптимизации с ограничениями ставится задача m-мерной оптимизации без ограничений. Формально эта разница сформулирована в следующем замечании.

***Замечание* 2.** *Пусть* q_t *— суррогат квадратичного вида. Тогда при постановке* (2.2) *для его хранения требуется порядка* $\mathcal{O}\left(n^2\right)$ *памяти, а для вычисления — порядка* $\mathcal{O}\left(n^3\right)$ *числа операций, в то время как при постановке* (2.3) *требуется лишь порядка* $\mathcal{O}\left(m^2\right)$ *памяти и порядка* $\mathcal{O}\left(m^3\right)$ *числа операций соответственно в худшем случае:* $q_t(\mathbf{D}_t\mathbf{z}) = \nabla f(\mathbf{x}_t)^\top \mathbf{D}_t\mathbf{z} + \mathbf{z}\mathbf{D}_t^\top\mathbf{B}_t\mathbf{D}_t\mathbf{z} = \mathbf{r}_t^\top\mathbf{z} + \mathbf{z}^\top\mathbf{Q}_t\mathbf{z}$, $\mathbf{r}_t \in \mathbb{R}^m$, $\mathbf{Q}_t \in \mathbb{R}^{m\times m}$.

Отметим свойство, общее для всех методов ППО: точность любого метода последовательной подпространственной оптимизации ограничена снизу качеством выбора подпространств, вне зависимости от способа решения подпространственной задачи, размера шага и пр.

***Замечание* 3.** *Рассмотрим процесс последовательной подпространственной оптимизации* $\mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{d}_t$, *где* $\mathbf{d}_t \in \mathcal{D}_t$. *Тогда*

$$\|\mathbf{x}_t - \mathbf{x}_\star\| \geqslant \|\mathbf{x}_0 - \mathbf{x}_\star\| - \|\mathcal{P}_{\cup_{0\leqslant j\leqslant t}\mathcal{D}_j}(\mathbf{x}_0 - \mathbf{x}_\star)\|,$$

где $\mathcal{P}_{\cup_{0\leqslant j\leqslant t}\mathcal{D}_j}(\mathbf{x})$ — *проекция вектора* $\mathbf{x}$ *на множество* $\cup_{0\leqslant j\leqslant t}\mathcal{D}_j$. *Более того, если столбцы матрицы* $\mathbf{D}_j$ *образуют ортонормированный базис пространства* $\mathcal{D}_j$ $\forall$ j, *то*

$$\|\mathbf{x}_t - \mathbf{x}_\star\| \geqslant \|\prod_{j=0}^{t}\left(\mathbf{I} - \mathbf{D}_j\mathbf{D}_j^\top\right)(\mathbf{x}_0 - \mathbf{x}_\star)\|.$$

Доказательство. Достаточно заметить, что $\mathbf{x}_t - \mathbf{x}_0 \in \operatorname{span}\{\mathbf{d}_0, \ldots, \mathbf{d}_t\}$.

□

Таким образом, Замечание 3 предлагает нижнюю границу нормы невязки процесса последовательной подпространственной оптимизации.

2.1.2 Квадратичный случай

В этом разделе свойства методов последовательной подпространственной оптимизации иллюстрируются на примере строговыпуклой квадратичной целевой функции. Рассмотрим функцию вида

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top\mathbf{A}\mathbf{x} + \mathbf{b}^\top\mathbf{x} + c, \tag{2.4}$$

где $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{A} \succ 0$, $\mathbf{b} \in \mathbb{R}^n$, $c \in \mathbb{R}$, коэффициенты $\mathbf{A}$, $\mathbf{b}$, c неизвестны. Обозначим $\mathbf{x}_\star = \mathbf{A}^{-1}\mathbf{b} = \operatorname{argmin} f$, зафиксируем матрицу $\mathbf{D}_t \in \mathbb{R}^{n\times m}$, $1 \leqslant m \leqslant n$ и $\mathbf{x}_t \in \mathbb{R}^n$. Рассмотрим задачу поиска минимума f в подпространстве $\mathcal{D}_t = \{\mathbf{x}_t + \mathbf{D}_t\mathbf{z} \ : \ \mathbf{z} \in \mathbb{R}^m\}$:

$$f\left(\mathbf{x}_t + \mathbf{D}_t\mathbf{z}\right) \to \min_{\mathbf{z}}. \tag{2.5}$$

Рассмотрим разложение $f(\mathbf{x}_t + \mathbf{D}_t\mathbf{z})$ по Тейлору в точке $\mathbf{x}_t$:

$$f(\mathbf{x}_t + \mathbf{D}_t\mathbf{z}) = f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top \mathbf{D}_t\mathbf{z} + \frac{1}{2}\mathbf{z}^\top \mathbf{D}_t^\top \nabla^2 f(\mathbf{x}_t)\mathbf{D}_t\mathbf{z} + o(\|\mathbf{D}_t\mathbf{z}\|_2^2). \tag{2.6}$$

Предположим, что суррогат q_t в точности совпадает с разложением Тейлора до второй степени включительно:

$$q_t(\mathbf{z}) = f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top \mathbf{D}_t\mathbf{z} + \frac{1}{2}\mathbf{z}^\top \mathbf{D}_t^\top \nabla^2 f(\mathbf{x}_t)\mathbf{D}_t\mathbf{z}. \tag{2.7}$$

Тогда точка минимума суррогата имеет вид $\mathbf{z}_t = \operatorname{argmin}_{\mathbf{z}} q_t = -\left(\mathbf{D}_t^\top \mathbf{A}\mathbf{D}_t\right)^{-1} \mathbf{D}_t^\top \nabla f\left(\mathbf{x}_t\right.$ в предположении, что матрица $\mathbf{D}_t$ невырожденна. Таким образом,

$$\mathbf{H}_t := \mathbf{D}_t \left(\mathbf{D}_t^\top \mathbf{A}\mathbf{D}_t\right)^{-1} \mathbf{D}_t^\top, \tag{2.8}$$

— ни что иное, как аппроксимация обратной к матрице Гессе. Несмотря на то, что вопрос близости полученной точки минимума суррогата (2.7) и решения задачи (2.5) остается открытым, модет быть продемонстрирована оптимальность аппроксимации обратной к матрице Гессе $\mathbf{H}_t$ в контексте выбранного подпространства $\mathcal{D}_t$.

***Замечание* 4.** *В предыдущих обозначениях, пусть* $\operatorname{rank} \mathbf{D}_t = m$. *Тогда матрица* $\mathbf{H}_t\mathbf{A} = \mathbf{D}_t \left(\mathbf{D}_t^\top \mathbf{A}\mathbf{D}_t\right)^{-1} \mathbf{D}_t^\top \mathbf{A}$ *имеет собственное число 1 кратности* m, *а соответствующие ему собственные вектора — столбцы матрицы* $\mathbf{D}_t$.

Доказательство. Достаточно заметить, что

$$\mathbf{H}_t\mathbf{A}\mathbf{D}_t = \mathbf{D}_t \left(\mathbf{D}_t^\top \mathbf{A}\mathbf{D}_t\right)^{-1} \mathbf{D}_t^\top \mathbf{A}\mathbf{D}_t = \mathbf{D}_t.$$

□

Матрица $\mathbf{D}_t \left(\mathbf{D}_t^\top \mathbf{A}\mathbf{D}_t\right)^{-1} \mathbf{D}_t^\top$ — не что иное, как аппроксимация матрицы $\mathbf{A}^{-1}$ по направлениям столбцов матрицы $\mathbf{D}_t$. Замечание 4 демонстрирует, что эта аппроксимация вдоль соответствующих направлений точна.

Далее, воспользовавшись аппроксимацией обратной к матрице Гессе (2.8), рассмотрим процесс минимизации f:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \mathbf{D}_t \left(\mathbf{D}_t^\top \mathbf{A}\mathbf{D}_t\right)^{-1} \mathbf{D}_t^\top \nabla f\left(\mathbf{x_t}\right) = \mathbf{x}_t - \mathbf{H}_t \nabla f\left(\mathbf{x_t}\right). \tag{2.9}$$

***Замечание* 5.** *В случае процесса оптимизации* (2.9) *имеют место следующие рекуррентные соотношения для оценок* $\mathbf{x}_t$ *и соответствующих значений градиента:*

$$\begin{aligned}\mathbf{x}_{t+1} - \mathbf{x}_\star &= (\mathbf{I} - \mathbf{H}_t\mathbf{A})\,(\mathbf{x}_t - \mathbf{x}_\star) \\ &= (\mathbf{I} - \mathbf{H}_t\mathbf{A})\left(\mathbf{I} - \mathbf{D}_t\mathbf{D}_t^\top\right)(\mathbf{x}_t - \mathbf{x}_\star),\end{aligned} \tag{2.10}$$

$$\nabla f\,(\mathbf{x}_{t+1}) = (\mathbf{I} - \mathbf{A}\mathbf{H}_t)\,\nabla f\,(\mathbf{x}_t)\,. \tag{2.11}$$

Доказательство. Оба соотношения следуют из уравнения хорд: $\nabla f\,(\mathbf{x}) = \nabla f\,(\mathbf{x}) - \nabla f\,(\mathbf{x}_\star) = \mathbf{A}\,(\mathbf{x} - \mathbf{x}_\star)$. Второе равенство в (2.10) дополнительно следует из Замечания 4: матрица $\mathbf{H}_t\mathbf{A}$ действует как единичная вдоль столбцов матрицы $\mathbf{D}_t$. □

Из Замечания 5 следует, что шаг по направлению $-\mathbf{H}_t\nabla f\,(\mathbf{x}_t)$, при обратной к матрице Гессе полученной по формуле (2.8), имеет следующий недостаток: из невязки $\mathbf{x}_t - \mathbf{x}_\star$ "удаляется" не вся составляющая подпространства $\mathcal{D}_t$: в общем случае $\mathbf{D}_t^\top\,(\mathbf{x}_{t+1} - \mathbf{x}_\star) \neq \mathbf{0}$. Это недостаток существенен, так как может приводить к проблеме зигзагообразных траекторий по аналогии с методом градиентного спуска [102; 103].

Шаг последовательной подпространственной оптимизации принадлежит линейной оболочке столбцов матрицы $\mathbf{D}_t$. Рассмотрим его разложение следующего вида:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \mathbf{D}_t\mathbf{D}_t^\top\mathbf{A}^{-1}\nabla f\,(\mathbf{x}_t) + \mathbf{D}_t\mathbf{D}_t^\top\boldsymbol{\xi}_t,$$

где $\mathbf{D}_t\mathbf{D}_t^\top\mathbf{A}^{-1}\nabla f\,(\mathbf{x}_t) = \mathbf{D}_t\mathbf{D}_t^\top(\mathbf{x}_t - \mathbf{x}_\star)$ — "оптимальный" шаг в том смысле, что он нивелирует невязку вдоль подпространства $\mathcal{D}_t$, а $\mathbf{D}_t\mathbf{D}_t^\top\boldsymbol{\xi}_t$ — соответствующая погрешность. Тогда, невязка на шаге $t+1$ раскладывается на две составляющие:

$$(\mathbf{x}_{t+1} - \mathbf{x}_\star) = (\mathbf{I} - \mathbf{H}_t\mathbf{A}_t)\,(\mathbf{x}_t - \mathbf{x}_\star) + \mathbf{H}_t\mathbf{A}_t\boldsymbol{\xi}_t,$$

— *ошибку проекции* $(\mathbf{I} - \mathbf{H}_t\mathbf{A}_t)\,(\mathbf{x}_t - \mathbf{x}_\star)$ и *ошибку аппроксимации* $\mathbf{H}_t\mathbf{A}_t\boldsymbol{\xi}_t$.

Следующая лемма характеризует влияние обеих ошибок на общую скорость сходимости.

***Лемма* 1.** *Рассмотрим коэффициенты* α_t *и* γ_t*, характеризующие относительные значения ошибки проекции и ошибки аппроксимации:*

$$\begin{aligned}\|(\mathbf{I} - \mathbf{H}_t\mathbf{A}_t)\,(\mathbf{x}_t - \mathbf{x}_\star)\|_2^2 &\leqslant (1 - \alpha_t)\|\mathbf{x}_t - \mathbf{x}_\star\|_2^2, \quad \alpha_t \in [0, 1], \\ \|\boldsymbol{\xi}_t\|_2^2 &\leqslant (1 - \gamma_t)\|\mathbf{x}_t - \mathbf{x}_\star\|_2^2, \quad \gamma_t \in [0, 1].\end{aligned} \tag{2.12}$$

Тогда, в зависимости от значений α_t и γ_t достигаются следующие скорости сходимости:

1. *При $\alpha_t, \gamma_t > 0$ имеет место сублинейная скорость сходимости;*
2. *При $\exists\, \varepsilon > 0$: $\alpha_t, \gamma_t \geqslant \varepsilon$ имеет место линейная скорость сходимости;*
3. *Для сверхлинейной сходимости достаточно выполнения одного из условий:*

   (a) $\begin{cases} \exists\, \alpha' > 0, t' < \infty : \forall t \geqslant t', \alpha_t \geqslant \alpha', \\ \gamma_t \to 1, \end{cases}$

   (b) $\begin{cases} \alpha_t \to 1, \\ \exists \gamma' > 0, t' < \infty : \forall t \geqslant t', \gamma_t \geqslant \gamma'. \end{cases}$

Доказательство. Рассмотрим, какой вклад ошибки проекции и аппроксимации вносят в общую скорость сходимости:

$$\begin{aligned} \|\mathbf{x}_{t+1} - \mathbf{x}_\star\|_2^2 &= \left\|\left(\mathbf{I} - \mathbf{D}_t\mathbf{D}_t^\top\right)(\mathbf{x}_t - \mathbf{x}_\star)\right\|_2^2 + \|\mathbf{D}_t\mathbf{D}_t^\top \boldsymbol{\xi}_t\|_2^2 \\ &\leqslant (1 - \alpha_t)\|\mathbf{x}_t - \mathbf{x}_\star\|_2^2 + \alpha_t(1 - \gamma_t)\|\mathbf{x}_t - \mathbf{x}_\star\|_2^2 \\ &= (1 - \alpha_t\gamma_t)\|\mathbf{x}_t - \mathbf{x}_\star\|_2^2. \end{aligned} \tag{2.13}$$

Утверждения Леммы следуют из уравнения (2.13) . □

Лемма 1 характеризует вклад ошибки аппроксимации и ошибки проекции в общую скорость сходимости методов последовательной подпространственной оптимизации с квадратичным суррогатом для случая квадратичной целевой функции. Примечателен тот факт, что роль обеих ошибок равноценна: если ошибка проекции велика, то уменьшение ошибки аппроксимации не повлияет качественно на скорость сходимости и наоборот.

2.1.3 Сильно выпуклый случай

Рассмотрим класс дважды дифференцируемых функций $f : \mathbb{R}^n \to \mathbb{R}$ с L-липшицевым градиентом и μ–сильной выпуклостью $\mathcal{F}_{\mu,L}$ (1.4). Обозначим $\mathbf{x}_\star = \operatorname{argmin}_{\mathbf{x}} f$ — точка минимума функции f, которая по строгой выпуклости существует и единственна.

В этом разделе исследованы достаточные условия различных скоростей сходимости для класса функций $\mathcal{F}_{\mu,L}$ в контексте подхода последовательной подпространственной оптимизации.

Лемма 2. *Пусть $f \in \mathcal{F}_{\mu,L}$. Рассмотрим метод с шагом $\mathbf{x}_{t+1} = \mathbf{x}_t - \alpha_t(\nabla f(\mathbf{x}_t) + \delta_t)$, где α_t — размер шага, а δ_t — погрешность, т.ч. $\delta_t^\top \nabla f(\mathbf{x}_t) = 0$, $\beta_t := \frac{\|\delta_t\|}{\|\nabla f(\mathbf{x}_t)\|}$. Тогда последовательность $\|\mathbf{x}_t - \mathbf{x}_\star\|$ ограничена снизу следующим рекуррентным соотношением:*

$$\|\mathbf{x}_{t+1} - \mathbf{x}_\star\|^2 \geqslant \|\mathbf{x}_t - \mathbf{x}_\star\|^2 \left(1 + \alpha_t^2(1+\beta_t^2)\mu^2 - 2\alpha_t L\right),$$
$$\|\mathbf{x}_{t+1} - \mathbf{x}_\star\|^2 \leqslant \|\mathbf{x}_t - \mathbf{x}_\star\|^2 \left(1 + \alpha_t^2(1+\beta_t^2)L^2 - 2\alpha_t\left(1 - \beta_t(L-\mu)\right)\right).$$

Доказательство. Введем $\triangle_t = \mathbf{x}_t - \mathbf{x}_\star$ и рассмотрим $\|\triangle_{t+1}\|^2$:

$$\begin{aligned}\|\triangle_{t+1}\|^2 &= \|\triangle_t\|^2 - 2\alpha_t \triangle_t^\top \nabla f(\mathbf{x}_t) + \alpha_t^2\|\nabla f(\mathbf{x}_t)\|^2 - 2\alpha_t \triangle_t^\top \delta_t \\ &\quad + 2\alpha_t^2 \delta_t^\top \nabla f(\mathbf{x}_t) + \alpha_t^2\|\delta_t\|^2 \\ &= \|\triangle_t\|^2 + \alpha_t^2(1+\beta_t^2)\|\nabla f(\mathbf{x}_t)\|^2 - 2\alpha_t \triangle_t^\top \nabla f(\mathbf{x}_t) \\ &\quad - 2\alpha_t \triangle_t^\top \delta_t.\end{aligned} \tag{2.14}$$

Воспользовавшись L-липшицевостью градиента и μ–выпуклостью функции, а так же тем фактом, что $\triangle_t^\top \delta_t \leqslant \beta_t \|\nabla f(\mathbf{x}_t)\| \|\triangle_t\| - \beta_t \triangle_t^\top \nabla f(\mathbf{x}_t)$, может быть получена оценка снизу на $\|\triangle_t\|$:

$$\begin{aligned}\|\triangle_{t+1}\|^2 &\geqslant \|\triangle_t\|^2 + \alpha_t^2(1+\beta_t^2)\|\nabla f(\mathbf{x}_t)\|^2 - 2\alpha_t(1-\beta_t)\triangle_t^\top \nabla f(\mathbf{x}_t) \\ &\quad - 2\alpha_t\beta_t\|\triangle_t\|\|\nabla f(\mathbf{x}_t)\| \\ &\geqslant \|\triangle_t\|^2\left(1 + \alpha_t^2(1+\beta_t^2)\mu^2 - 2\alpha_t L\right).\end{aligned}$$

Далее, рассмотрим верхнюю границу на $\|\triangle_t\|$. Основной интерес представляет слагаемое $\triangle_t^\top \delta_t$, которое в общем случае может быть ограничено снизу как $\beta_t \triangle_t^\top \nabla f(\mathbf{x}_t) - \|\triangle_t\|\|\nabla f(\mathbf{x}_t)\|$. Тогда

$$\begin{aligned}\|\triangle_{t+1}\|^2 &\leqslant \|\triangle_t\|^2 + \alpha_t^2(1+\beta_t^2)\|\nabla f(\mathbf{x}_t)\|^2 - 2\alpha_t(1+\beta_t)\triangle_t^\top \nabla f(\mathbf{x}_t) \\ &\quad + 2\alpha_t\beta_t\|\triangle_t\|\|\nabla f(\mathbf{x}_t)\| \\ &\leqslant \|\triangle_t\|^2\left(1 + \alpha_t^2(1+\beta_t^2)L^2 - 2\alpha_t\left(1 - \beta_t(L-\mu)\right)\right).\end{aligned}$$

□

***Замечание* 6.** *В условиях Леммы 2, скорость сходимости нормы невязки* $\|\triangle_t\|$ *не превосходит линейную, если* $\exists\varepsilon > 0$*:* $\beta_t \geqslant \frac{L^2}{\mu^2} - 1 + \varepsilon$*, либо* $\beta_t \leqslant \frac{L^2}{\mu^2} - 1 - \varepsilon$*:*

$$\alpha_t \in \left[\varepsilon, \frac{L - \sqrt{L^2 - (1+\beta_t^2)\mu^2}}{(1+\beta_t^2)\mu^2} - \varepsilon\right] \cup \left[\frac{L + \sqrt{L^2 - (1+\beta_t^2)\mu^2}}{(1+\beta_t^2)\mu^2} + \varepsilon, \infty\right].$$

Доказательство. Достаточно заметить, что корни уравнения $1 + \alpha_t^2(1+\beta_t^2)\mu^2 - 2\alpha_t L = 0$ имеют вид:

$$\alpha_\pm = \frac{L \pm \sqrt{L^2 - (1+\beta_t^2)\mu^2}}{(1+\beta_t^2)\mu^2}.$$

□

Полученные в Лемме 2 результаты представляют лишь теоретический интерес из-за сложности получения оценок констант липшицевости и строгой выпуклости L и μ. Следующая теорема частично исправляет этот недостаток, накладывая более сильные условия на шаг $\mathbf{d}_t$.

***Лемма* 3.** *Рассмотрим итеративный процесс оптимизации функции* $f \in \mathcal{F}_{\mu,L}$ *вида* $\mathbf{x}_{t+1} = \mathbf{x}_t - \alpha_t \mathbf{d}_t$*, где* $\mathbf{d}_t$ *— направление, а* α_t *— размер шага. Обозначим* $\delta_t = \mathbf{d}_t - \frac{\mathbf{d}_t^\top \nabla f(\mathbf{x}_t)}{\|\nabla f(\mathbf{x}_t)\|^2}\nabla f(\mathbf{x}_t)$*,* $\beta_t := \frac{\|\delta_t\|}{\|\nabla f(\mathbf{x}_t)\|}$ *и предположим, что* $(\mathbf{x}_t - \mathbf{x}_\star)^\top \delta_t \geqslant 0$*. Тогда, последовательность* $\|\mathbf{x}_t - \mathbf{x}_\star\|$ *ограничена сверху следующими рекуррентными соотношениями:*

– *при* $\alpha_t \in \left(0, \frac{2}{(1+\beta_t^2)(\mu+L)}\right]$*:*

$$\|\triangle_{t+1}\|^2 \leqslant \|\triangle_t\|^2 \left(1 + \alpha_t^2(1+\beta_t^2)\mu^2 - 2\alpha_t\mu\right), \tag{2.15}$$

– *при* $\alpha_t \in \left(\frac{2}{(1+\beta_t^2)(\mu+L)}, \frac{2}{(1+\beta_t^2)L}\right]$*:*

$$\|\triangle_{t+1}\|^2 \leqslant \|\triangle_t\|^2 \left(1 + \alpha_t^2(1+\beta_t^2)L^2 - 2\alpha_t L\right). \tag{2.16}$$

Доказательство. Обозначим $\delta_t = \mathbf{d}_t - \frac{\mathbf{d}_t^\top \nabla f(\mathbf{x}_t)}{\|\nabla f(\mathbf{x}_t)\|^2}\nabla f(\mathbf{x}_t)$, тогда $\delta_t^\top \nabla f(\mathbf{x}_t) = 0$ и аналогично уравнению (2.14) может быть получено разложение невязки:

$$\|\triangle_{t+1}\|^2 = \|\triangle_t\|^2 + \alpha_t^2(1+\beta_t^2)\|\nabla f(\mathbf{x}_t)\|^2 - 2\alpha_t \triangle_t^\top \nabla f(\mathbf{x}_t) - 2\alpha_t \triangle_t^\top \delta_t.$$

По условию теоремы $\delta_t^\top \triangle_t \geqslant 0$, следовательно:

$$\begin{aligned}
\|\triangle_{t+1}\|^2 &\leqslant \|\triangle_t\|^2 \left(1 - \frac{2\alpha_t\mu L}{\mu+L}\right) + \alpha_t\|\nabla f(\mathbf{x}_t)\|^2 \left(\alpha_t(1+\beta_t^2) - \frac{2}{\mu+L}\right) \\
&\leqslant^1 \|\triangle_t\|^2 \left(1 - \frac{2\alpha_t\mu L}{\mu+L} + \alpha_t^2(1+\beta_t^2)\mu^2 - \frac{2\alpha_t\mu^2}{\mu+L}\right) \\
&= \|\triangle_t\|^2 \left(1 + \alpha_t^2(1+\beta_t^2)\mu^2 - 2\alpha_t\mu\right)
\end{aligned}$$

где 1–ое неравенство, а с ним и линейная сходимость $\|\triangle_t\|$ выполняются при следующих условиях на α_t:

$$0 < \alpha_t \leqslant \frac{2}{(1+\beta_t^2)(\mu+L)}.$$

Если же $\alpha_t \geqslant \frac{2}{(1+\beta_t^2)(\mu+L)}$, то аналогичным образом может быть получено следующее рекуррентное неравенство:

$$\|\triangle_{t+1}\|^2 \leqslant \|\triangle_t\|^2 \left(1 + \alpha_t^2(1+\beta_t^2)L^2 - 2\alpha_t L\right),$$

из которого следует, что монотонная сходимость также имеет место при следующих α_t:

$$\frac{2}{(1+\beta_t^2)(\mu+L)} \leqslant \alpha_t < \frac{2}{(1+\beta_t^2)L}.$$

□

По сравнению с Леммой 2, Лемма 3 представляет более удобный на практике результат. Для удовлетворения ограничения на α_t достаточно получить верхнюю оценку на L исходя из физического смысла задачи, либо используя приближенные вычислительные процедуры (смотри, например, степенной метод [104], § 53, либо метод итераций с отношениями Рэлея в [105], а так же в [106], § 4.6). Значения β_t не имеют явных ограничений, неявно же величины β_t ограничивает необходимость выбора соответствующих α_t.

***Теорема* 2.** *Рассмотрим $f \in \mathcal{F}_{\mu,L}$, процесс оптимизации $\mathbf{x}_{t+1} = \mathbf{x}_t - \alpha_t \mathbf{d}_t$, где α_t — размер шага, а $\mathbf{d}_t$ — его направление. Обозначим $\gamma_t := \frac{\mathbf{d}_t^\top \nabla f(\mathbf{x}_t)}{\nabla f(\mathbf{x}_t)^\top \nabla f(\mathbf{x}_t)}$, $\beta_t := \frac{\|\mathbf{d}_t/\gamma_t - \nabla f(\mathbf{x}_t)\|}{\|\nabla f(\mathbf{x}_t)\|}$ и допустим $\beta_t \leqslant \frac{\rho}{L}$, где $\rho \in (0,1)$ и $\alpha_t = \frac{1-\rho}{\gamma_t(L^2+\rho^2)}$. Тогда последовательность $\|\mathbf{x}_t - \mathbf{x}_\star\|$ сходится к нулю линейно:*

$$\|\mathbf{x}_{t+1} - \mathbf{x}_\star\|^2 \leqslant \left(1 - \frac{1-\rho}{L^2+\rho^2}\right) \|\mathbf{x}_t - \mathbf{x}_\star\|^2. \tag{2.17}$$

Доказательство. Согласно Лемме 2 выполняется следующее рекуррентное мажорирующее соотношение для $\|\mathbf{x}_t - \mathbf{x}_\star\|$:

$$\|\mathbf{x}_{t+1} - \mathbf{x}_\star\|^2 \leqslant \|\mathbf{x}_t - \mathbf{x}_\star\|^2 \left(1 + \overline{\alpha}_t^2(1+\beta_t^2)L^2 - 2\overline{\alpha}_t\left(1 - \beta_t(L-\mu)\right)\right),$$

где $\overline{\alpha}_t = \alpha_t \gamma_t$.

Подставив верхнюю границу на β_t, получим

$$\|\mathbf{x}_{t+1} - \mathbf{x}_\star\|^2 \leqslant \|\mathbf{x}_t - \mathbf{x}_\star\|^2 \left(1 + \overline{\alpha}_t^2(L^2 + \rho^2)L^2 - 2\overline{\alpha}_t\,(1-\rho)\right).$$

Легко заметить, что минимум $\overline{\alpha}_t^2(L^2 + \rho^2)L^2 - 2\overline{\alpha}_t\,(1-\rho)$ достигается при $\overline{\alpha_t} = \frac{1-\rho}{L^2+\rho^2}$ и равен $-\frac{1-\rho}{L^2+\rho^2}$. Отсюда $\alpha_t = \frac{1-\rho}{\gamma_t(L^2+\rho^2)}$.

□

Теорема 2 демонстрирует, что линейная сходимость для строго выпуклых функций с липшицевым градиентом, являющаяся известным фактом для градиентного спуска (смотри, например, Теорему 2.1.15 в [62], либо Теорему 2, § 4, Гл. 1 в [65]), сохраняется и при выборе направлений равномерно ограниченно отличающихся от направления градиента. Это отличие характеризуется коэффициентами β_t и γ_t. Примечательно, что γ_t входит лишь в размер шага алгоритма, но отсутствует в оценке скорости сходимости — таким образом отличие размера выбранного направления $\mathbf{d}_t$ от градиента в ℓ_2–норме компенсируется размером шага. В оценку скорости сходимости входит лишь параметр ρ, который ограничивает разницу между выбранным направлением и направлением градиента.

Получив условия на линейную скорость сходимости, перейдем к изучению условий, необходимых для обеспечения суперлинейной скорости сходимости. Следующая лемма предоставляет верхнюю границу невязки для квазиньютоновских методов.

***Лемма* 4.** *Пусть $f \in \mathcal{F}_{\mu,L}$ и $\nabla^2 f$ удовлетворяет условию Липшица с константой L'. Рассмотрим метод с шагом $\mathbf{x}_{t+1} = \mathbf{x}_t - \alpha_t \mathbf{H}_t \nabla f(\mathbf{x}_t)$, тогда норма невязки удовлетворяет следующему рекуррентному соотношению:*

$$\begin{aligned}\|\mathbf{x}_{t+1} - \mathbf{x}_\star\| \leqslant \alpha_t \frac{L}{2}\|\mathbf{H}_t\|\|\mathbf{x}_{t+1} - \mathbf{x}_\star\|^2 + (1-\alpha_t)\|\mathbf{x}_t - \mathbf{x}_\star\| \\ + \alpha_t\|(\mathbf{I} - \mathbf{H}_t\nabla^2 f(\mathbf{x}_t))(\mathbf{x}_t - \mathbf{x}_\star)\|.\end{aligned}$$

Доказательство.

$$\begin{aligned}\|\mathbf{x}_{t+1}-\mathbf{x}_\star\| &= \|\mathbf{x}_t-\mathbf{x}_\star-\alpha_t\mathbf{H}_t\nabla^2 f(\mathbf{x}_t)(\mathbf{x}_t-\mathbf{x}_\star)+\alpha_t\mathbf{H}_t\nabla^2 f(\mathbf{x}_t)(\mathbf{x}_t-\alpha_t\mathbf{x}_\star)\\ &\quad-\alpha_t\mathbf{H}_t\nabla f\left(\mathbf{x}_t\right)\|\\ &\leqslant \alpha_t\|\left(\mathbf{I}-\mathbf{H}_t\nabla^2 f(\mathbf{x}_t)\right)\left(\mathbf{x}_t-\mathbf{x}_\star\right)\|+(1-\alpha_t)\|\mathbf{x}_t-\mathbf{x}_\star\|\\ &\quad+\alpha_t\|\mathbf{H}_t\left(\nabla^2 f(\mathbf{x}_t)(\mathbf{x}_t-\mathbf{x}_\star)-\nabla f\left(\mathbf{x}_t\right)\right)\|,\end{aligned}$$

согласно утверждению 2:

$$\begin{aligned}&\leqslant \alpha_t\|\left(\mathbf{I}-\mathbf{H}_t\nabla^2 f(\mathbf{x}_t)\right)\left(\mathbf{x}_t-\mathbf{x}_\star\right)\|+(1-\alpha_t)\|\mathbf{x}_t-\mathbf{x}_\star\|\\ &\quad+\alpha_t\|\mathbf{H}_t\|\frac{L'}{2}\|\mathbf{x}_t-\mathbf{x}_\star\|^2.\end{aligned}$$

□

В следующей Теореме, использующей результаты Леммы (4), предоставляется характеристика скорости сходимости методов последовательной подпространственной оптимизации с квазиньютоновским шагом вида (2.9).

Теорема 3. *Пусть $f\in\mathcal{F}_{\mu,L}$ и $\nabla^2 f$ удовлетворяет условию Липшица с константой L'. Рассмотрим метод с шагом $\mathbf{x}_{t+1}=\mathbf{x}_t-\alpha_t\mathbf{H}_t\nabla f\left(\mathbf{x}_t\right)$, где $\mathbf{H}_t=\mathbf{D}_t\mathbf{Q}_t^{-1}\mathbf{D}_t^\top$. Допустим, выполняются следующие условия:*

1. *матрицы $\mathbf{D}_t$ выбираются таким образом, что $\mathbf{D}_t\mathbf{D}_t^\top\nabla f\left(\mathbf{x}_t\right)=\nabla f\left(\mathbf{x}_t\right)$, а $\mathbf{D}_t^\top\mathbf{D}_t=\mathbf{I}$;*
2. *шаг α_t выбирается таким образом, что $1-\|\nabla f\left(\mathbf{x}_t\right)\|\leqslant\alpha_t\leqslant 1$;*
3. *матрицы $\mathbf{D}_t$ выбираются таким образом, что $\exists\, c_0, C_0>0$: $\alpha_t\|(\mathbf{I}-\mathbf{D}_t\mathbf{D}_t^\top)\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1}\nabla f\left(\mathbf{x}_t\right)\|\leqslant C_0\|\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1}\nabla f\left(\mathbf{x}_t\right)\|^{1+c_0}$;*
4. *точность решения задачи подпространственной оптимизации прогрессирует с суперлинейной скоростью относительно нормы градиента: $\exists\, c_1, C_1>0$: $\|\left[\mathbf{D}_t^\top\nabla^2 f(\mathbf{x}_t)\mathbf{D}_t\right]^{-1}\mathbf{D}_t^\top\nabla f\left(\mathbf{x}_t\right)-\mathbf{Q}_t^{-1}\mathbf{D}_t^\top\nabla f\left(\mathbf{x}_t\right)\|\leqslant C_1\|\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1}\nabla f\left(\mathbf{x}_t\right)\|^{1+c_1}$.*

Тогда, метод сходится суперлинейно:

$$\exists C>0, c>0:\quad \|\mathbf{x}_{t+1}-\mathbf{x}_\star\|\leqslant C\|\mathbf{x}_t-\mathbf{x}_\star\|^{1+c},$$

где $C=\min\left(C_0, 2C_1, \left(L'\|\mathbf{Q}_t^{-1}\|+\frac{L'^2}{2}+L\right)\right)$, $c=\min(c_0,c_1,1)$.

Доказательство. Подставим в выражение границы ошибки из Леммы 4 матрицу $\mathbf{H}_t = \mathbf{D}_t\mathbf{Q}_t^{-1}\mathbf{D}_t^\top$

$$\|\mathbf{x}_{t+1} - \mathbf{x}_\star\| \leqslant \alpha_t\|\left(\mathbf{I} - \mathbf{D}_t\mathbf{Q}_t^{-1}\mathbf{D}_t^\top\nabla^2 f(\mathbf{x}_t)\right)(\mathbf{x}_t - \mathbf{x}_\star)\| + (1-\alpha_t)\|\mathbf{x}_t - \mathbf{x}_\star\| + \alpha_t\|\mathbf{D}_t\mathbf{Q}_t^{-1}\mathbf{D}_t^\top\|\frac{L'}{2}\|\mathbf{x}_t - \mathbf{x}_\star\|^2.$$

Заметим, что $\|\mathbf{D}_t\mathbf{Q}_t^{-1}\mathbf{D}_t^\top\| \leqslant \|\mathbf{Q}_t^{-1}\|$, и рассмотрим $\|\left(\mathbf{I} - \mathbf{H}_t\nabla^2 f(\mathbf{x}_t)\right)(\mathbf{x}_t - \mathbf{x}_\star)\|$

$$\begin{aligned}
&\|\left(\mathbf{I} - \mathbf{H}_t\nabla^2 f(\mathbf{x}_t)\right)(\mathbf{x}_t - \mathbf{x}_\star)\| \\
&= \|(\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1} - \mathbf{D}_t\mathbf{Q}_t^{-1}\mathbf{D}_t^\top)\nabla^2 f(\mathbf{x}_t)(\mathbf{x}_t - \mathbf{x}_\star)\| \\
&\leqslant \|(\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1} - \mathbf{D}_t\mathbf{Q}_t^{-1}\mathbf{D}_t^\top)\nabla f(\mathbf{x}_t)\| \\
&\quad + \|\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1} - \mathbf{D}_t\mathbf{Q}_t^{-1}\mathbf{D}_t^\top\|\frac{L'}{2}\|\mathbf{x}_t - \mathbf{x}_\star\|^2 \\
&\leqslant \|(\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1} - \mathbf{D}_t\mathbf{Q}_t^{-1}\mathbf{D}_t^\top)\nabla f(\mathbf{x}_t)\| + \left(\frac{1}{\mu} + \|\mathbf{Q}_t^{-1}\|\right)\frac{L'}{2}\|\mathbf{x}_t - \mathbf{x}_\star\|^2.
\end{aligned}$$

Рассмотрим $\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1}\nabla f(\mathbf{x}_t) - \mathbf{D}_t\mathbf{Q}_t^{-1}\mathbf{D}_t^\top\nabla f(\mathbf{x}_t)$:

$$\begin{aligned}
&\|\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1}\nabla f(\mathbf{x}_t) - \mathbf{D}_t\mathbf{Q}_t^{-1}\mathbf{D}_t^\top\nabla f(\mathbf{x}_t)\| \\
&\qquad \leqslant \|\mathbf{D}_t\mathbf{D}_t^\top\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1}\nabla f(\mathbf{x}_t) - \mathbf{D}_t\mathbf{Q}_t^{-1}\mathbf{D}_t^\top\nabla f(\mathbf{x}_t)\| \\
&\qquad + \|(\mathbf{I} - \mathbf{D}_t\mathbf{D}_t^\top)\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1}\nabla f(\mathbf{x}_t)\| \\
&\qquad \leqslant \|\mathbf{D}_t^\top\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1}\nabla f(\mathbf{x}_t) - \mathbf{Q}_t^{-1}\mathbf{D}_t^\top\nabla f(\mathbf{x}_t)\| \\
&\qquad + \|(\mathbf{I} - \mathbf{D}_t\mathbf{D}_t^\top)\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1}\nabla f(\mathbf{x}_t)\| \\
&\qquad \leqslant \|\mathbf{D}_t^\top\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1}\nabla f(\mathbf{x}_t) - \left[\mathbf{D}_t^\top\nabla^2 f(\mathbf{x}_t)\mathbf{D}_t\right]^{-1}\mathbf{D}_t^\top\nabla f(\mathbf{x}_t)\| \\
&\qquad + \|\left[\mathbf{D}_t^\top\nabla^2 f(\mathbf{x}_t)\mathbf{D}_t\right]^{-1}\mathbf{D}_t^\top\nabla f(\mathbf{x}_t) - \mathbf{Q}_t^{-1}\mathbf{D}_t^\top\nabla f(\mathbf{x}_t)\| \\
&\qquad + \|(\mathbf{I} - \mathbf{D}_t\mathbf{D}_t^\top)\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1}\nabla f(\mathbf{x}_t)\| \\
&\qquad \leqslant \|\left[\mathbf{D}_t^\top\nabla^2 f(\mathbf{x}_t)\mathbf{D}_t\right]^{-1}\mathbf{D}_t^\top\nabla f(\mathbf{x}_t) - \mathbf{Q}_t^{-1}\mathbf{D}_t^\top\nabla f(\mathbf{x}_t)\| \\
&\qquad + \left(1 + \frac{L}{\mu}\right)\|(\mathbf{I} - \mathbf{D}_t\mathbf{D}_t^\top)\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1}\nabla f(\mathbf{x}_t)\|.
\end{aligned}$$

Поясним последнее неравенство

$$\begin{aligned}
&\|\mathbf{D}_t^\top\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1}\nabla f(\mathbf{x}_t) - \left[\mathbf{D}_t^\top\nabla^2 f(\mathbf{x}_t)\mathbf{D}_t\right]^{-1}\mathbf{D}_t^\top\nabla f(\mathbf{x}_t)\| \\
&\quad \leqslant \|\left[\mathbf{D}_t^\top\nabla^2 f(\mathbf{x}_t)\mathbf{D}_t\right]^{-1}\mathbf{D}_t\nabla^2 f(\mathbf{x}_t)(\mathbf{I} - \mathbf{D}_t\mathbf{D}_t^\top)\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1}\nabla f(\mathbf{x}_t)\| \\
&\quad \leqslant \|\left[\mathbf{D}_t^\top\nabla^2 f(\mathbf{x}_t)\mathbf{D}_t\right]^{-1}\mathbf{D}_t\nabla^2 f(\mathbf{x}_t)\|\|(\mathbf{I} - \mathbf{D}_t\mathbf{D}_t^\top)\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1}\nabla f(\mathbf{x}_t)\| \\
&\quad \leqslant \frac{L}{\mu}\|(\mathbf{I} - \mathbf{D}_t\mathbf{D}_t^\top)\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1}\nabla f(\mathbf{x}_t)\|.
\end{aligned}$$

Вновь подставляя полученные границы в выражение невязки из Леммы 4, получим:

$$
\begin{aligned}
\|\mathbf{x}_{t+1} - \mathbf{x}_\star\| &\leqslant \alpha_t \|\left(\mathbf{I} - \mathbf{D}_t\mathbf{Q}_t^{-1}\mathbf{D}_t^\top \nabla^2 f(\mathbf{x}_t)\right)(\mathbf{x}_t - \mathbf{x}_\star)\| + (1-\alpha_t)\|\mathbf{x}_t - \mathbf{x}_\star\| \\
&+ \alpha_t \|\mathbf{D}_t\mathbf{Q}_t^{-1}\mathbf{D}_t^\top\| \frac{L'}{2}\|\mathbf{x}_t - \mathbf{x}_\star\|^2, \\
&\text{в силу полученных выше ограничений и условий на } \alpha_t \\
&\leqslant \alpha_t \|\left[\mathbf{D}_t^\top \nabla^2 f(\mathbf{x}_t)\mathbf{D}_t\right]^{-1}\mathbf{D}_t^\top \nabla f(\mathbf{x}_t) - \mathbf{Q}_t^{-1}\mathbf{D}_t^\top \nabla f(\mathbf{x}_t)\| \\
&+ 2\alpha_t \|(\mathbf{I} - \mathbf{D}_t\mathbf{D}_t^\top)\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1}\nabla f(\mathbf{x}_t)\| \\
&+ \alpha_t \frac{L'}{2}\left(\frac{1}{\mu} + \|\mathbf{Q}_t^{-1}\|\right)\|\mathbf{x}_t - \mathbf{x}_\star\|^2 + L\|\mathbf{x}_t - \mathbf{x}_\star\|^2 \\
&+ \alpha_t \frac{L'}{2}\|\mathbf{Q}_t^{-1}\|\|\mathbf{x}_t - \mathbf{x}_\star\|^2, \\
&\text{подставив ограничения из условий Теоремы:} \\
&\leqslant \alpha_t C_0 \|\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1}\nabla f(\mathbf{x}_t)\|^{1+c_0} + 2\alpha_t C_1 \|\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1}\nabla f(\mathbf{x}_t)\|^{1+c_1} \\
&+ \left(\alpha_t L'\|\mathbf{Q}_t^{-1}\| + \alpha_t\frac{L'^2}{2} + L\right)\|\mathbf{x}_t - \mathbf{x}_\star\|^2 \\
&\leqslant \min(C_0, 2C_1)\|\mathbf{x}_t - \mathbf{x}_\star\|^{1+\min(c_0,c_1)} + \alpha_t C_0\|\mathbf{x}_t - \mathbf{x}_\star\|^{2+2c_0} \\
&+ 2\alpha_t C_1\|\mathbf{x}_t - \mathbf{x}_\star\|^{2+2c_1} + \left(\alpha_t L'\|\mathbf{Q}_t^{-1}\| + \alpha_t\frac{L'^2}{2} + L\right)\|\mathbf{x}_t - \mathbf{x}_\star\|^2.
\end{aligned}
$$

□

Теорема 3 предоставляет достаточные условия для суперлинейной (а в зависимости от значений c_0 и c_1 — и квадратичной) сходимости. Она наглядно показывает, что скорость сходимости методов последовательной подпространственной оптимизации в равной степени зависит от качества выбираемых подпространств и точности решения подпространственной задачи. Условия (1) и (2) представляются наиболее простыми в достижении. Так, выбрав $\alpha_t = 1$, и $\mathbf{D}_t = [\nabla f(\mathbf{x}_t)/\|\nabla f(\mathbf{x}_t)\|]$ они очевидно будут соблюдены. Условия же (3) — характеризующее угол между $\mathcal{D}_t$ и вектором $\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1}\nabla f(\mathbf{x}_t)$ и (4) — характеризующее точность решения подпространственной задачи, представляют наибольший интерес, а вопрос их выполнения остается открытым. В последующих разделах мы рассмотрим несколько подходов как к построению матриц $\mathbf{D}_t$, так и к решению подпространственной задачи (т.е. оценке вектора $\mathbf{Q}_t^{-1}\mathbf{D}_t^\top \nabla f(\mathbf{x}_t)$), в контексте полученных в Теореме 2 и Лемме 3 результатов.

2.2 Элементы методов последовательной подпространственной оптимизации

Как было упомянуто выше, формирования подпространств, а также способ решения подпространственной задачи оптимизации — две основные характеристики методов ППО. В этом разделе подробно рассмотрены подходы к каждой из них.

2.2.1 Оценка шага в подпространстве

Принимая во внимание разложение по Тейлору (2.6), естественно рассмотреть квадратичный суррогат вида:

$$q_t(\mathbf{z}) = f(\mathbf{x}_t) + \nabla f\left(\mathbf{x}_t\right)^\top \mathbf{D}_t\mathbf{z} + \frac{1}{2}\mathbf{z}^\top\mathbf{D}_t^\top\nabla^2 f(\mathbf{x}_t)\mathbf{D}_t\mathbf{z}. \tag{2.18}$$

Поиск точки оптимума суррогата (2.18) эквивалентен решению системы линейных уравнений:

$$\mathbf{D}_t^\top\nabla^2 f(\mathbf{x}_t)\mathbf{D}_t\mathbf{z} = \mathbf{D}_t^\top\nabla f\left(\mathbf{x}_t\right), \tag{2.19}$$

которое при условии существования обратной матрицы $\left[\mathbf{D}_t^\top\nabla^2 f(\mathbf{x}_t)\mathbf{D}_t\right]^{-1}$ достигается в следующей точке:

$$\mathbf{z}_\star^{(t)} = \left[\mathbf{D}_t^\top\nabla^2 f(\mathbf{x}_t)\mathbf{D}_t\right]^{-1}\mathbf{D}_t^\top\nabla f\left(\mathbf{x}_t\right). \tag{2.20}$$

Вычисление матрицы $\mathbf{D}_t^\top\nabla^2 f(\mathbf{x}_t)\mathbf{D}_t$ в общем случае требует вычисления Гессиана исходной функции $\nabla^2 f(\mathbf{x}_t)$, что может быть затруднительно из-за ограничений по памяти, времени вычислений, или отсутствия явного выражения Гессиана. В силу этих причин, вместо оптимального (в терминах Теоремы 3) суррогата (2.18) рассмотрим суррогат с неизвестной матрицей квадратичных коэффициентов $\mathbf{Q}_t$:

$$q_t(\mathbf{z}) = f(\mathbf{x}_t) + \nabla f\left(\mathbf{x}_t\right)^\top \mathbf{D}_t\mathbf{z} + \frac{1}{2}\mathbf{z}^\top\mathbf{Q}_t\mathbf{z}. \tag{2.21}$$

В свою очередь, вместо системы (2.19) рассмотрим систему

$$\mathbf{Q}_t\mathbf{z} = \mathbf{D}_t^\top\nabla f\left(\mathbf{x}_t\right), \tag{2.22}$$

решение которой аналогично (2.20) задается следующим образом:

$$\mathbf{z}^{(t)} = \mathbf{Q}_t^{-1}\mathbf{D}_t^\top \nabla f\left(\mathbf{x}_t\right), \tag{2.23}$$

где единственным неизвестным элементом является обратная матрица Гессе $\mathbf{Q}_t^{-1}$. Определив $\mathbf{z}_t$, очередное направление шага алгоритма рассчитывается следующим образом:

$$\mathbf{d}_t = \mathbf{D}_t\mathbf{Q}_t^{-1}\mathbf{D}_t^\top \nabla f\left(\mathbf{x}_t\right). \tag{2.24}$$

В разделе 2.1 продемонстрированы некоторые свойства методов последовательной подпространственной, основанных на суррогатах подобного вида. Далее рассмотрены несколько способов построения вектора 2.23 напрямую, либо посредством оценки матрицы $\mathbf{Q}_t^{-1}$. При этом, целью является получение оценки близкой к оптимальной (2.20). В качестве исходных данных для получения оценок использованы историю аргументов $\{\mathbf{x}_t\}$, соответствующих значений градиентов $\nabla f\left(\mathbf{x}_t\right)$ и значений функции $f(\mathbf{x}_t)$. Заметим, что конечной целью является получение оценки вектора (2.20) вида (2.23): построение оценки матрицы $\mathbf{Q}_t^{-1}$ не необходимо и методы, получающие оценку (2.23) без явного построения $\mathbf{Q}_t^{-1}$, представляют особый интерес.

Далее рассмотрим несколько подходов к построению оценок вектора $\mathbf{z}_\star^{(t)}$ (2.20) вида (2.23) на основе использования *усеченной истории* значений суррогата q_t, его аргументов и градиентов: $\{\mathbf{z}_j^{(t)}\}_{i=1}^K \subset \mathbb{R}^m$, $\{q_t(\mathbf{z}_j^{(t)})\}_{i=1}^K \subset \mathbb{R}$ и $\{\nabla q_t(\mathbf{z}_j^{(t)})\}_{i=1}^K \subset \mathbb{R}^m$. Заметим, что матрица $\mathbf{Q}_t$ неизвестна, а следовательно неизвестны значения суррогата $\{q_t(\mathbf{z}_j^{(t)})\}_{j=1}^K \subset \mathbb{R}$ и его градиента $\{\nabla q_t(\mathbf{z}_j^{(t)})\}_{j=1}^K \subset \mathbb{R}^m$. Вместо них рассмотрим следующие приближения:

$$\mathbf{z}_j^{(t)} = \mathbf{D}_t^\top(\mathbf{w}_j^{(t)} - \mathbf{x}_t), \tag{2.25}$$

$$q_t(\mathbf{z}_j^{(t)}) \approx f(\mathbf{w}_j), \tag{2.26}$$

$$\nabla q_t(\mathbf{z}_j^{(t)}) \approx \mathbf{D}_t^\top \nabla f(\mathbf{w}_j^{(t)}), \tag{2.27}$$

где в качестве $\mathbf{w}_j^{(t)}$ может быть выбран, например предшествующий вектор $\mathbf{x}_{t-j}$, либо случайный вектор из $\mathbb{R}^n$.

***Замечание* 7.** *Заметим, что приближения* (2.26) *и* (2.27) *не являются точными даже при условии оптимальной оценки* $\mathbf{Q}_t$. *Так, приближенные неравенства становятся точными лишь при условии* $\mathbf{Q}_t = \mathbf{D}_t^\top \nabla^2 f(\mathbf{x}_t)\mathbf{D}_t$ *и квадратичной функции*

f. Действительно, пусть $\mathbf{Q}_t = \mathbf{D}_t^\top \nabla^2 f(\mathbf{x}_t)\mathbf{D}_t$, тогда

$$\nabla q_t(\mathbf{z}_j^{(t)}) - \mathbf{D}_t^\top \nabla f(\mathbf{w}_j^{(t)}) = \mathbf{D}_t^\top \nabla^2 f(\mathbf{x}_t)(\mathbf{I} - \mathbf{D}_t\mathbf{D}_t^\top)(\mathbf{w}_j^{(t)} - \mathbf{x}_t) + o(\|\mathbf{w}_j^{(t)} - \mathbf{x}_t\|),$$
$$q_t(\mathbf{z}_j^{(t)}) - f(\mathbf{w}_j^{(t)}) = \frac{1}{2}(\mathbf{w}_j^{(t)} - \mathbf{x}_t)^\top \left(\nabla^2 f(\mathbf{x}_t) - \mathbf{D}_t\mathbf{D}_t^\top \nabla^2 f(\mathbf{x}_t)\mathbf{D}_t\mathbf{D}_t^\top\right)(\mathbf{w}_j^{(t)} - \mathbf{x}_t)$$
$$+\nabla f\left(\mathbf{x}_t\right)^\top (\mathbf{I} - \mathbf{D}_t\mathbf{D}_t^\top)(\mathbf{w}_j^{(t)} - \mathbf{x}_t) + o(\|\mathbf{x}_t - \mathbf{w}_j^{(t)}\|^2).$$

Если дополнительно $\mathbf{D}_t\mathbf{D}_t^\top(\mathbf{w}_j^{(t)} - \mathbf{x}_t) = (\mathbf{w}_j^{(t)} - \mathbf{x}_t)$, то

$$\nabla q_t(\mathbf{z}_j^{(t)}) - \mathbf{D}_t^\top \nabla f(\mathbf{w}_j^{(t)}) = o(\|\mathbf{w}_j^{(t)} - \mathbf{x}_t\|),$$
$$q_t(\mathbf{z}_j^{(t)}) - f(\mathbf{w}_j^{(t)}) = o(\|\mathbf{w}_j^{(t)} - \mathbf{x}_t\|^2).$$

Наконец, если выполняется равенство $\mathbf{D}_t\mathbf{D}_t^\top(\mathbf{w}_j^{(t)} - \mathbf{x}_t) = (\mathbf{w}_j^{(t)} - \mathbf{x}_t)$ и функция f – квадратичная, то

$$\nabla q_t(\mathbf{z}_j^{(t)}) - \mathbf{D}_t^\top \nabla f(\mathbf{w}_j^{(t)}) = 0,$$
$$q_t(\mathbf{z}_j^{(t)}) - f(\mathbf{w}_j^{(t)}) = 0.$$

Доказательство. Разница между градиентом суррогата $\nabla q_t(\mathbf{z}_j^{(t)})$ и проекцией $\mathbf{D}_t^\top \nabla f(\mathbf{w}_j^{(t)})$ может быть выражена следующим образом:

$$\begin{aligned}
\nabla q_t(\mathbf{z}_j^{(t)}) - \mathbf{D}_t^\top \nabla f(\mathbf{w}_j^{(t)}) &= \mathbf{D}_t^\top \nabla f\left(\mathbf{x}_t\right) + \mathbf{D}_t^\top \nabla^2 f(\mathbf{x}_t)\mathbf{D}_t\mathbf{z}_j^{(t)} + o(\|\mathbf{w}_j^{(t)} - \mathbf{x}_t\|) \\
&- \mathbf{D}_t^\top \nabla f\left(\mathbf{x}_t\right) - \mathbf{D}_t^\top \nabla^2 f(\mathbf{x}_t)(\mathbf{w}_j^{(t)} - \mathbf{x}_t) \\
&= \mathbf{D}_t^\top \nabla^2 f(\mathbf{x}_t)\mathbf{D}_t\mathbf{z}_j^{(t)} - \mathbf{D}_t^\top \nabla^2 f(\mathbf{x}_t)(\mathbf{w}_j^{(t)} - \mathbf{x}_t) \\
&+ o(\|\mathbf{w}_j^{(t)} - \mathbf{x}_t\|) \\
&= \mathbf{D}_t^\top \nabla^2 f(\mathbf{x}_t)(\mathbf{I} - \mathbf{D}_t\mathbf{D}_t^\top)(\mathbf{w}_j^{(t)} - \mathbf{x}_t) + o(\|\mathbf{w}_j^{(t)} - \mathbf{x}_t\|).
\end{aligned}$$

Аналогичным образом, выражается разница между значением суррогата и значением функции:

$$\begin{aligned}
q_t(\mathbf{z}_j^{(t)}) - f(\mathbf{w}_j^{(t)}) &= f(\mathbf{x}_t) + \nabla f\left(\mathbf{x}_t\right)^\top \mathbf{D}_t\mathbf{z}_j^{(t)} + \frac{1}{2}\mathbf{z}_j^{(t)\top}\mathbf{D}_t^\top \nabla^2 f(\mathbf{x}_t)\mathbf{D}_t\mathbf{z}_j^{(t)} - f(\mathbf{x}_t) \\
&- \nabla f\left(\mathbf{x}_t\right)^\top (\mathbf{w}_j^{(t)} - \mathbf{x}_t) - \frac{1}{2}(\mathbf{w}_j^{(t)} - \mathbf{x}_t)^\top \nabla^2 f(\mathbf{x}_t)(\mathbf{w}_j^{(t)} - \mathbf{x}_t) \\
&+ o(\|\mathbf{x}_t - \mathbf{w}_j^{(t)}\|^2) \\
&= \nabla f\left(\mathbf{x}_t\right)^\top (\mathbf{I} - \mathbf{D}_t\mathbf{D}_t^\top)(\mathbf{w}_j^{(t)} - \mathbf{x}_t) + o(\|\mathbf{x}_t - \mathbf{w}_j^{(t)}\|^2) \\
&+ \frac{1}{2}(\mathbf{w}_j^{(t)} - \mathbf{x}_t)^\top \left[\nabla^2 f(\mathbf{x}_t) - \mathbf{D}_t\mathbf{D}_t^\top \nabla^2 f(\mathbf{x}_t)\mathbf{D}_t\mathbf{D}_t^\top\right](\mathbf{w}_j^{(t)} - \mathbf{x}_t).
\end{aligned}$$

□

2.2.2 Шаг в подпространстве через решение уравнения хорд

Уравнения хорд зачастую используются для оценки матрицы вторых производных. Например, уравнения хорд для градиента используются для получения оценки обратной к матрице Гессе в методах Давидона–Флетчера–Паувелла [14; 18], SR-1 [19], Бройдена–Флетчера–Гольдфарба–Шэнно [15; 18], L-BFGS [17], Барзилая-Борвейна [71] и др. Рассмотрим функцию $q : \mathbb{R}^m \to \mathbb{R}$ и две точки $\mathbf{z}, \mathbf{z}_0 \in \mathbb{R}^m$, тогда соответствующее уравнение хорд градиента функции ∇q выглядит следующим образом:

$$\nabla^2 q(\mathbf{z}')(\mathbf{z} - \mathbf{z}_0) = \nabla q(\mathbf{z}) - \nabla q(\mathbf{z}_0),$$

где $\mathbf{z}' \in \text{span}\{\mathbf{z},\mathbf{z}_0\}$. Заметим, что из разложения Тейлора легко получить приближенный вариант уравнения хорд: $\nabla^2 q(\mathbf{z})(\mathbf{z} - \mathbf{z}_0) = \nabla q(\mathbf{z}) - \nabla q(\mathbf{z}_0) + o(\|\mathbf{z} - \mathbf{z}_0\|)$. Более точная оценка отклонения возможна при условии L-липшицевости градиента: $\|\nabla^2 q(\mathbf{z})(\mathbf{z} - \mathbf{z}_0) - (\nabla q(\mathbf{z}) - \nabla q(\mathbf{z}_0))\|_2^2 \leqslant \frac{L}{2}\|\mathbf{z} - \mathbf{z}_0\|^2$ (смотри Лемму 1.2.3 в [62]).

Рассмотрим систему из уравнений хорд для q_t вида (2.21)

$$\mathbf{Q}_t \left(\mathbf{z}_0^{(t)} - \mathbf{z}_j^{(t)}\right) = \nabla q_t(\mathbf{z}_0^{(t)}) - \nabla q_t(\mathbf{z}_j^{(t)}), \quad j = 1 \dots K - 1, \tag{2.28}$$

где соответствующие значения аргументов градиентов оцениваются приближенно по формулам (2.26) и (2.27):

$$\begin{aligned} \mathbf{z}_j^{(t)} - \mathbf{z}_0^{(t)} &= \mathbf{z}_j^{(t)} = \mathbf{D}_t^\top \mathbf{w}_j^{(t)} \\ \nabla q_t(\mathbf{z}_j^{(t)}) - \nabla q_t(\mathbf{z}_0^{(t)}) &= \mathbf{D}_t^\top \left(\nabla f(\mathbf{w}_j^{(t)}) - \nabla f\left(\mathbf{x}_t\right)\right). \end{aligned}$$

Рассмотрим матрицы $\mathbf{Z}_t$ и $\mathbf{G}_t$, чьи строки составлены из разниц значений аргументов и градиентов соответственно:

$$\begin{aligned} \mathbf{Z}_t &= \left[\mathbf{D}_t^\top \mathbf{w}_1^{(t)}, \dots, \mathbf{D}_t^\top \mathbf{w}_K^{(t)}\right]^\top, \\ \mathbf{G}_t &= \left[\mathbf{D}_t^\top \left(\nabla f(\mathbf{w}_1^{(t)}) - \nabla f\left(\mathbf{x}_t\right)\right), \dots, \mathbf{D}_t^\top \left(\nabla f(\mathbf{w}_K^{(t)}) - \nabla f\left(\mathbf{x}_t\right)\right)\right]^\top. \end{aligned} \tag{2.29}$$

Тогда уравнения хорд, записанные в матричном виде $\mathbf{Z}_t\mathbf{Q}_t = \mathbf{G}_t$, будут служить приближением системы уравнений (2.28). Добавив естественное условие симметричности матрицы $\mathbf{Q}_t$ и домножив матричное уравнение хорд на $\mathbf{Q}_t^{-1}$ справа,

получим систему уравнений с ограничениями:

$$\begin{cases} \mathbf{G}_t \mathbf{Q}_t^{-1} = \mathbf{Z}_t \\ \mathbf{Q}_t^{-1} = \left[\mathbf{Q}_t^{-1}\right]^\top. \end{cases} \tag{2.30}$$

Следующая лемма предлагает способ решения полученной системы.

Лемма 5. *Следующая формула задает решение системы* (2.30) *:*

$$\begin{aligned} \hat{\mathbf{Q}}^{-1} = \mathbf{G}_t^+ \mathbf{Z}_t \\ + \left(\mathbf{I} - \mathbf{G}_t^+ \mathbf{G}_t\right) \left(\mathbf{G}_t^+ \mathbf{Z}_t\right)^\top, \end{aligned} \tag{2.31}$$

где $\mathbf{G}_t^+$ *— псевдообратная матрицы* $\mathbf{G}$*:* $\mathbf{G}_t^+ = \mathbf{G}_t^\top \left(\mathbf{G}_t \mathbf{G}_t^\top\right)^{-1}$.

Доказательство. Подставив в условия Теоремы 1 матрицу $\mathbf{G}_t^\top$ вместо $\mathbf{A}$ и $\mathbf{Z}_t$ вместо матрицы $\mathbf{B}$ заметим, что равенство $\mathbf{G}_t^\top \mathbf{Z}_t = \mathbf{Z}_t^\top \mathbf{G}_t$ следует из квадратичности q_t, а псевдообратная $\mathbf{G}_t^+$ удовлетворет условию $\mathbf{G}_t \mathbf{G}_t^+ = \mathbf{I}$ по определению. Таким образом, условия Теоремы 1 выполняются и уравнение (2.31) дает решение системы (2.30). □

Замечание 8. *Вычислительная сложность формулы* (1.8) *составляет* $\mathcal{O}\left(m^2 K + mK^2 + m^3\right)$ *по числу операций и* $\mathcal{O}\left(K^2 + mK + m^2\right)$ *по используемой памяти:*

- *для расчета* $\left(\mathbf{G}_t \mathbf{G}_t^\top\right)^{-1}$ *необходимо вычислить полное сингулярное разложение матрицы* $\mathbf{G}_t \in \mathbb{R}^{m \times K}$, $m \leqslant K$ *что требует* $\mathcal{O}\left(mK^2 + m^3\right)$ *операций и* $\mathcal{O}\left(K^2 + mK + m^2\right)$ *памяти (смотри [107], сводная таблица 8.6.1 на стр. 493), получив которое, достаточно возвести в квадрат и обратить диагональную матрицу с сингулярными числами, что потребует лишь* $\mathcal{O}(m)$ *операций;*
- $\mathcal{O}\left(m^2 K + mK^2\right)$ *операций и* $\mathcal{O}\left(mK + m^2\right)$ *памяти на прочие матричные операции.*

2.2.3 Шаг в подпространстве через прямое восстановление квазиньютоновского направления

Рассмотрим систему аналогичную (2.30), опустив требование симметричности матрицы $\mathbf{Q}_t$:

$$\mathbf{G}_t\mathbf{Q}_t^{-1} = \mathbf{Z}_t, \tag{2.32}$$

Заметим, что искомой величиной является не матрица $\mathbf{Q}_t^{-1}$, а ее произведение на проекцию градиента: $\mathbf{Q}_t^{-1}\mathbf{D}_t^\top\nabla f\left(\mathbf{x}_t\right)$. Домножив уравнение (2.32) справа на проекцию градиента, получим:

$$\mathbf{G}_t\mathbf{z}_t = \mathbf{Z}_t\mathbf{D}_t^\top\nabla f\left(\mathbf{x}_t\right), \tag{2.33}$$

— система линейных уравнений относительно искомого вектора $\mathbf{z}_t = \mathbf{Q}_t^{-1}\mathbf{D}_t^\top\nabla f\left(\mathbf{x}_t\right)$. Решение этой системы может быть получено методом наименьших квадратов:

$$\widehat{\mathbf{z}}_t = \left(\mathbf{G}_t^\top\mathbf{G}_t\right)^{-1}\mathbf{G}_t^\top\mathbf{Z}_t\mathbf{D}_t^\top\nabla f\left(\mathbf{x}_t\right). \tag{2.34}$$

***Замечание* 9.** *Вычислительная сложность уравнения* (2.34) *составляет* $\mathcal{O}\left(m^3 + m^2K\right)$, *где* $\mathcal{O}\left(m^3\right)$ *возникает из обращения квадратной* $m\times m$ *матрицы* $\mathbf{G}_t^\top\mathbf{G}_t$, *а* m^2K *из перемножения прямоугольных матриц* $\mathbf{G}_t^\top\mathbf{G}_t$ *и* $\mathbf{G}_t^\top\mathbf{Z}_t$.

2.2.4 Оценка матрицы Гессе регрессионным методом

Рассмотрим следующую систему уравнений:

$$\mathbf{z}_i^{(t)\top}\mathbf{Q}_t\mathbf{z}_i^{(t)} = f(\mathbf{x}_{t-i}) - \nabla f\left(\mathbf{x}_t\right)^\top\mathbf{D}_t\mathbf{z}_i^{(t)} - f(\mathbf{x}_t), \quad i = 0\ldots K-1. \tag{2.35}$$

Заметим, что выражение (2.35) эквивалентно приближению q_t (2.26) и что матрица $\mathbf{Q}_t$ входит в уравнения (2.35) линейно. Подобные уравнения могут быть решены методом наименьших квадратов. Для модели квадратичной регрессии введем оператор, устанавливающий биекцию между пространством симметричных матриц размера $m\times m$ и векторным пространством размерности $\frac{m(m+1)}{2}$:

$\text{vech} : \mathbb{R}^{m\times m} \to \mathbb{R}^{\frac{m(m+1)}{2}}$, и соответствующий обратный ему оператор $\text{vech}^{-1} : \mathbb{R}^{\frac{m(m+1)}{2}} \to \mathbb{R}^{m\times m}$:

$$\begin{aligned}\text{vech}\left(\mathbf{M}\right) = &\left[\mathbf{M}_{1,1},\ldots,\mathbf{M}_{m,1},\mathbf{M}_{2,2},\right.\\ &\left.\ldots,\mathbf{M}_{m,2},\ldots,\mathbf{M}_{m-1,m-1},\mathbf{M}_{m,m-1},\mathbf{M}_{m,m}\right]^\top\\ \text{vech}^{-1}\left(\mathbf{v}\right) = &\mathbf{M}:\ \mathbf{M}_{i,j} = \mathbf{v}_k,\\ &k = m(\min(i,j)-1)+\max(i,j)-1.\end{aligned}$$

Рассмотрим вектор $\mathbf{q}_t = \text{vech}\left(\mathbf{Q}_t\right)$, вектора $\psi_i^{(t)} = \text{vech}\left(\mathbf{z}_i^{(t)}\mathbf{z}_i^{(t)\top}\right)$ и скаляры $y_i^{(t)} = f(\mathbf{x}_{t-i}) - \nabla f\left(\mathbf{x}_t\right)^\top \mathbf{D}_t\mathbf{z}_i^{(t)} - f(\mathbf{x}_t)$, $i = 1..K$. Сформировав матрицу независимых переменных $\Psi_t = \left[\psi_1^{(t)},\ldots,\psi_K^{(t)}\right]^\top$ и вектор зависимой переменной $\mathbf{y}_t = \left(y_1^{(t)},\ldots,y_K^{(t)}\right)$ получим систему линейных уравнений относительно $\mathbf{q}_t$: $\Psi_t\mathbf{q}_t = \mathbf{y}_t$. Решение этой системы методом наименьших квадратов: $\widehat{\mathbf{q}}_t = (\Psi_t^\top\Psi_t)^{-1}\Psi_t^\top\mathbf{y}_t$. Соответствующая ей матрица может быть получена применением оператора vech^{-1}:

$$\widehat{\mathbf{Q}}_t = \text{vech}^{-1}\left((\Psi_t^\top\Psi_t)^{-1}\Psi_t^\top\mathbf{y}_t\right). \tag{2.36}$$

В том случае, когда функция f — полином второй степени вида (2.4), а разницы $\mathbf{x}_t - \mathbf{x}_{t-i}$ принадлежат линейной оболочке столбцов матрицы $\mathbf{D}_t$, то матрица $\mathbf{D}_t^\top\nabla^2 f(\mathbf{x}_t)\mathbf{D}_t$ является решением уравнения (2.35), так как выполняется следующее равенство:

$$f(\mathbf{x}_{t-i}) - \nabla f\left(\mathbf{x}_t\right)^\top \mathbf{D}_t\mathbf{z}_i^{(t)} - f(\mathbf{x}_t) = \frac{1}{2}\mathbf{z}_i^{(t)\top}\mathbf{D}_t^\top\mathbf{A}\mathbf{D}_t\mathbf{z}_i^{(t)}.$$

***Замечание* 10.** *Затраты на вычисление уравнения* (2.36) *и конструирования матриц* $\mathbf{\Psi}_t$ *и вектора* $\mathbf{y}_t$ *составляют* $\mathcal{O}\left(m^2K + m^4\right)$ *по памяти и* $\mathcal{O}\left(m^4K + m^6\right)$ *по числу операций.*

2.2.5 Построение подпространств на основе истории градиентов

Рассмотрим подпространство $\mathcal{D}_t$, заданное линейной оболочкой предшествующих значений градиента:

$$\mathcal{D}_t = \text{span}\left\{\nabla f\left(\mathbf{x}_{t-T}\right),\ldots,\nabla f\left(\mathbf{x}_t\right)\right\}. \tag{2.37}$$

Следующая лемма демонстрирует линейную скорость сходимости по подпространствам, образованным в том числе последним значением градиента $\nabla f(\mathbf{x}_t)$.

Лемма 6. *Пусть функция* $f \in \mathcal{F}_{\mu,L}$, *а* $\mathbf{D}_t \in \mathbb{R}^{d\times m}$, $m \geqslant 1$: $\mathbf{D}_t^\top \mathbf{D}_t = \mathbf{I}$ *и* $\mathbf{D}_t\mathbf{D}_t^\top \nabla f(\mathbf{x}_t) = \nabla f(\mathbf{x}_t)$. *Тогда*

$$\|(\mathbf{I} - \mathbf{D}_t^\top \mathbf{D}_t)(\mathbf{x}_t - \mathbf{x}_\star)\| \leqslant \left(1 - \frac{\mu}{L}\right)\|\mathbf{x}_t - \mathbf{x}_\star\|.$$

Доказательство. Заметим, что

$$\begin{aligned}\|\mathbf{D}_t^\top \mathbf{D}_t(\mathbf{x}_t - \mathbf{x}_\star)\| &\geqslant \|\|\nabla f(\mathbf{x}_t)\|^{-2}\nabla f(\mathbf{x}_t)\,\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{x}_\star)\| \\ &\geqslant \frac{\|\nabla f(\mathbf{x}_t)\,\mu\|\mathbf{x}_t - \mathbf{x}_\star\|^2\|}{\|\nabla f(\mathbf{x}_t)\|^2} = \frac{\mu\|\mathbf{x}_t - \mathbf{x}_\star\|^2}{\|\nabla f(\mathbf{x}_t)\|} \geqslant \frac{\mu}{L}\|\mathbf{x}_t - \mathbf{x}_\star\|.\end{aligned}$$

Остается лишь воспользоваться ортонормированностью столбцов $\mathbf{D}_t$:

$$\begin{aligned}\|(\mathbf{I} - \mathbf{D}_t^\top \mathbf{D}_t)(\mathbf{x}_t - \mathbf{x}_\star)\| &= \|\mathbf{x}_t - \mathbf{x}_\star\| - \|\mathbf{D}_t^\top \mathbf{D}_t(\mathbf{x}_t - \mathbf{x}_\star)\| \\ &\leqslant \left(1 - \frac{\mu}{L}\right)\|\mathbf{x}_t - \mathbf{x}_\star\|\end{aligned}$$

□

Таким образом, Лемма 6 демонстрирует линейную сходимость для подпространств вида (2.37) в терминах Теоремы 3, если столбцы матрицы $\mathbf{D}_t$ — ортонормированный базис пространства $\mathcal{D}_t$.

Рассмотрим рекуррентный алгоритм построения матрицы $\mathbf{D}_t$, чьи столбцы образуют множество (2.37).

Algorithm 2 updateSubspace1($\mathbf{D}$, $\mathbf{g}$, m)

$\mathbf{m_0} \leftarrow$ # of columns in $\mathbf{D}$
$\mathbf{g} \leftarrow \frac{\mathbf{g}}{\|\mathbf{g}\|}$
for i from 1 to m_0 **do**
  $\mathbf{D}_{\cdot,i} \leftarrow \mathbf{D}_{\cdot,i} - \mathbf{D}_{\cdot,i}^\top \mathbf{g}$
  $\mathbf{D}_{\cdot,i} \leftarrow \frac{\mathbf{D}_{\cdot,i}}{\|\mathbf{D}_{\cdot,i}\|}$
if $m_0 < m$ **then**
  $\mathbf{D} \leftarrow [\mathbf{D}_{\cdot,1}, \ldots, \mathbf{D}_{\cdot,m_0}, \mathbf{g}]$
else
  $\mathbf{D} \leftarrow [\mathbf{D}_{\cdot,2}, \ldots, \mathbf{D}_{\cdot,m_0}, \mathbf{g}]$
return $\mathbf{D}$

Благодаря условию 6 в Алгоритме 2 число столбцов матрицы $\mathbf{D}$ будет расти первые $m-1$ итераций пока не достигнет значения m. Заметим, что если на вход алгоритма передаются последовательно вектора $\mathbf{g}_1, \ldots, \mathbf{g}_m$, то столбцы итоговой матрицы $\mathbf{D}$ будут ортонормированным базисом линейной оболочки векторов $\{\mathbf{g}_i\}_1^m$.

***Замечание* 11.** *Вычислительная сложность Алгоритма 2 составляет $\mathcal{O}(mn)$ по используемой памяти и $\mathcal{O}(mn)$ по числу операций.*

2.3 Методы последовательной подпространственной оптимизации

В этом разделе приведены конкретные реализации методов последовательной подпространственной оптимизации, основанные на полученных в предыдущих разделах результатах, а так же исследованы их свойства: вычислительная сложность и условия на различные скорости сходимости.

2.3.1 Корректирующий метод ППО

В Алгоритме 1 представлена общая схема алгоритмов последовательной подпространственной оптимизации. В столь общей постановке затруднительно гарантировать сколь-нибудь хорошее поведение оптимизационного процесса. Однако, модифицировав существующий процесс оптимизации, можно добиться как минимум линейной скорости сходимости. Так в Алгоритме 3 приведен метод *корректировки* алгоритмов ППО, базирующийся на Теореме 2: если отличие шага сгенерированного методом ППО от градиента слишком велико, это отличие нивелируется (строки алгоритма 10–11), а если шаг ортогонален градиенту, то он пропускается и выполняется шаг в направлении аинтиградиента (строки 5–7).

***Замечание* 12.** *Порождаемая Алгоритмом 3 последовательность $\{\mathbf{x}_t\}_{t\geqslant 0}$ удовлетворяет условиям Теоремы 2 при любых процедурах subspaceUpdate и subspaceSearch.*

Algorithm 3 SSO_correction($\mathbf{x}_0$, ρ, L, subspaceUpdate, subspaceSearch)

$\mathbf{D}_t \leftarrow$ subspaceUpdate($\mathbf{D}_{t-1}$, $\nabla f(\mathbf{x}_t)$)
$\mathbf{z}_t \leftarrow$ subspaceSearch(f, $\mathbf{D}_t$, $\mathbf{x}_t$)
$\mathbf{d}_t \leftarrow \mathbf{D}_t\mathbf{z}_t$
$\gamma_t \leftarrow \frac{\mathbf{d}_t^\top \nabla f(\mathbf{x}_t)}{\|\nabla f(\mathbf{x}_t)\|^2}$
if $\gamma_t = 0$ **then**
    $\mathbf{d}_t \leftarrow \nabla f(\mathbf{x}_t)$
    $\alpha_t \leftarrow \frac{1}{L}$
else
    $\beta_t \leftarrow \frac{\|\mathbf{d}_t/\gamma_t - \nabla f(\mathbf{x}_t)\|}{\|\nabla f(\mathbf{x}_t)\|}$
    if $\beta_t > \frac{\rho}{L}$ **then**
        $\mathbf{d}_t \leftarrow \gamma_t \nabla f(\mathbf{x}_t) + \frac{\rho}{L\beta_t}(\mathbf{d}_t - \gamma_t \nabla f(\mathbf{x}_t))$
    $\alpha_t \leftarrow \frac{1-\rho}{\gamma_t(L^2+\rho^2)}$
$\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \alpha_t \mathbf{d}_t$
$t \leftarrow t+1$; go to 1

2.3.2 Квазиньютоновский метод ППО

В Алгоритме 4 приведён алгоритм последовательной подпространственной оптимизации, комбинирующий квазиньютоновский шаг и шаг метода сопряженных градиентов, опирающийся на метод конструирования подпространств (матриц $\mathbf{D}_t$), описанный в разделе 2.2.5 и изложенный в разделе 2.2.3 способ оценки квазиньютоновского направления.

***Замечание* 13.** *На каждой итерации Алгоритма 4 вектор $\mathbf{d}_t$ принадлежит линейной оболочке столбцов соответствующих матриц $\mathbf{D}_t$:* $\mathbf{d}_t \in \mathcal{D}_t = \operatorname{span}\{\nabla f\left(\mathbf{x}_{t-\min(m-1,t)}\right), \ldots, \nabla f(\mathbf{x}_t)\}$.

***Замечание* 14.** *Пусть $f \in \mathcal{F}_{\mu,L}$ — квадратичная функция. Тогда, последовательность оценок $\mathbf{x}_0, \ldots, \mathbf{x}_t, \ldots$, сгенерированных Алгоритмом 4, сходится к $\mathbf{x}_\star$ с линейной скоростью сходимости не более чем за n шагов.*

Доказательство. Заметим, что Алгоритм 4 совпадает с шагом метода сопряженных градиентов на первом шаге: $\mathbf{d}_0 = \mathbf{d}_0^{CG} = -\nabla f(\mathbf{x}_0)$. Предположим, что $\mathbf{d}_j = \mathbf{d}_j^{CG}\ \forall j = 0 \ldots t-1$ и пользуясь тем, что для метода сопряженных градиентов

Algorithm 4 L-QNSSO_step(f, $\mathbf{x}_0$, $\mathbf{D}_{t-1}$, m)

$\mathbf{D}_t \leftarrow \text{subspaceUpdate1}\left(\mathbf{D}_{t-1}, \nabla f\left(\mathbf{x}_t\right), m\right)$

if $t = 0$ **then**

$\mathbf{d}_t \leftarrow -\nabla f\left(\mathbf{x}_t\right)$; переход к шагу 11

$T \leftarrow \min(t, m-1)$

$\mathbf{Z}_t \leftarrow [\mathbf{D}_t^\top(\mathbf{x}_t - \mathbf{x}_{t-1}), \ldots, \mathbf{D}_t^\top(\mathbf{x}_t - \mathbf{x}_{t-T})]^\top$

$\mathbf{G}_t \leftarrow [\mathbf{D}_t^\top(\nabla f_t - \nabla f_{t-1}), \ldots, \mathbf{D}_t^\top(\nabla f_t - \nabla f_{t-T})]^\top$

$\mathbf{d}_t^{QN} \leftarrow -\mathbf{D}_t\left(\mathbf{G}_t^\top \mathbf{G}_t\right)^{-1}\mathbf{G}_t^\top \mathbf{Z}_t \mathbf{D}_t^\top \nabla f\left(\mathbf{x}_t\right)$

$\beta_t \leftarrow \frac{\nabla f(\mathbf{x}_t)^\top(\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1}))}{\nabla f(\mathbf{x}_{t-1})^\top \nabla f(\mathbf{x}_{t-1})}$

$\mathbf{d}_t^{CG} \leftarrow \left(\mathbf{I} - \mathbf{D}_t\mathbf{G}_t^\top\left(\mathbf{G}_t\mathbf{G}_t^\top\right)^{-1}\mathbf{G}_t\mathbf{D}_t^\top\right)\left(-\nabla f\left(\mathbf{x}_t\right) + \beta_t \mathbf{d}_{t-1}\right)$

$\mathbf{d}_t \leftarrow \mathbf{d}_t^{QN} + \mathbf{d}_t^{CG}$

$\alpha_t \leftarrow \operatorname{argmin} f(\mathbf{x}_t + \alpha \mathbf{d}_t)$

$\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t + \alpha_t \mathbf{d}_t$

$t \leftarrow t + 1$; переход к шагу 1

все предыдущие сопряженные направления ортогональны текущему значению градиента $\nabla f\left(\mathbf{x}_j\right)^\top \mathbf{d}_i^{CG} = 0 \; \forall \; 0 \leqslant i < j$, получим $\mathbf{d}_t^{QN} = 0$, так как

$$\begin{aligned}
&\mathbf{Z}_t\mathbf{D}_t^\top \nabla f\left(\mathbf{x}_t\right) = \\
&\quad = \left[(\mathbf{x}_t - \mathbf{x}_{t-1})\mathbf{D}_t\mathbf{D}_t^\top \nabla f\left(\mathbf{x}_t\right), \ldots, (\mathbf{x}_t - \mathbf{x}_{t-T})\mathbf{D}_t\mathbf{D}_t^\top \nabla f\left(\mathbf{x}_t\right)\right] \\
&\quad = \left[\mathbf{d}_{t-1}^\top \nabla f\left(\mathbf{x}_t\right), \ldots, \mathbf{d}_{t-T} \nabla f\left(\mathbf{x}_t\right)\right] = \mathbf{0}.
\end{aligned}$$

Далее, по свойству сопряженных направлений $\mathbf{d}_i^{CG\,\top} \nabla^2 f \, \mathbf{d}_j^{CG} = 0, \forall \; 0 \neq j$:

$$\begin{aligned}
&\mathbf{G}_t\mathbf{D}_t^\top\left(\nabla f\left(\mathbf{x}_t\right) + \beta_t \mathbf{d}_{t-1}\right) = \\
&\qquad = \sum_{j=0}^{t-1} c_j \mathbf{d}_j^{CG\,\top} \nabla^2 f \; \left(\nabla f\left(\mathbf{x}_t\right) + \beta_t \mathbf{d}_{t-1}\right) = 0.
\end{aligned}$$

Для квадратичного случая направления, сгенерированные Алгоритмом 4, совпадают с направлениями метода сопряженных градиентов, а следовательно обладают теми же свойствами, в том числе сходимостью не более чем за n итераций с линейной скоростью согласно Утверждению 3. □

Таким образом, для Алгоритма 4 показана конечная сходимость в квадратичном случае. В матрицах $\mathbf{G}_t$ и $\mathbf{Z}_t$ накапливается информация о кривизне функции, за счет чего делается квазиньютоновский шаг вдоль $\mathcal{G}_t =$

$\operatorname{span}\{\nabla f_t - \nabla f_{t-1}, \ldots, \nabla f_t - \nabla f_{t-T})\}$ — линейной оболочки разности градиентов. Дополнительно, выполняется шаг методом сопряженного градиента в направлении, ортогональном $\mathcal{G}_t$. Таким образом, шаг в квазиньютоновском направлении *может* нивелировать накопленную методом сопряженных градиентов ошибку, содержащуюся в $\mathcal{G}_t$. Отметим, что для обеспечения линейной сходимости в сильно выпуклом случае можно воспользоваться корректирующим алгоритмом подпространственной оптимизации 3

2.4 Метод модифицированных знако–возмущенных сумм

Предложенный в работе [41] метод знако–возмущённых сумм даёт точное доверительное множество при условии независимых помех, распределённых одинаково и симметрично относительно нуля. Данное ограничение делает метод ЗВС неприменимым при поставленных в Разделе 1.3.3 условиях независимых с входами а в остальном произвольных помех и симметрично распределённых относительно известного среднего входах. В этом разделе изложен метод *модифицированных знако-возмущённых сумм* (МВЗС), вместо накладывания ограничения на природу помех, основывающийся на симметричности распределения входов системы вокруг известного среднего. Стоит отметить, что схожие идеи используются и в некоторых других рандомизированных методах, работающих при почти произвольных помехах (смотри [36]).

Рассмотрим задачу оценки параметра линейной модели входа–выхода с аддитивными помехами в наблюдениях выхода аналогиную модели (1.14):

$$y_i = \boldsymbol{\varphi}_i^T \mathbf{x}_\star + \varepsilon_i, \quad i = 1..N,$$

где N — число наблюдений, $y_i \in \mathbb{R}$ — наблюдаемые выходы системы, $\boldsymbol{\varphi}_i \in \mathbb{R}^n$ — наблюдаемые входы системы, $\mathbf{x}_\star \in \mathbb{R}^n$ — искомый вектор параметров, а $\varepsilon_i \in \mathbb{R}$ — неизвестные аддитивные помехи. Предположим, что вектора входов $\boldsymbol{\varphi}_i$ независимы и одинаково распределены вокруг известного среднего $m_\varphi \in \mathbb{R}^n$. Введем

обозначения:

$$S_k(\mathbf{x}) = \sum_{i=1}^{N} a_{i,k}\Delta_i(y_i - \boldsymbol{\varphi}_i\mathbf{x}), \tag{2.38}$$

$$S_0(\mathbf{x}) = \sum_{i=1}^{N} \Delta_i(y_i - \boldsymbol{\varphi}_i\mathbf{x}), \tag{2.39}$$

где $\Delta_i = \boldsymbol{\varphi}_i - m_\varphi$ — центрированные значения вектров входов, $a_{i,k}$ — так называемые *случайные знаки*, определяемые по формуле:

$$a_{k,i} = \begin{cases} 1 & \text{с вероятностью } \frac{1}{2}, \\ -1 & \text{с вероятностью } \frac{1}{2}. \end{cases}$$

По аналогии с *знако-возмущенными суммами*, введенными в работе [41], $S_k(\mathbf{x})$ именуются *модифицированными знако-возмущенными суммами*. Модификация заключается в введении *рандомизированных* множителей Δ_i.

В Алгоритме 5 в виде псевдокода изложен основанный на суммах (2.38) и (2.39) метод модифицированных знако-возмущенных сумм построения точного доверительного множества параметра линейной модели (1.14) $\mathbf{x}^\star$ в контексте сделанных в Разделе 1.3.3 предположений. Основная идея приведенного Алгоритма 5 заключается в том, что из независимости и симметричности распределения Δ_i следует совпадение распределений сумм $S_k(\mathbf{x}^\star) = \sum_{i=1}^{N} a_{k,i}\Delta_i\varepsilon_i$, $k = 0, \ldots, M-1$. Таким образом, обозначив $Z_k = ||S_k(\mathbf{x}^\star)||$, случайная величина Z_0 примет любую позицию в отсортированном по возрастанию списке $Z_{(0)}, \ldots, Z_{(M-1)}$ с вероятностью $\frac{1}{M}$. Следовательно, Z_0 *не* окажется среди $q \leqslant M$ наибольших значений $\{Z_k\}_{k=0}^{M-1}$ с вероятностью $1 - \frac{q}{M}$.

Формально обоснуем приведенные рассуждения. Следующая теорема утверждает, что множество значений $\mathbf{x}$, для которых Алгоритм 5 возвращает $TRUE$, формируют точное $(1 - \frac{q}{M})$-доверительное множество для параметра $\mathbf{x}^\star$.

***Теорема* 4.** *Рассмотрим линейную модель с аддитивными помехами* (1.14). *Пусть входы* $\{\boldsymbol{\varphi}_i\}_1^N$ *и помехи* $\{\varepsilon_i\}_1^N$ *независимы как между собой так и друг с другом. Предположим, что входы* $\{\boldsymbol{\varphi}_i\}_1^N$ *одинаково и симметрично распределены относительно известного вектора* $m_\varphi \in \mathbb{R}^n$. *Зафиксируем* $M \geqslant q \geqslant 0$ — *два натуральных числа,* $\{S_k(\mathbf{x})\}_{k=0}^{M-1}$ — M *модифицированных знако-возмущённых суммы, полученных по формулам* (2.38) *и* (2.39). *Обозначим* $Z_k(\mathbf{x}) = ||S_k(\mathbf{x})||_2^2$ *и*

$$\mathcal{X}_{M,q} = \{\mathbf{x} \in \mathbb{R}^p : |\{k \in \{1, \ldots, M-1\} : Z_0(\mathbf{x}) < Z_k(\mathbf{x})\}| \geqslant 1\}. \tag{2.40}$$

Algorithm 5 MSPS($\{y_i\}_{i=1}^N$, $\{\varphi\}_{i=1}^N$, m_φ, M, q, $\mathbf{x}$)

for i in $1 \dots N$ **do**
  $n_i \leftarrow y_i - \varphi_i^T \cdot \mathbf{x}$
  $\Delta_i \leftarrow \varphi_i - m_\varphi$
for k in $1 \dots M-1$ **do**
  $S_k \leftarrow 0 \in \mathbb{R}^p$
  for i in $1 \dots N$ **do**
    $S_k \leftarrow S_k + \Delta_i \cdot n_i \cdot RANDOMSIGN()$
  $Z_k \leftarrow ||S_k||_2^2$
$S_0 \leftarrow 0 \in \mathbb{R}^p$
for i in $1 \dots N$ **do**
  $S_0 \leftarrow S_0 + \Delta_i \cdot n_i$
$Z_0 \leftarrow ||S_0||_2^2$
$r \leftarrow 0$
for k in $1 \dots M-1$ **do**
  if $Z_0 > Z_k$ **then**
    $r \leftarrow r + 1$
if $r >= q$ **then**
  return $FALSE$
else
  return $TRUE$
end procedure

Тогда $\mathcal{X}_{M,q}$ будет точным $(1 - \frac{q}{M})$-доверительным множеством для параметра $\mathbf{x}^\star$:

$$P_{\varepsilon_i, a_{k,i}} (\mathbf{x}^\star \in \mathcal{X}_{M,q}) = 1 - \frac{q}{M}.$$

Доказательство. Достаточно показать, что $\mathbf{x}^\star \in \mathcal{X}_{M,q}$ с вероятностью $1 - \frac{q}{M}$. Это эквивалентно тому, что $Z_0(\mathbf{x}^\star)$ меньше как минимум q различных $Z_k(\mathbf{x}^\star)$ из $\{Z_k(\mathbf{x}^\star)\}_{k=1}^{M-1}$ с вероятностью $1 - \frac{q}{M}$, что в свою очередь выполняется при условии равномерной упорядоченности $\{Z_k(\mathbf{x}^\star)\}_{k=0}^{M-1}$. Заметим, что $Z_0(\mathbf{x}^\star)$ и $Z_k(\mathbf{x}^\star)$ могут быть записаны в виде:

$$Z_k(\mathbf{x}^\star) = || \sum_{i=1}^N a_{k,i} \Delta_i \varepsilon_i ||, \quad Z_0(\mathbf{x}^\star) = || \sum_{i=1}^N \Delta_i \varepsilon_i ||.$$

Согласно Утверждению 5 $a_{k,i}\Delta_i$ и Δ_i независимо и одинаково распределены $\forall\, i = 1..N$, $k = 0..M-1$. Следовательно, независимо и одинаково распределены величины $\{Z_k(\mathbf{x}^\star)\}_{k=0}^{M-1}$. Кроме того, они распределены непрерывно, а значит и равномерно упорядоченны в силу непрерывности распределения Δ_i и Утверждения 6. Таким образом, $Z_0(\mathbf{x}^\star)$ не среди q наибольших значений из $\{Z_k(\mathbf{x}^\star)\}_{k=1}^{M-1}$ с вероятностью $1 - \frac{q}{M}$. □

Следующее замечание характеризует область применимости Алгоритма 5.

***Замечание* 15.** *Доказательство Теоремы 4 в существенной степени основывается на случайной природе векторов входов $\{\boldsymbol{\varphi}_i\}_1^N$: вероятность принадлежности истинного значения параметра $\mathbf{x}^\star$ множеству $\mathcal{X}_{M,q}$ в Теореме 4 понимается как по случайным знакам $\{\alpha_{k,i}\}$, так и по векторам входов $\{\boldsymbol{\varphi}_i\}_1^N$. Таким образом, к области применения Алгоритма 5 относятся ситуации в которых входы модели генерируются случайным образом, их распределение симметрично, а их среднее известно — например, задаются экспериментатором.*

Приведенные в данном разделе Теорема 4 и Алгоритм 5 основаны на методе знако-возмущённых сумм и соответствующей теореме, приведенныхв статье [41]. Тем не менее, изложенный метод модифицированных знако-возмущённых использует отличную идею рандомизации входов, что позволяет расширить область применения алгоритма, ослабив ограничения на природу помех.

Стоит отметить, что множество $\mathcal{X}_{M,q}$, полученное методом знако–возмущённых сумм, является решением поставленной в Разделе 1.3.3 задачи при $\alpha = 1 - \frac{q}{M}$.

2.5 Свойства доверительного множества метода МЗВС

В этом разделе продемонстрированны некоторые свойства доверительного множества $\mathcal{X}_{M,q}$, полученного Алгоритмом 5.

Выражение доверительного множества в явном виде по формуле (2.40) неудобно, так как требует применения приближенно-численных процедур. Кроме того, отсутствие аналитического выражения границ доверительного множества $\mathcal{X}_{M,q}$ затрудняет возможность исследования их теоретических свойств.

***Замечание* 16.** *В условиях Теоремы 4, множество $\mathcal{X}_{M,q}$, определяемое по формуле (2.40), допускает альтернативное выражение:*

$$\mathcal{X}_{M,q} = \bigcup_{\mathbb{I}\subset\{1,..,M-1\}:|\mathbb{I}|=q} \left(\bigcap_{k\in\mathbb{I}} \{\mathbf{x} : Z_0(\mathbf{x}) < Z_k(\mathbf{x})\} \right) \neq \varnothing.$$

Доказательство. Достаточно заметить, что множество $\mathcal{X}_{M,q}$ состоит из точек $\mathbf{x}$, для которых $Z_0(\mathbf{x})$ меньше как минимум q различных $Z_k(\mathbf{x})$, т.е. по меньшей мере q различных неравенств $Z_0(\mathbf{x}) < Z_k(\mathbf{x})$, $k = 1..M-1$ выполняются. □

Важность Замечания 16 состоит в том, что вычисление множества $\mathcal{X}_{M,q}$ может быть сведено к вычислению более простых множеств $\{\mathbf{x} : Z_0(\mathbf{x}) < Z_k(\mathbf{x})\}$ в многомерном случае.

Рассмотрим одномерный случай линейной модели (1.14)

$$y_i = \varphi_i x + \varepsilon, \qquad i = 1,\ldots,N, \tag{2.41}$$

где $y_i, \varphi_i, x, \varepsilon_i \in \mathbb{R}\ \forall\ i = 1,\ldots,N$. В следующей лемме обосновывается аналитическая форма множеств $\{\mathbf{x} : Z_0(\mathbf{x}) < Z_k(\mathbf{x})\}$ для одномерного случая.

***Лемма* 7.** *Пусть $d = 1$. Тогда множество $\{x : Z_0(x) < Z_k(x)\}$ допускает следующее аналитическое выражение:*

$$\{x : Z_0(x) < Z_k(x)\} = (B_k^{min}, B_k^{max}), \quad \textit{где}$$

$$B_k^{(1)} = \frac{\sum_{i=1}^N (1 - a_{k,i})\varphi_i y_i}{\sum_{i=1}^N (1 - a_{k,i})\varphi_i^2}, \qquad B_k^{(2)} = \frac{\sum_{i=1}^N (1 + a_{k,i})\varphi_i y_i}{\sum_{i=1}^N (1 + a_{k,i})\varphi_i^2},$$

$$B_k^{min} = \min(B_k^{(1)}, B_k^{(2)}), \qquad B_k^{max} = \max\left(B_k^{(1)}, B_k^{(2)}\right).$$

Доказательство. Рассмотрим неравенство $Z_0(x^\star) < Z_k(x^\star)$

$$(\sum_{i=1}^{N} \Delta_i(y_i - \varphi_i x))^2 < (\sum_{i=1}^{N} a_{k,i}\Delta_i(y_i - \varphi_i x))^2$$

$$(\sum_{i=1}^{N} \Delta_i(y_i - \varphi_i x))^2 - (\sum_{i=1}^{N} a_{k,i}\varphi_i(y_i - \varphi_i x))^2 < 0,$$

$$(\sum_{i=1}^{N}(1 - a_{k,i})\Delta_i(y_i - \varphi_i x))(\sum_{i=1}^{N}(1 + a_{k,i})\Delta_i(y_i - \varphi_i x)) < 0,$$

$$\left(\frac{\sum_{i=1}^{N}(1 - a_{k,i})\Delta_i(\varepsilon_i + \Delta_i x^\star)}{\sum_{i=1}^{N}(1 - a_{k,i})\Delta_i^2} - x\right) \cdot$$

$$\cdot \left(\frac{\sum_{i=1}^{N}(1 + a_{k,i})\Delta_i(\varepsilon_i + \Delta_i x^\star)}{\sum_{i=1}^{N}(1 + a_{k,i})\Delta_i^2} - x\right) < 0.$$

□

Стоит отметить, что аналитическое выражение границ множеств $\{x : Z_0(x) < Z_k(x)\}$, полученное в Лемме (7) значительно эффективнее поточечного вычисления с помощью Алгоритма 5.

***Лемма* 8.** *Пусть B_k является $B_k^{(1)}$ или $B_k^{(2)}$ и в дополнении к предположениям, изложенным в 1.3.3 выполняются следующие условия:*

- *$\exists$ второй и четвертый моменты для $\{\varphi_i\}$;*
- *$\{\varepsilon_i\}$ является либо случайной величиной с вторым нецентрированным моментом, ограниченным константой $C < \infty$, либо детерминированной последовательностью, равномерно ограниченной константой $\sqrt{C}$.*

Тогда

$$E_{\varepsilon,\varphi,a}[B_k] = x^\star,$$

$$E_{\varepsilon,\varphi,a}[B_k - E_{\varepsilon,\varphi,a}[B_k]]^2 \xrightarrow[N\to\infty]{} 0,$$

где $E_{\varepsilon,\varphi,a}$ — математическое ожидание совместного распределения $\{\varepsilon_i\}_{i=1}^N, \{\varphi_i\}_{i=1}^N$ и $\{a_i\}_{i=1}^N$.

Доказательство. Рассмотрим математическое ожидание границы множества

$$E_{\varepsilon,\varphi,a}B_{\{a_{k,i}\}} = E_{\varepsilon,\varphi,a}\left[\frac{\sum_{i=1}^{N}(1 \pm a_{k,i})\Delta_i(\varepsilon_i + \Delta_i x^\star)}{\sum_{i=1}^{N}(1 \pm a_{k,i})\Delta_i^2}\right] =$$

$$= x^\star + E_{\varepsilon,\varphi,a}\left[\frac{\sum_{i=1}^N (1 \pm a_{k,i})\Delta_i \varepsilon_i}{\sum_{i=1}^N (1 \pm a_{k,i})\Delta_i^2}\right].$$

Так как Δ_i независимы с ε_i и среднее первых равно нулю, то

$$E_{\varepsilon,\varphi,a}\left[\frac{\sum_{i=1}^N (1 \pm a_{k,i})\Delta_i \varepsilon_i}{\sum_{i=1}^N (1 \pm a_{k,i})\Delta_i^2}\right] = \sum_{i=1}^N |E_{\varepsilon,\varphi,a}\left[\frac{(1 \pm a_{k,i})\Delta_i \varepsilon_i}{\sum_{j=1}^N (1 \pm a_{k,j})\Delta_j^2}\right]| \leqslant$$
$$\leqslant \sum_{i=1}^N |E_{\varepsilon,\varphi,a}\left[\frac{(1 \pm a_{k,i})\Delta_i \varepsilon_i}{\sum_{j\neq i}^N (1 \pm a_{k,j})\Delta_j^2}\right]| = 0.$$

Таким образом первая часть утверждения доказана. Для доказательства второй части заметим, что $\forall \{x_i\}_1^N$ верно неравенство $2\sum_1^N x_i^2 \geqslant (\sum_1^N x_i)^2$. Следовательно,

$$E_{\varepsilon,\varphi,a}[B_{\{a_{k,i}\}} - E_{\varepsilon,\varphi,a}B_{\{a_{k,i}\}}]^2 \leqslant E_{\varepsilon,\varphi,a}\left[\frac{(\sum_{i=1}^N (1 \pm a_{k,i})\Delta_i \varepsilon_i)^2}{(\sum_{i=1}^N (1 \pm a_{k,i})\Delta_i^2)^2}\right] \leqslant$$

$$\leqslant C E_{\varepsilon,\varphi,a}\left[\frac{(\sum_{i=1}^N (1 \pm a_{k,i})\Delta_i \varepsilon_i)^2}{(\sum_{i=1}^N (1 \pm a_{k,i})\Delta_i)^4}\right] \xrightarrow[N\to\infty]{} 0.$$

Так как ε_i независимы с Δ_i, математическое ожидание по ним может быть взять отдельно. Таким образом, используя наложенные ограничения:

$$E_{\varepsilon,\varphi,a}[B_{\{a_{k,i}\}} - E_{\varepsilon,\varphi,a}B_{\{a_{k,i}\}}]^2 \leqslant C E_{\varepsilon,\varphi,a}\left[\frac{(\sum_{i=1}^N (1 \pm a_{k,i})\Delta_i)^2}{(\sum_{i=1}^N (1 \pm a_{k,i})\Delta_i)^4}\right] =$$

$$= C E_{\varepsilon,\varphi,a}\left[\frac{1}{(\sum_{i=1}^N (1 \pm a_{k,i})\Delta_i)^2}\right] \xrightarrow[N\to\infty]{} 0.$$

Следовательно, верно и второе утверждение. □

Лемма 8 даёт два важных следствия. Во-первых, B_k — границы интервала $\mathcal{X}_{M,q}$ — в среднем равны истинному значению параметра $x^\star$. Во-вторых, эти границы стремятся к $x^\star$ с увеличением N. Резюмируя, доверительный интервал сходится $x^\star$ при $N \to \infty$. Таким образом, по аналогии с оценками параметров, доверительный интервал $\mathcal{X}_{M,q}$ можно назвать *состоятельным*.

Глава 3. Сравнительный анализ методов оптимизации и оценивания параметров

3.1 Анализ метода МЗВС на модельных данных

В этом разделе демонстрируются особенности доверительных множеств, полученных методом модифицированных знако–возмущённых сумм в сравнении с методом знако-возмущённых сумм и ассимптотическими доверительными множествами, полученным по формуле (1.15). В силу того, что перечисленные методы имеют разные области применимости: нормально распределённые помехи, центрированные симметрично рапределённые помехи и произвольные помехи, — результаты сравнения в каждом из указанных случаев предсказуемы. Таким образом, привёденные далее примеры носят исключительно иллюстративный характер.

3.1.1 Описание модельных данных

В привёденных в Разделе 1.3.3 условиях, наибольший интерес представляет случай малого числа имзерений. Для удобства визуализации рассмотрим двумерный случай $n = 2$. Положим $\theta^\star = (-1, 2)^T$ и смоделируем значения $\boldsymbol{\varphi}_i$ согласно следующему распределению:

$$\varphi_i \sim N(\mu_\varphi, \Sigma_\varphi), \quad \mu_\varphi = (1, -1)^T, \quad \Sigma_\varphi = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{pmatrix},$$

Таким образом, неизвестными остаются число наблюдений N и природа помех $\{\varepsilon_i\}$, различные комбинации которых будут рассмотрены в последующих подразделах.

3.1.2 Случай большого числа измерений

Рассмотрим два типа помех:

1. *несмещенные симметричные*: $\varepsilon_i \sim N(0,1)$,
2. *смещённые ассиметричные*: $\varepsilon_i \sim Exp(\lambda = 1) + 4$.

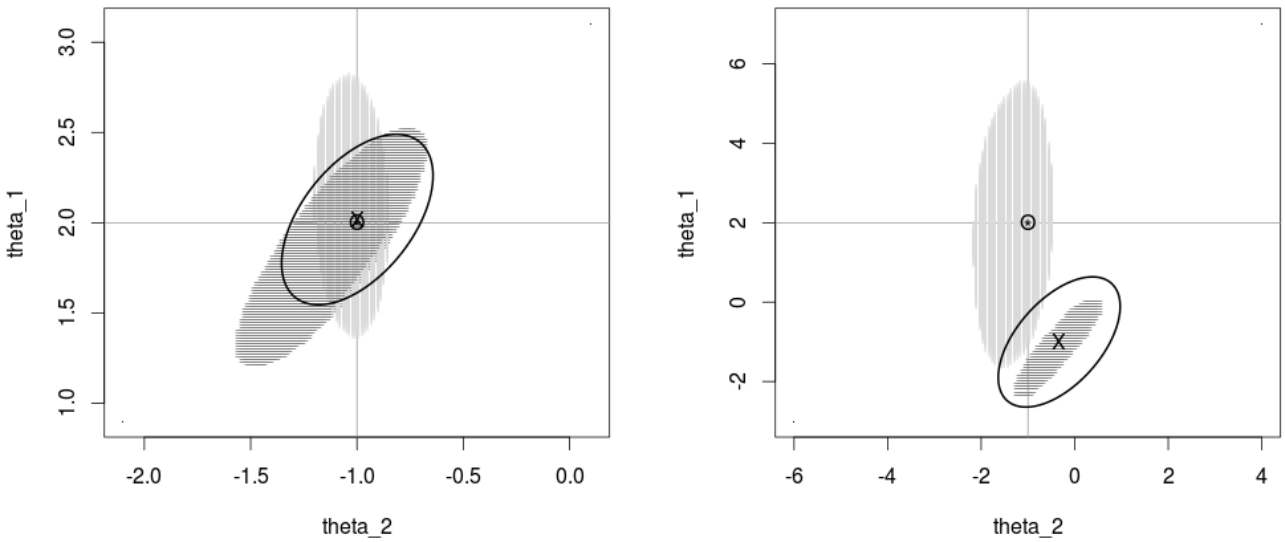
На Рисунке 3.1 (a) приведены доверительные множества для случая числа наблюдений $N = 50$ и несмещённых симметричных помех. Все три доверительных множества содержат как истинное значение параметра, так и его оценку, полученную методом наименьших квадратов. Отметим, что методы МЗВС и ЗВЗ дают доверительные множества значительно большие по объем, чем множество, полученное по формуле (1.15). На Рисунке 3.1 (b) приведены доверительные множества для случая числа наблюдений $N = 50$ и смещённых асимметричных помех. Несмотря на то, что доверительное множество, полученное методом МЗВС, значительно превосходит по объему доверительные множества, полученные другими методами, оно единственное содержит истинное значение параметра. Доверительные множества, полученные методом ЗВС и по формуле (1.15), смещены вместе с оценкой МНК из-за асимметричности и смещенности помех.

3.1.3 Случай малого числа измерений

Рассмотрим два типа помех:

1. *несмещённые асимметричные* $\varepsilon_i \sim Exp(\lambda = 1) - 1$,
2. *смещённые асимметричные* $\varepsilon_i \sim Exp(\lambda = 1) + 4$.

На Рисунке 3.2 (a) приведены доверительные множества для случая числа наблюдений $N = 15$ и несмещённых асимметричных помех. На Рисунке 3.2 (b) приведены доверительные множества для случая числа наблюдений $N = 15$ и смещённых асимметричных помех. Доверительное множество метода МЗВС — единственное, содержащие истинное значение параметра. При этом, доверительное множество, полученное методом ЗВС вновь неограничено. Примечательно, что полученные методом ЗВС доверительные множества вырождаются в неограниченное.

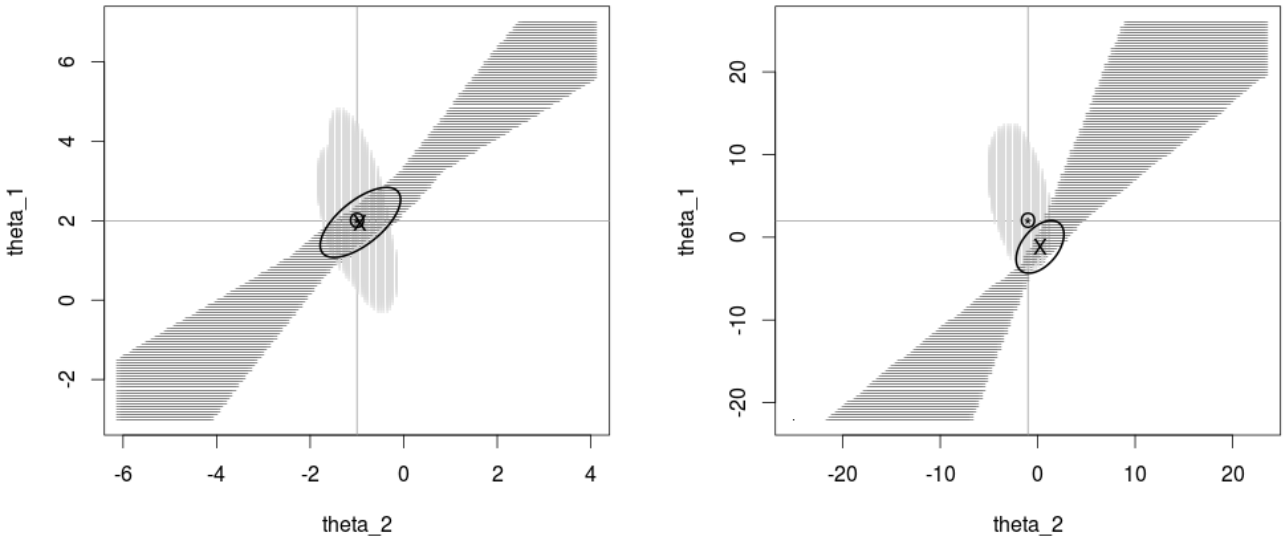


а) Несмещенные симметричные помехи б) Смещенные асимметричные помехи
Истинное значение параметра отмечено кругом с точкой, оценка МНК отмечена крестом

Рисунок 3.1 — Доверительные множества, полученные методом ЗВС (темные горизонтальные линии), методом МЗВС (светлые вертикальные линии) и по формуле (1.15) (эллипсоид) при большом количестве наблюдений.

3.1.4 Выводы

В предыдущих разделах проиллюстрированны доверительные множества, полученные методоами ЗВС, МЗВС и по формуле (1.15) для четырёх случаев: несмещённые симметричные помехи при большом числе наблюдений, смещённые асимметричные помехи при малом числе наблюдений, несмещённые асимметричные помехи при большом числе наблюдений и смещённые асимметричные помехи при малом числе наблюдений. Приведённые примеры согласуются с полученными в Разделе 2.4 теоретическими результатами: метод модифицированных знако–возмущённых сумм даёт точное доверительное множество в том числе в условиях смещённых асимметричных помех и малого числа измерений.



а) Несмещенные асимметричные помехи

б) Смещенные асимметричные помехи

Истинное значение параметра отмечено кругом с точкой, оценка МНК отмечена крестом

Рисунок 3.2 — Доверительные множества, полученные методом ЗВС (темные горизонтальные линии), методом МЗВС (светлые вертикальные линии) и по формуле (1.15) (эллипсоид) при малом количестве наблюдений.

3.2 Сравнительный анализ методов оптимизации

В этом разделе приведено сравнение методов последовательной подпространственной оптимизации и распространенных аналогов. Рассматриваются следующие методы оптимизации:

- CG: метод сопряженных градиентов (1.10) с выбором шага по формуле Полака–Рибьере–Поляка (1.11);
- BPCG: метод Била–Пауэлла (1.12), в качестве параметров c_1, c_2, c_3 использовались значения 0.2, 0.8 и 1.2 соответственно;
- L-BFGS(m): метод Бройдена–Флетчера–Гольдфарба–Шэнно, хранящий историю значений аргумента и градиента за последние m шагов (смотри Алгоритм 7.5 в [70]), в качестве начального приближения обратной матрицы Гессе использовалась единичная матрица;

– L-QNSSO(m): предложенный Алгоритм 4, хранящий историю значений аргумента и градиента за последние m шагов.

Для определения размера шага каждого из методов использовался приближенный линейный поиск методом Брента [108]. Все методы реализованы на языке программирования Python с использованием библиотек линейной алгебры и научных расчётов NumPy [109] и SciPy [110]. Исходный код выложен в публичный доступ на сервисе GitHub, адрес репозитория: https://github.com/obus/optimus.

3.2.1 Квадратичная функция

Рассмотрим функцию вида

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} + \mathbf{b}^\top \mathbf{x}, \tag{3.1}$$

где $\mathbf{x} \in \mathbb{R}^n$, $n = 1000$, $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{b} = \mathbf{1} \in \mathbb{R}^n$. При этом, спектр матрицы $\mathbf{A}$ имеет следующий вид:

$$\rho(\mathbf{A}) = \left\{1 + \frac{i-1}{n-1}(\lambda_{max} - 1)\right\}_{i=1}^{n} \subset [1, \lambda_{max}]\,. \tag{3.2}$$

Сравнение методов оптимизации проведено следующим образом:

1. генерируется начальная точка $\mathbf{x}_0$ из распределения $\mathcal{N}(\mathbf{x}_\star, \mathbf{I})$;
2. для каждого алгоритма запускается процесс оптимизации, и считается минимальное значение функции $f(\mathbf{x}_t)$ за тысячу итераций;
3. этапы 1-2 повторяется тысячу раз, получив таким образом выборку $\min\limits_{0\leqslant t\leqslant n} f(\mathbf{x}_t)$ для каждого из алгоритмов;
4. этапы 1-3 повторяются для значений λ_{max} = 10, 100, 1000.

Распределения полученных выборок проиллюстрированы посредством “ящиков с усами” на Рисунках 3.3, 3.4 и 3.5 для случаев λ_{max} = 10, 100 и 1000 соответственно. Во всех трёх случаях алгоритм L-QNSSO(2) демонстрирует наименьшую ошибку. Заметим, что с ростом λ_{max} увеличивается превосходство алгоритма L-QNSSO(2) над методом сопряженных градиентов. Это можно объяснить тем, что плохая обусловленность задачи увеличивает накапливаемую методом сопряженных градиентов ошибку и лишает соответствующие направления свойств сопряженности.

Метод L-QNSSO нивелируют эту погрешность за счет дополнительного квазиньютоновского шага.

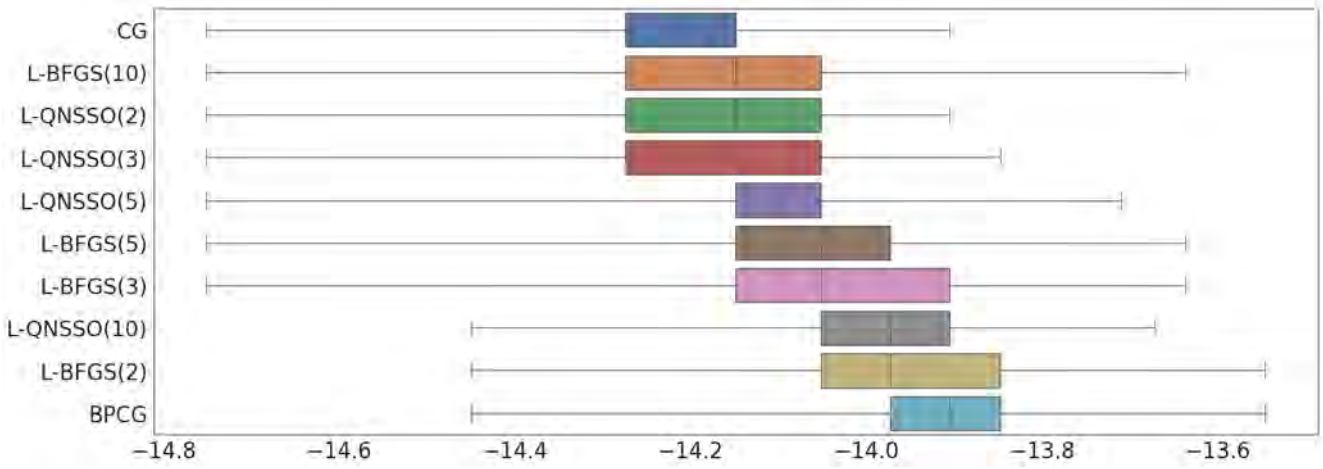


Рисунок 3.3 — Ящики с усами статистики $\min\limits_{0\leqslant t\leqslant n} f(\mathbf{x}_t)$ для квадратичной функции (3.1) с $\Lambda_{max} = 10$. По оси абсцисс отмечено значение функции в логарифмической шкале. Левая граница, отрезок внутри и правая граница ящика — 25%, 50% и 75% персентили соответственно, границы усов соответствуют минимальным и максимальным значениям.

3.2.2 Функция Розенброка

Рассмотрим обобщение предложенной в работе [111] функции Розенброка на случай n переменных:

$$f(\mathbf{x}) = \sum_{i=1}^{n-1} \left[100(x_{i+1} - x_i^2)^2 + (1 - x_i)^2\right], \quad -2.048 \leqslant x_i \leqslant 2.048 \quad \forall\, i = 1..n. \tag{3.3}$$

Глобальный минимум функции (3.3) находится в точке $\mathbf{x}_\star = \mathbf{1}\ \forall n \geqslant 2$, а для $n \geqslant 4$ известно о существовании по меньшей мере одного локального минимума в окрестности точки $(-1, 1, \ldots, 1)$ [112]. При этом, предполагается, что седловая точка находится в окрестности точки $(-0.555, 0.322, 0.115, 0.024, 0.011, 0.010, \ldots, 0.010, 0.0001)^\top$ [113]. В связи с этим, сравнение методов оптимизации проведено следующим образом:

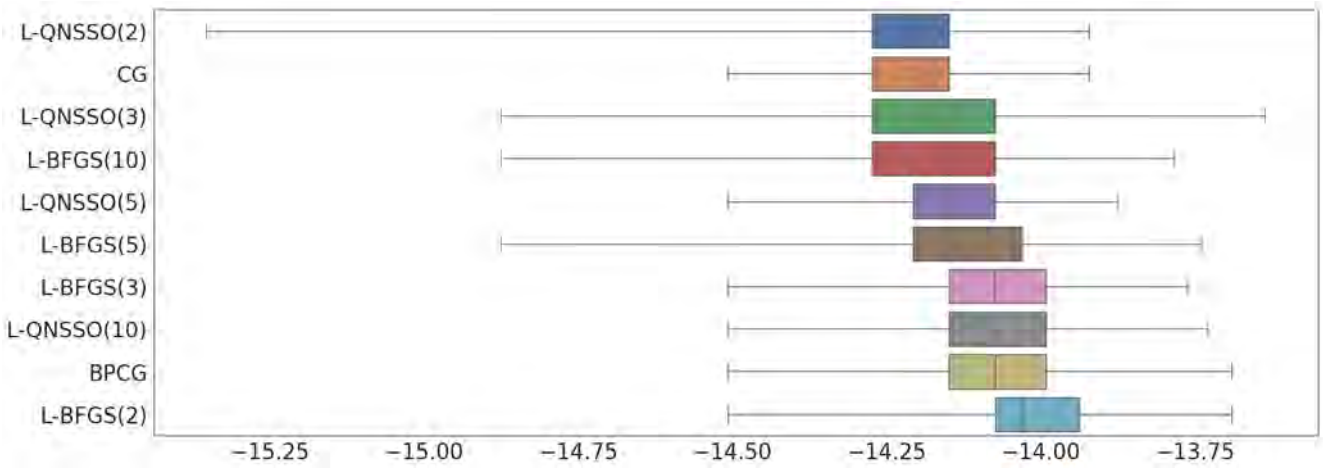


Рисунок 3.4 — Ящики с усами статистики $\min\limits_{0 \leqslant t \leqslant n} f(\mathbf{x}_t)$ для квадратичной функции (3.1) с $\lambda_{max} = 100$. По оси абсцисс отмечено значение функции в логарифмической шкале. Левая граница, отрезок внутри и правая граница ящика — 25%, 50% и 75% персентили соответственно, границы усов соответствуют минимальным и максимальным значениям.

1. генерируется начальная точка $\mathbf{x}_0$ согласно распределению

$$\begin{aligned} \mathbf{x}_0 &= (-0.555, |x_0^{(2)}|, \ldots, |x_0^{(n)}|)^\top, \\ x_0^{(i)} &\sim \mathcal{N}(0,1), \qquad \forall\, i = 2..n; \end{aligned} \tag{3.4}$$

2. для каждого алгоритма запускается процесс оптимизации и считается минимальное значение функции $f(\mathbf{x}_t)$ за T итераций;
3. этапы 1-2 повторяются тысячу раз, в результате чего получим выборку значений $\min\limits_{0 \leqslant t \leqslant T} f(\mathbf{x}_t)$ для каждого из алгоритмов.

Распределения полученных выборок проиллюстрированы посредством “ящиков с усами” на Рисунках 3.6 и 3.7 для числа итераций T равному 50 и n соответственно. Алгоритм L-QNSSO(2) зачастую достигает наилучшего качества на первых пятидесяти итерациях, но в последствии L-QNSSO уступает методам L-BFGS и BPCG. Тем не менее, метод L-QNSSO демонстрирует результаты превосходящие результаты метод сопряженных градиентов.

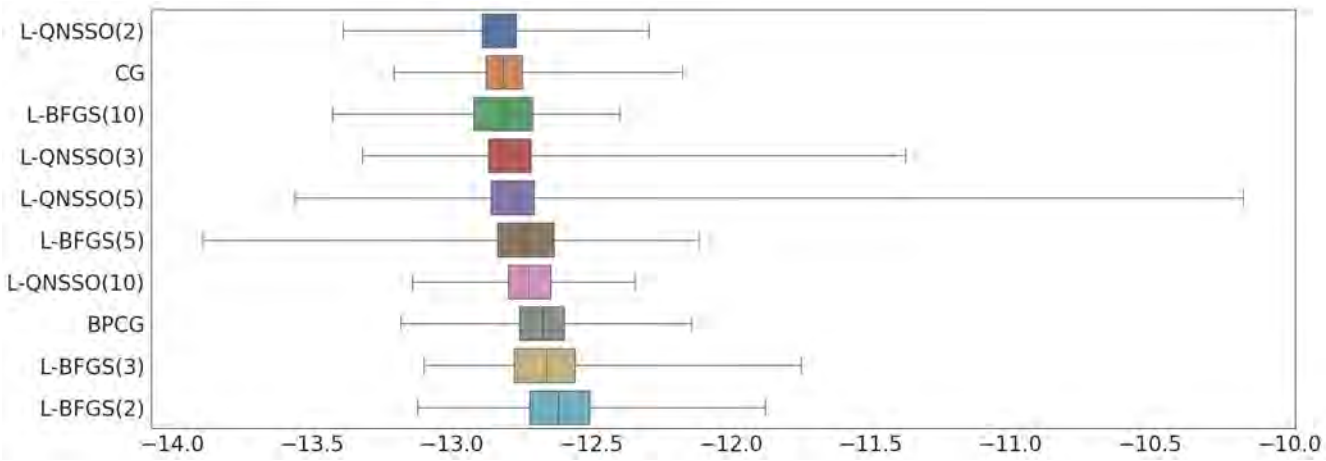


Рисунок 3.5 — Ящики с усами статистики $\min\limits_{0\leqslant t\leqslant n} f(\mathbf{x}_t)$ для квадратичной функции (3.1) с $\lambda_{max} = 1000$. По оси абсцисс отмечено значение функции в логарифмической шкале. Левая граница, отрезок внутри и правая граница ящика — 25%, 50% и 75% персентили соответственно, границы усов соответствуют минимальным и максимальным значениям.

3.2.3 Линейная регрессия с регуляризацией по Тихонову

Рассмотрим линейную модель входа–выхода с аддитивными помехами аналогичную модели (1.14):

$$y_i = \boldsymbol{\varphi}_i^\top \mathbf{x}_\star + \varepsilon_i, \quad i = 1 \ldots N, \tag{3.5}$$

где $y_i \in \mathbb{R}$ — выход, $\boldsymbol{\varphi}_i \in \mathbb{R}^n$ — вектора входа, $\mathbf{x} \in \mathbb{R}^n$ — искомый вектор параметров, а помехи распределены по нормальному закону: $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)\ \forall, i = 1 \ldots N$. Положим $n = 10^6$, $N = 10$. Отметим, что модель вида (3.5) актуальна для многих практических задач, в том числе в рекомендательных системах [45].

В подобных случаях задача поиска параметра $\mathbf{x}_\star$ зачастую формулируется как задача линейной регрессии с регуляризацией по Тихонову, также известной как гребневая регрессия [60] — частный случай задачи распознавания образов (смотри Раздел 1.1):

$$f(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \boldsymbol{\varphi}_i^\top \mathbf{x})^2 + \lambda\|\mathbf{x}\|^2 \to \min_{\mathbf{x}}, \tag{3.6}$$

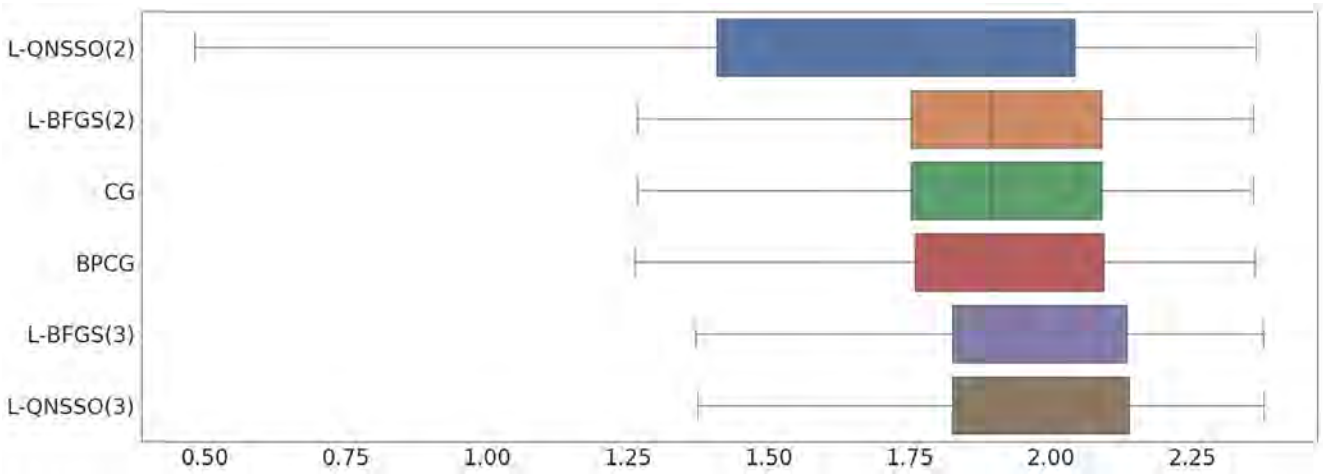


Рисунок 3.6 — Ящики с усами статистики $\min\limits_{0\leqslant t\leqslant 50} f(\mathbf{x}_t)$ для функции Розенброка (3.3). По оси абсцисс отмечено значение функции в логарифмической шкале. Левая граница, отрезок внутри и правая граница ящика — 25%, 50% и 75% персентили соответственно, границы усов соответствуют минимальным и максимальным значениям.

где $\lambda > 0$ — коэффициент регуляризации. Подобная постановка задачи позволяет получить единственное решение (так как функция (3.6) строго выпукла), а так же помогает бороться с помехами.

Для сравнения методов оптимизации использовался коэффициент регуляризации $\lambda = 10^{-6}$ и следующая процедура

1. генерируются наблюдения $\{y_i, \boldsymbol{\varphi}_i\}_1^N$ и начальная точка $\mathbf{x}_0$ согласно следующим распределениям:

$$\boldsymbol{\varphi}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \varepsilon_i \sim \mathcal{N}(\mathbf{0}, 10^{-2}), \quad \mathbf{x}_0 \sim \mathcal{N}(\mathbf{x}_\star, \mathbf{I});$$

2. для каждого алгоритма $\mathcal{A}$ запускается процесс оптимизации, и берется минимальное значение функции по аргументу за 100 шагов $f_{min}^{\mathcal{A}} := \min\limits_{0\leqslant t\leqslant 100} f(\mathbf{x}_t)$;
3. считается статистика $\Upsilon(\mathcal{A})$:

$$\Upsilon(\mathcal{A}) = |\{\mathcal{A}' : \ f_{min}^{\mathcal{A}'} < f_{min}^{\mathcal{A}}\}|; \tag{3.7}$$

4. этапы 1-3 повторяются тысячу раз, в результате чего получается выборка значений $\Upsilon(\mathcal{A})$ для каждого из алгоритмов.

Поясним выбор $\Upsilon(\mathcal{A})$ как метрики оценивания качества алгоритма. В силу того, что при каждом повторении этапов (1) и (2) используется разные функции f,

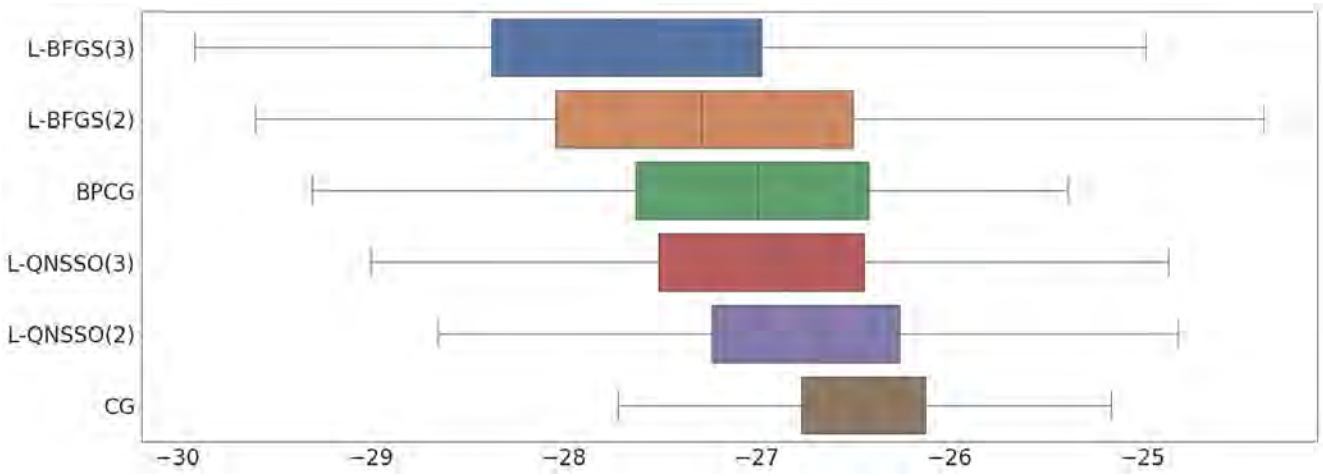


Рисунок 3.7 — Ящики с усами статистики $\min\limits_{0 \leqslant t \leqslant n} f(\mathbf{x}_t)$ для функции Розенброка (3.3). По оси абсцисс отмечено значение функции в логарифмической шкале. Левая граница, отрезок внутри и правая граница ящика — 25%, 50% и 75% персентили соответственно, границы усов соответствуют минимальным и максимальным значениям.

абсолютные значения минимума функции малоинформативны. Удобнее рассматривать относительный показатель качества алгоритмов — число алгоритмов, чей результат превосходит результат $\mathcal{A}$, — чем и является $\Upsilon(\mathcal{A})$.

В силу высокой размерности задачи, в сравнении участвовали лишь методы с незначительными требованиями к потребляемой памяти: CG, BPCG, L-BFGS(2), L-BFGS(3), L-QNSSO(2), L-QNSSO(3). Результаты эксперимента — сумма значений статистик $\Upsilon(\mathcal{A})$ для каждого из алгоритмов — представлены в Таблице 1. Согласно этим результатам алгоритмы L-QNSSO значительно превоходят конкуретнов. В частности, алгоритм L-QNSSO(2) доставляет минимальное значение функции более чем в половине экспериментов. Стоит так же отметить, что на этой задаче размерность подпространства играет негативную роль как для метода L-QNSSO, так и для метода L-BFGS.

3.2.4 Логистическая регрессия для классификации химических соединений

Рассмотрим набор данных DOROTHEA [114], содержащий информацию о 1950-ти лекарствах (химических соединениях), разделенных на три множе-

Таблица 1 — Сумма $\Upsilon(\mathcal{A})$ (3.7) за 1000 экспериментов.

Алгоритм	Сумма $\Upsilon(\mathcal{A})$
L-QNSSO(2)	425
L-QNSSO(3)	1498
CG	1896
L-BFGS(2)	3520
L-BFGS(3)	3808
BPCG	3853

ства: тренировочное (800), проверочное (350) и тестовое (800). Каждое лекарство представлено *вектором признаков* — бинарным вектором размерности 10^5 (половина из элементов которого умышленно сгенерированны случайным образом для усложнения задачи) и флагом активности, характеризующим связываемость лекарства с тромбином. Ставится следующая задача классификации: на основе тренировочного множества построить алгоритм предсказания флага активности по химического соединения по его вектору признаков.

Логистическая регрессия — один из методов решения поставленной задачи — это статистическая модель прогнозирования вероятности, зачастую используемая для задач классификации [115], которая представляет собой применение логистической функции к линейной модели. Решающая функция имеет следующий вид:

$$h_i(\mathbf{x}) = h(\boldsymbol{\varphi}_i, \mathbf{x}) = \frac{1}{1 + \exp\left(-\boldsymbol{\varphi}_i^\top \mathbf{x}\right)},$$

где вектор φ_i — вектор признаков i-го объекта, а $\mathbf{x}$ — вектор параметров модели. Отметим, что размерности вектора признаков и вектора параметров модели совпадают. Один из наиболее распространенных способов оценки параметров логистической регресии — максимизация функции правдоподобия — частный случай задачи минимизации функционала эмпирического риска (1.1):

$$L(\mathbf{x}) = \sqrt[\frac{1}{N}]{\prod_{i=1}^{N} h_i(\mathbf{x})^{y_i}(1 - h_i(\mathbf{x}))^{(1-y_i)}} \to \max_{\mathbf{x}},$$

что эквивалентно минимизации её логарифма со знаком минус:

$$\begin{aligned} f(\mathbf{x}) &= -\log L(\mathbf{x}) \\ &= -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log h_i(\mathbf{x}) + (1-y_i)\log(1-h_i(\mathbf{x}))\right] \to \min_{\mathbf{x}}. \end{aligned} \tag{3.8}$$

Градиент функции (3.8) по $\mathbf{x}$ имеет вид:

$$\nabla f(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}(h_i(\mathbf{x}) - y_i)\boldsymbol{\varphi}_i. \tag{3.9}$$

В контексте оптимизации целью является минимизация значения функции (3.8) на тренировочном наборе данных. Максимизация качества классификации на тестовом наборе данных является задачей распознавания образов и потому не рассматривается. Сравнение алгоритмов проведено следующим образом:

1. генерируется начальная точка согласно распределению $\mathbf{x}_0 \sim \mathcal{N}\left(\mathbf{0}, \sigma_x^2\mathbf{I}\right)$, $\sigma_x^2 = 10^{-6}$;
2. для каждого алгоритма запускается процесс оптимизации функции (3.8) на тренировочном наборе данных, и считается минимальное значение функции за первые 100 итераций $\min\limits_{0\leqslant t\leqslant 100} f(\mathbf{x}_t)$;
3. этапы 1-2 повторяются тысячу раз, таким образом получается выборка значений $\min\limits_{0\leqslant t\leqslant 100} f(\mathbf{x}_t)$ для каждого из алгоритмов.

Распределения полученных выборок для каждого из алгоритма проиллюстрированы в виде “ящиков с усами” на Рисунке 3.8. Легко заметить, что методы L-QNSSO и CG значительно превосходят методы L-BFGS и BPCG. При этом, наилучшее качество как в смысле 50% и 25% персентилей, так и в смысле минимального значения достигается алгоритмом L-QNSSO(5). Примечательно, что зависимость между размером истории и качеством метода L-QNSSO на данной задаче немонотонна. Отметим, наконец, что в контексте задачи распознавания образов распространён прием повторного перезапуска процесса оптимизации с разными начальными условиями для получения наилучшего результата. В этом ключе наиболее релевантны оптимистичные характеристики распределения полученных выборок: 25% персентиль и минимальное значение, по которым алгоритмы L-QNSSO демонстрируют превосходящие результаты.

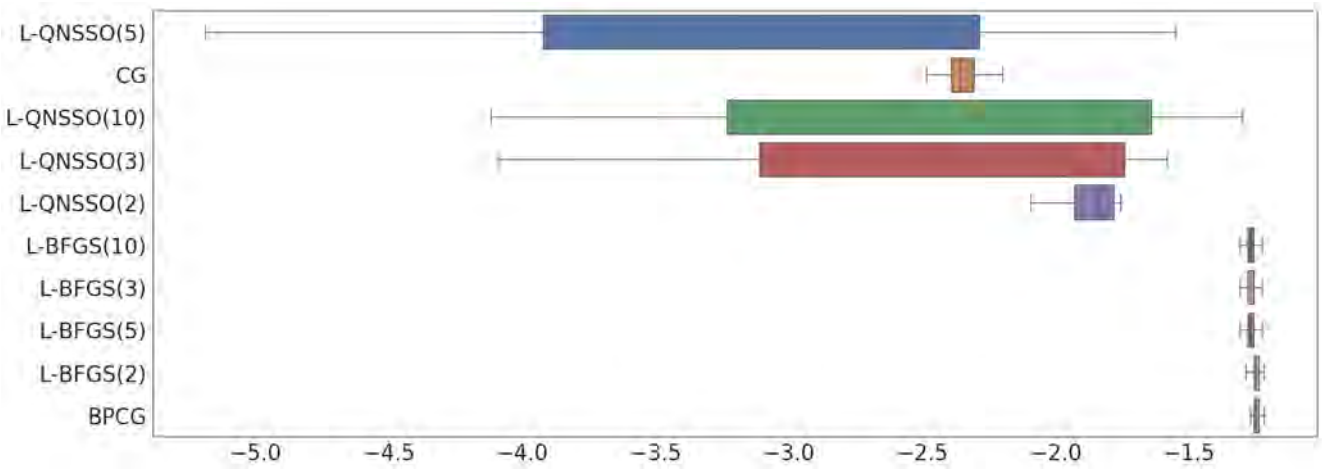


Рисунок 3.8 — Ящики с усами статистики $\min\limits_{0 \leqslant t \leqslant 100} f(\mathbf{x}_t)$ функции потерь логистической регресии (3.8) на тренировочном наборе данных. По оси абсцисс отмечено значение функции в логарифмической шкале. Левая граница, отрезок внутри и правая граница ящика — 25%, 50% и 75% персентили соответственно, границы усов соответствуют минимальным и максимальным значениям.

Заключение

Основные результаты работы заключаются в следующем.

– Получена характеристика скорости сходимости методом последовательной подпространственной оптимизации с квадратичным суррогатом для квадратичного строго выпуклого случая через ошибку проекции и ошибку аппроксимации (Раздел 2.1.2, Лемма 1) [56];
– Установлены критерии линейной и суперлинейной скоростей сходимости методов последовательной подпространственной оптимизации с квадратичным суррогатом для случая строго выпуклой целевой функции (Раздел 2.1.3, Теорема 3) [57];
– Разработан корректирующий метод последовательной подпространственной оптимизации с линейной скоростью сходимости (Раздел 2.3.1, Теорема 2) [57], разработан метод последовательной подпространственной оптимизации с квазиньютоновским шагом, сходящийся с конечно число итераций с линейной скоростью сходимости (Раздел 2.3.2, Замечание 14) для случая строго выпуклой квадратичной целевой функции [57];
– Разработан метод модифицированных знако–возмущённых сумм для определения точного доверительного множества параметров линейной модели при условии независимых друг с другом и с входами модели, а в остальном произвольных аддитивных помех в наблюдениями (Раздел 2.4, Теорема 4), для одномерного случая получено аналитическое выражение границ доверительного интервала и условия их состоятельности (Раздел 2.5, Лемма 7 и Лемма 8) [45—47].

Список литературы

1. *Ljung*, *L.* System identification (2nd ed.): theory for the user / L. Ljung. — Upper Saddle River, NJ, USA : Prentice Hall PTR, 1999.
2. *Boyd*, *S.* Convex Optimization / S. Boyd, L. Vandenberghe. — Cambridge university press, 2004.
3. *Граничин*, *О. Н.* Рандомизированные алгоритмы оценивания и оптимизации при почти произвольных помехах / О. Н. Граничин, Б. Т. Поляк. — М.: Наука, 2003.
4. *Вапник*, *В. Н.* Теория распознавания образов: статистические проблемы обучения / В. Н. Вапник, А. Я. Червоненкис. — Наука. Гл. ред. физ.-мат. лит., 1974.
5. *Bishop*, *C. M.* Pattern Recognition and Machine Learning / C. M. Bishop. — Springer, 2006.
6. *Vapnik*, *V.* The Nature of Statistical Learning Theory / V. Vapnik. — Springer science & business media, 2013.
7. *Goodfellow*, *I.* Deep Learning / I. Goodfellow, Y. Bengio, A. Courville. — MIT press, 2016.
8. *Hestenes*, *M. R.* Methods of conjugate gradients for solving linear systems / M. R. Hestenes, E. Stiefel // Journal of Research of the National Bureau of Standards. — 1952. — Vol. 49, no. 6. — P. 409—436.
9. *Fletcher*, *R.* Function minimization by conjugate gradients / R. Fletcher, C. M. Reeves // The Computer Journal. — 1964. — Vol. 7, no. 2. — P. 149—154.
10. *Polak*, *E.* Note sur la convergence de méthodes de directions conjuguées / E. Polak, G. Ribiere // ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique. — 1969. — Vol. 3, R1. — P. 35—43.
11. *Polyak*, *B. T.* The conjugate gradient method in extremal problems / B. T. Polyak // USSR Computational Mathematics and Mathematical Physics. — 1969. — Vol. 9, no. 4. — P. 94—112.

12. *Golub*, *G. H.* Some history of the conjugate gradient and Lanczos algorithms: 1948–1976 / G. H. Golub, D. P. O'Leary // SIAM Review. — 1989. — Vol. 31, no. 1. — P. 50—102.

13. *Cohen*, *A. I.* Rate of convergence of several conjugate gradient algorithms / A. I. Cohen // SIAM Journal on Numerical Analysis. — 1972. — Vol. 9, no. 2. — P. 248—259.

14. *Davidon*, *W. C.* Variable metric method for minimization / W. C. Davidon // SIAM Journal on Optimization. — 1991. — Vol. 1, no. 1. — P. 1—17.

15. *Broyden*, *C. G.* The convergence of a class of double-rank minimization algorithms 1. general considerations / C. G. Broyden // IMA Journal of Applied Mathematics. — 1970. — Vol. 6, no. 1. — P. 76—90.

16. *Davidon*, *W. C.* New least-square algorithms / W. C. Davidon // Journal of Optimization Theory and Applications. — 1976. — Vol. 18, no. 2. — P. 187—197.

17. *Nocedal*, *J.* Updating quasi-Newton matrices with limited storage / J. Nocedal // Mathematics of Computation. — 1980. — Vol. 35, no. 151. — P. 773—782.

18. *Fletcher*, *R.* Practical Methods of Optimization (2nd ed.) / R. Fletcher. — New Your: John Wiley, 1987.

19. *Conn*, *A. R.* Convergence of quasi-Newton matrices generated by the symmetric rank one update / A. R. Conn, N. I. Gould, P. L. Toint // Mathematical Programming. — 1991. — Vol. 50, no. 1—3. — P. 177—195.

20. *Miele*, *A.* Study on a memory gradient method for the minimization of functions / A. Miele, J. Cantrell // Journal of Optimization Theory and Applications. — 1969. — Vol. 3, no. 6. — P. 459—470.

21. *Cragg*, *E.* Study on a supermemory gradient method for the minimization of functions / E. Cragg, A. Levy // Journal of Optimization Theory and Applications. — 1969. — Vol. 4, no. 3. — P. 191—205.

22. *Fletcher*, *R.* A limited memory steepest descent method / R. Fletcher // Mathematical Programming. — 2012. — Vol. 135, no. 1/2. — P. 413—436.

23. On iterated-subspace minimization methods for nonlinear optimization / A. R. Conn [et al.] // Linear and Nonlinear Conjugate-Gradient Related Methods. — SIAM, 1994. — P. 50—78.

24. *Narkiss*, *G.* Sequential Subspace Optimization Method for Large-Scale Unconstrained Optimization : tech. rep. / G. Narkiss, M. Zibulevsky ; Technion-IIT, Department of Electrical Engineering. — 2005. — P. 31.

25. *Yuan*, *Y.-X.* A Review on Subspace Methods for Nonlinear Optimization : tech. rep. / Y.-X. Yuan. — 2014.

26. *Chouzenoux*, *E.* A majorize–minimize strategy for subspace optimization applied to image restoration / E. Chouzenoux, J. Idier, S. Moussaoui // IEEE Transactions on Image Processing. — 2010. — Vol. 20, no. 6. — P. 1517—1528.

27. *Narkiss*, *G.* Support Vector Machine via Sequential Subspace Optimization / G. Narkiss, M. Zibulevsky. — Technion-IIT, Department of Electrical Engineering, 2005.

28. *Zheng*, *Y.* Efficient variational Bayesian approximation method based on subspace optimization / Y. Zheng, A. Fraysse, T. Rodet // IEEE Transactions on Image Processing. — 2014. — Vol. 24, no. 2. — P. 681—693.

29. SEBOOST-Boosting stochastic learning using subspace optimization techniques / E. Richardson [et al.] // Advances in Neural Information Processing Systems. — 2016. — P. 1534—1542.

30. *Zibulevsky*, *M.* L1-L2 optimization in signal and image processing / M. Zibulevsky, M. Elad // IEEE Signal Processing Magazine. — 2010. — Vol. 27, no. 3. — P. 76—88.

31. *Граничин*, *О. Н.* Оценивание параметров линейной регрессии при произвольных помехах / О. Н. Граничин // Автоматика и телемеханика. — 2002. — № 1. — С. 30—41.

32. *Granichin*, *O.* Linear regression and filtering under nonstandard assumptions (Arbitrary noise) / O. Granichin // IEEE Transactions on Automatic Control. — 2004. — Vol. 49, no. 10. — P. 1830—1837.

33. *Goldenshluger*, *A. V.* Estimation of regression parameters with arbitrary noise / A. V. Goldenshluger, B. T. Polyak // Mathematical Methods of Statistics. — 1993. — Vol. 2, no. 1. — P. 18—29.

34. *Граничин*, *О. Н.* Алгоритм стохастической аппроксимации с возмущением на входе для идентификации статического нестационарного дискретного объекта / О. Н. Граничин // Вестник Санкт-Петербургского университета. Серия 1. Математика. Механика. Астрономия. — 1988. — № 3. — С. 92—93.

35. *Граничин*, *О. Н.* Адаптивное управление с использованием пробных сигналов в канале обратной связи / О. Н. Граничин, В. Н. Фомин // Автоматика и телемеханика. — 1986. — № 2. — С. 100—112.

36. *Граничин*, *О. Н.* Рандомизированные алгоритмы оценивания и оптимизации при почти произвольных помехах / О. Н. Граничин, Б. Т. Поляк. — Москва: Наука, 2003.

37. Two procedures with randomized controls for the parameters' confidence region of linear plant under external arbitrary noise / K. Amelin [et al.] // Proc. of the IEEE Int. Symposium on Intelligent Control (ISIC). — IEEE. 2012. — P. 1226—1231.

38. Combined procedure with randomized controls for the parameters' confidence region of linear plant under external arbitrary noise / K. Amelin [et al.] // Proc. of the 51st Conference on Decision and Control (CDC2012). — IEEE. 2013. — P. 2134—2139.

39. *Amelin*, *K.* Randomized controls for linear plants and confidence regions for parameters under external arbitrary noise / K. Amelin, O. Granichin // Proc. of the American Control Conference (ACC). — IEEE. 2012. — P. 0743—1619.

40. *Campi*, *M. C.* Guaranteed non-asymptotic confidence regions in system identification / M. C. Campi, E. Weyer // Automatica. — 2005. — Vol. 41, no. 10. — P. 1751—1764.

41. *Csaji*, *B. C.* Non-asymptotic confidence regions for the least-squares estimate / B. C. Csaji, M. C. Campi, E. Weyer // Proceedings of the 16th IFAC Symposium on System Identification (SYSID 2012). — 2012. — July. — P. 227—232.

42. *Granichin*, *O. N.* The nonasymptotic confidence set for parameters of a linear control object under an arbitrary external disturbance / O. N. Granichin // Automation and Remote Control. — 2012. — Vol. 73, no. 1. — P. 20—30.

43. Патент: Программа для оптического распознавания визуальной текстовой информации на арабском языке (ОРТА-2Б.ГРС) / О. А. Берникова [и др.]. — 2013.

44. Методы оптического распознавания текста на арабском языке / О. А. Берникова [и др.] // Стохастическая оптимизация в информатике. — 2013. — Т. 9, № 2. — С. 3—20.

45. *Сенов*, *А. А.* Доверительные множества при почти произвольных помехах в контексте линейных моделей рекомендательных систем / А. А. Сенов // Стохастическая оптимизация в информатике. — 2013. — Т. 9, № 1. — С. 68—86.

46. Exact confidence regions for linear regression parameter under external arbitrary noise / A. Senov [et al.] // American Control Conference (ACC), 2014. — IEEE. 2014. — P. 5097—5102.

47. *Сенов*, *А. А.* Идентификация параметров линейной регрессии при произвольных внешних помехах в наблюдениях / А. А. Сенов, О. Н. Граничин // XII всероссийское совещание по проблемам управления ВСПУ-2014. — 2014. — С. 2708—2719.

48. *Senov*, *A.* Improving distributed stochastic gradient descent estimate via loss function approximation / A. Senov // IFAC-PapersOnLine. — 2015. — Vol. 48, no. 25. — P. 292—297.

49. *Сенов*, *А. А.* Квадратичная проективная регрессия как метод обучения в разреженных пространствах высокой размерности / А. А. Сенов // Эвристические Алгоритмы и Распределенные Вычисления. — 2015. — Т. 2, № 4. — С. 73—92.

50. *Сенов*, *А. А.* Улучшение оценки распределенного стохастического градиентного спуска через аппроксимацию функции потерь / А. А. Сенов // Стохастическая оптимизация в информатике. — 2015. — Т. 11, № 1. — С. 103—126.

51. *Boiarov*, *A.* Arabic manuscript author verification using deep convolutional networks / A. Boiarov, A. Senov, A. Knysh // 2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR). — 2017. — P. 1—5.

52. *Senov*, *A.* Accelerating gradient descent with projective response surface methodology / A. Senov // International Conference on Learning and Intelligent Optimization. — Springer. 2017. — P. 376—382.

53. *Senov*, *A.* Projective approximation based gradient descent modification / A. Senov, O. Granichin // IFAC-PapersOnLine. — 2017. — Vol. 50, no. 1. — P. 3899—3904.

54. *Senov*, *A.* Projective approximation based quasi-Newton methods / A. Senov // International Workshop on Machine Learning, Optimization, and Big Data. — Springer. 2017. — P. 29—40.

55. *Сенов*, *А. А.* Глубокое обучение в задаче реконструкции суперразрешения изображений / А. А. Сенов // Стохастическая оптимизация в информатике. — 2017. — Т. 13, № 2. — С. 38—57.

56. *Сенов*, *А. А.* О методах последовательной подпространственной оптимизации / А. А. Сенов // Стохастическая оптимизация в информатике. — 2018. — Т. 14, № 2. — С. 40—61.

57. *Сенов*, *А. А.* Квазиньютоновские методы последовательной подпространственной оптимизации с квадратичным суррогатом ми- нимизации строго выпуклых функций / А. А. Сенов // Стохастическая оптимизация в информатике. — 2019. — Т. 15, № 1. — С. 20—68.

58. *Montgomery*, *D. C.* Introduction to Linear Regression Analysis / D. C. Montgomery, E. A. Peck, G. G. Vining. — Wiley-Interscience, 2007.

59. *Фомин*, *В. Н.* Рекуррентное оценивание и адаптивная фильтрация / В. Н. Фомин. — Москва: Наука, 1984.

60. *Hastie*, *T.* The Elements of Statistical Learning: Data Mining, Inference, and Prediction / T. Hastie, R. Tibshirani, J. H. Friedman. — New York, NY: Springer, 2009.

61. *Vapnik*, *V. N.* The nature of statistical learning theory / V. N. Vapnik. — Springer–Verlag, 1995.

62. *Нестеров*, *Ю. Е.* Введение в выпуклую оптимизацию / Ю. Е. Нестеров. — МЦНМО, 2010.

63. *Granichin*, *O. N.* Stochastic approximation search algorithms with randomization at the input / O. N. Granichin // Automation and Remote Control. — 2015. — Vol. 76, no. 5. — P. 762—775.

64. *Nesterov*, *Y.* Introductory Lectures on Convex Programming Volume I: Basic Course / Y. Nesterov. — Citeseer, 1998.

65. *Поляк*, *Б. Т.* Введение в оптимизацию / Б. Т. Поляк. — Наука. Гл. ред. физ.-мат. лит., 1983.

66. *Ypma*, *T. J.* Historical development of the Newton–Raphson method / T. J. Ypma // SIAM review. — 1995. — Vol. 37, no. 4. — P. 531—551.

67. *Polyak*, *B. T.* Newton's method and its use in optimization / B. T. Polyak // European Journal of Operational Research. — 2007. — Vol. 181, no. 3. — P. 1086—1096.

68. *Cauchy*, *A.-L.* Méthode générale pour la résolution des systemes d'équations simultanées / A.-L. Cauchy // Comp. Rend. Sci. Paris. — 1847. — Vol. 25, no. 1847. — P. 536—538.

69. *Crockett*, *J. B.* Gradient methods of maximization / J. B. Crockett, H. Chernoff // Pacific Journal of Mathematics. — 1955. — Vol. 5, no. 1. — P. 33—50.

70. *Nocedal*, *J.* Numerical Optimization / J. Nocedal, S. J. Wright. — Springer, 2006.

71. *Barzilai*, *J.* Two-point step size gradient methods / J. Barzilai, J. M. Borwein // IMA Journal of Numerical Analysis. — 1988. — Vol. 8, no. 1. — P. 141—148.

72. *Dai*, *Y.-H.* A new analysis on the Barzilai-Borwein gradient method / Y.-H. Dai // Journal of the Operations Research Society of China. — 2013. — Vol. 1, no. 2. — P. 187—198.

73. *Wei*, *Z.* New quasi-Newton methods for unconstrained optimization problems / Z. Wei, G. Li, L. Qi // Applied Mathematics and Computation. — 2006. — Vol. 175, no. 2. — P. 1156—1188.

74. *Zhang*, *J. Z.* New quasi-Newton equation and related methods for unconstrained optimization / J. Z. Zhang, N. Y. Deng, L. H. Chen // Journal of Optimization Theory and Applications. — 1999. — Vol. 102, no. 1. — P. 147—167.

75. *Don*, *F. J. H.* On the symmetric solutions of a linear matrix equation / F. J. H. Don // Linear Algebra and its Applications. — 1987. — Vol. 93. — P. 1—7.

76. *Dai*, *Y.-H.* Nonlinear conjugate gradient methods / Y.-H. Dai // Wiley Encyclopedia of Operations Research and Management Science. — 2010.

77. *Powell*, *M. J. D.* Restart procedures for the conjugate gradient method / M. J. D. Powell // Mathematical programming. — 1977. — Vol. 12, no. 1. — P. 241—254.

78. *Dai*, *Y.* Convergence properties of Beale-Powell restart algorithm / Y. Dai, Y. Yuan // Science in China Series A: Mathematics. — 1998. — Vol. 41, no. 11. — P. 1142—1150.

79. *Nemirovski, A.* Orth-method for smooth convex optimization / A. Nemirovski // Izvestia AN SSSR, Transl.: Eng. Cybern. Soviet J. Comput. Syst. Sci. — 1982. — Vol. 2. — P. 937—947.

80. *Shewchuk, J. R.* An introduction to the conjugate gradient method without the agonizing pain / J. R. Shewchuk. — 1994. — Carnegie-Mellon University. Department of Computer Science.

81. *Немировский, А. С.* Методы оптимизации, адаптивные к «существенной» размерности задачи / А. С. Немировский, Д. Б. Юдин // Автоматика и телемеханика. — 1977. — № 4. — С. 75—87.

82. *Powell, M. J.* A new algorithm for unconstrained optimization / M. J. Powell // Nonlinear Programming. — Elsevier, 1970. — P. 31—65.

83. *Wang, Z.-H.* A subspace implementation of quasi-Newton trust region methods for unconstrained optimization / Z.-H. Wang, Y.-X. Yuan // Numerische Mathematik. — 2006. — Vol. 104, no. 2. — P. 241—269.

84. *Yuan, Y.-X.* A subspace study on conjugate gradient algorithms / Y.-X. Yuan, J. Stoer // ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik. — 1995. — Vol. 75, no. 1. — P. 69—77.

85. *Wolfe, P.* Convergence conditions for ascent methods / P. Wolfe // SIAM Review. — 1969. — Vol. 11, no. 2. — P. 226—235.

86. *Wolfe, P.* Convergence conditions for ascent methods. II: Some corrections / P. Wolfe // SIAM Review. — 1971. — Vol. 13, no. 2. — P. 185—188.

87. *Yuan, Y.-X.* Subspace techniques for nonlinear optimization / Y.-X. Yuan // Some Topics in Industrial and Applied Mathematics. — World Scientific, 2007. — P. 206—218.

88. *Yuan, Y.-X.* Subspace methods for large scale nonlinear equations and nonlinear least squares / Y.-X. Yuan // Optimization and Engineering. — 2009. — Vol. 10, no. 2. — P. 207—218.

89. *LeCun, Y.* The MNIST Dataset Of Handwritten Digits / Y. LeCun, C. Cortes, C. J. Burges. — 1999. — URL: http://yann.lecun.com/exdb/mnist.

90. *Krizhevsky, A.* Learning multiple layers of features from tiny images : tech. rep. / A. Krizhevsky, G. Hinton ; Citeseer. — 2009.

91. *Hartley, H. O.* Exact confidence regions for the parameters in non-linear regression laws / H. O. Hartley // Biometrika. — 1964. — Vol. 51, no. 3/4. — P. 347—353.

92. *Draper, N. R.* Applied Regression Analysis. Vol. 326 / N. R. Draper, H. Smith. — John Wiley & Sons, 1998.

93. *Davison, A. C.* Bootstrap Methods and Their Application / A. C. Davison, D. V. Hinkley. — Cambridge university press, 1997.

94. Bootstrapping regression models / D. A. Freedman [et al.] // The Annals of Statistics. — 1981. — Vol. 9, no. 6. — P. 1218—1228.

95. *Fox, J.* Bootstrapping regression models / J. Fox // An R and S-PLUS Companion to Applied Regression: A Web Appendix to the Book. Sage, Thousand Oaks, CA. URL: http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-bootstrapping.pdf. — 2002.

96. *Bai, E.-W.* Membership set estimators: size, optimal inputs, complexity and relations with least squares / E.-W. Bai, R. Tempo, H. Cho // IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications. — 1995. — Vol. 42, no. 5. — P. 266—277.

97. *Vicino, A.* Sequential approximation of feasible parameter sets for identification with set membership uncertainty / A. Vicino, G. Zappa // IEEE Transactions on Automatic Control. — 1996. — Vol. 41, no. 6. — P. 774—785.

98. *Dalai, M.* Parameter identification for nonlinear systems: guaranteed confidence regions through LSCR / M. Dalai, E. Weyer, M. C. Campi // Automatica. — 2007. — Vol. 43, no. 8. — P. 1418—1425.

99. *Csáji, B. C.* Sign-Perturbed Sums: A new system identification approach for constructing exact non-asymptotic confidence regions in linear regression models / B. C. Csáji, M. C. Campi, E. Weyer // IEEE Transactions on Signal Processing. — 2014. — Vol. 63, no. 1. — P. 169—181.

100. *Kieffer, M.* Guaranteed characterization of exact non-asymptotic confidence regions as defined by LSCR and SPS / M. Kieffer, E. Walter // Automatica. — 2014. — Vol. 50, no. 2. — P. 507—512.

101. *Волкова, М. В.* Рандомизированные алгоритмы оценивания параметров инкубационных процессов в условиях неопределённостей и конечного числа наблюдений: дис. …канд. физ.-мат. наук: 01.01.09 / М. В. Волкова. — 2018. — Санкт-Петербургский Государственный Университет, СПб.

102. *Akaike, H.* On a successive transformation of probability distribution and its application to the analysis of the optimum gradient method / H. Akaike // Annals of the Institute of Statistical Mathematics. — 1959. — Vol. 11, no. 1. — P. 1—16.

103. *Forsythe, G. E.* On the asymptotic directions of thes-dimensional optimum gradient method / G. E. Forsythe // Numerische Mathematik. — 1968. — Vol. 11, no. 1. — P. 57—76.

104. *Фаддеев, Д. К.* Вычислительные методы линейной алгебры. Т. 1 / Д. К. Фаддеев, В. Н. Фаддеева. — Физматгиз Москва, 1960.

105. *O'Leary, D. P.* Estimating the largest eigenvalue of a positive definite matrix / D. P. O'Leary, G. Stewart, J. S. Vandergraft // Mathematics of Computation. — 1979. — Vol. 33, no. 148. — P. 1289—1292.

106. *Парлетт, Б. Н.* Симметричная проблема собственных значений: Численные методы / Б. Н. Парлетт, Х. Д. Икрамов, Ю. А. Кузнецов. — Мир, 1983.

107. *Golub, G. H.* Matrix Computations / G. H. Golub, C. F. Van Loan. — 4th. — JHU press, 2012.

108. *Brent, R. P.* An algorithm with guaranteed convergence for finding a zero of a function / R. P. Brent // The Computer Journal. — 1971. — Vol. 14, no. 4. — P. 422—425.

109. *Van Der Walt, S.* The NumPy array: a structure for efficient numerical computation / S. Van Der Walt, S. C. Colbert, G. Varoquaux // Computing in Science & Engineering. — 2011. — Vol. 13, no. 2. — P. 22.

110. SciPy: Open source scientific tools for Python / E. Jones, T. Oliphant, P. Peterson, [et al.]. — 2001–. — URL: http://www.scipy.org/ ; Просмотрено: 2019-09-01.

111. *Rosenbrock, H. H.* An automatic method for finding the greatest or least value of a function / H. H. Rosenbrock // The Computer Journal. — 1960. — Vol. 3, no. 3. — P. 175—184.

112. *Shang, Y.-W.* A note on the extended Rosenbrock function / Y.-W. Shang, Y.-H. Qiu // Evolutionary Computation. — 2006. — Vol. 14, no. 1. — P. 119—126.

113. *Quapp, W.* Searching minima of an n-dimensional surface: A robust valley following method / W. Quapp // Computers & Mathematics with Applications. — 2001. — Vol. 41, no. 3/4. — P. 407—414.

114. Result analysis of the NIPS 2003 feature selection challenge / I. Guyon [et al.] // Advances in Neural Information Processing Systems. — 2005. — P. 545—552.

115. *Cramer, J. S.* The early origins of the logit model / J. S. Cramer // Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences. — 2004. — Vol. 35, no. 4. — P. 613—626.

SAINT PETERSBURG STATE UNIVERSITY

Senov Aleksandr Alekseevich

Methods of optimization and parameter estimation in multidimensional problems with arbitrary noise

Specialization 01.01.09 —
«Discrete Mathematics
and Mathematical Cybernetics»

Dissertation is submitted for the
degree Candidate of Physical and Mathematical Sciences

Translated from Russian

Scientific advisor:
Professor, Doctor of Science in Physics and Mathematics
Granichin Oleg Nikolaevich

Saint Petersburg — 2020

Table of contents

Introduction

The solution of many problems of adaptive control, pattern recognition, modeling, signal processing is reduced to solving corresponding problems of nonlinear optimization and parameter estimation [1–4]. The growth of the amount and sources of data, the development of computer technology and constantly increasing requirements to the quality of models form a need for the development of methods applicable in high-dimensional spaces. The pattern recognition problem is one of the examples: there is a direct relationship between the number of parameters of the model and its generalization capability, and at the same time the complex nature of input objects, such as images and sound, leads to the fact that the dimension of the parameter space can be counted in millions [5–7].

Iterative convex optimization methods are widely used in adaptive control, pattern recognition, modeling, data analysis and signal processing [2]. The first formulations of iterative optimization methods are often attributed to the works of I. Newton, and convex optimization formation as an independent discipline and the study of the convergence properties of methods are attributed to the middle of the 20-th century. The gradient descent method and the Newton – Raphson method are probably the best known optimization methods. Under proper conditions, the Newton-Raphson method converges to the minimum point of the function with a quadratic convergence speed that is sufficient for the majority of practical problems. However, the need to calculate and store an inverted matrix of second derivatives makes it inapplicable in the case of high dimensional problems. By contrast, the gradient descent method has low resource intensity, but also has a low convergence rate. That ”gap” between the gradient descent and Newton-Raphson methods have been actively filled up since the mid 20-th century by the development of such optimization methods as: conjugate gradient methods [8–13], quasi-Newton methods [14–19], memory gradient methods [20–22], and others.

Sequential subspace optimization (SSO) is another direction proposed at the end of the 20-th century in the work of R. Conn, N. Gould, A. Sartenaer and Ph. Toint [23] and further developed in the works of G. Narkiss, M. Zhibulevsky, Yu. Yuan, E. Chouzenoux et al. [24–26]. The main idea of the SSO approach is the sequential formation of a subspace of significantly smaller dimension than the original one and subsequent optimization of the target function along the selected subspace. Translation of the problem into a subspace of smaller dimension allows to reduce utilization

of computational resources, which is especially important in high-dimensional problems. Sequential subspace optimization methods are successfully applied in practice, including but not limited to the problems of pattern recognition [27–29] and image analysis [26; 30].

One of the major drawbacks of the sequential subspace optimization is poorly studied theoretical properties of the methods: only for some of them convergence guarantees have been proved, while certain general theoretical properties, such as the impact of the choice of subspaces and the quality of the solution of the subspace optimization problem on the convergence, the lower bound of convergence, etc. — have never been properly studied. *Thus, the study of properties and synthesis of sequential subspace optimization methods are of essential relevance for high-dimensional optimization problems*.

If the presence of uncertainties estimation methods are an alternative to solving the optimization problem. The development of the theory of estimation in conditions of centered additive noise was originated by N. Wiener, A. N. Kolmogorov, R. Kalman and R. Bucy in the middle of the XX-th century. Further development of estimation methods in the area of almost arbitrary noises was made by V. N. Fomin, A. L. Fradkov, V. A. Yakubovich. The work of A. B. Tsybakov, A.V. Goldenschluger, J. Spall, B. T. Polyak and O. N. Granichin on randomized stochastic approximation methods and linear models parameters estimation under arbitrary external noise also is worth noting. However, in the case of a small number of observations and the presence of systematic errors, the estimations obtained by applying optimization methods and point estimation methods can lead to unsatisfactory results. In such a conditions, an approach based on the evaluation of *confidence regions* that contain the true value of the parameter with a given probability is more reasonable. In context of small number of observations, the problem of constructing an *exact* confidence set, which contains the true value of the parameter exactly with the given probability regardless of the number of observations is of particular interest. Until recently, only the methods for asymptotic confidence regions construction were known, which are only applicable under sufficiently strong assumptions about the noise distribution [1]. A significant disadvantage of these methods is that they can guarantee the result only when the number of measurements tends to infinity, while for a small number of measurements the results may be unsatisfactory.

Randomization is a common approach to solve estimation problems in the context of small number of measurements and a high degree of uncertainty. The main idea of it is to inject random but controlled perturbations. Randomized methods are successfully

used in the problems of parameter estimation with arbitrary noise [31–36], as well as for system parameters confidence regions construction [37–39]. For example, in the work of M. Campi and E. Weyer [40] an approach of sign-dominant correlation regions exclusion (LSCR, leave-out sign-dominant correlation regions) was proposed for finding the exact confidence intervals of a linear system in the one-dimensional case. Subsequently, on the basis of that approach the same authors proposed the Sign-Perturbed Sums (SPS) method to determine the exact confidence regios of the linear regression parameters for centered and symmetrically distributed additive noise [41]. One of the main limitations of the LSCR and forSPS methods is the requirement of the noise distribution symmetry around zero. In many practical problems noises can not only be biased, but also have a non-random nature, that makes both LSCR and SPS inapplicable. In [42] O. N. Granichin proposed a modification of the LSCR method for the case of unknown and almost arbitrary noises but controlled inputs. While it significantly relaxes the conditions on the noise distribution, it still can be only applied to the one-dimensional case . *Thus, the problem of construction and characterizing the exact confidence region for the parameters of a linear model under almost arbitrary noises remains open.*

The dissertation's **objectives** consist of two parts: (1) development and of sequential subspace optimization methods and investigation of their properties for the strongly convex differentiable function optimization problem and (2) development of methods for constructing exact confidence regions of a linear model parameter in case a small number of observations and almost arbitrary noises. To achieve these objectives the following key goals have been set and achieved:

- To determine convergence conditions for sequential subspace optimization methods;
- To develop sequential subspace optimization methods applicable in high-dimensional problems;
- To develop a method for constructing exact confidence regions for linear model parameters under arbitrary external additive noise.

The novelty of the work consists of the following main **results, submitted to the defence**:

1. Criteria of sublinear, linear and superlinear convergence rates for sequential subspace optimization methods with quadratic surrogate for cases of the quadratic and strictly convex objective function are established;
2. A sequential subspace optimization method that converges in a finite number of iterations with a linear speed in the quadratic case is developed;

3. A method for exact confidence regions construction of linear model parameters in the case of independent and otherwise arbitrary additive noise is developed, for the one-dimensional case, an analytical expression of the confidence interval boundaries and the conditions of their consistency are obtained.

The theoretical value and practical significance The sequential subspace optimization is an actively developing area of mathematical programming aimed at high-dimensional problems. The approach demonstrates state of the art results in problems of practical importance, such as image processing and pattern recognition. The development of sequential subspace optimization methods and their properties contributes to the theory (in the field of nonlinear optimization) and practice (the mentioned problems of image analysis and pattern recognition, as well as data analysis, adaptive control and signal processing). The main theoretical contribution consists of the obtained conditions of sublinear, linear and superlinear convergence rate and the expression of the convergence rate through the chosen subspaces; the quality of the solution of the subspace optimization problem and of the two proposed sequential subspace optimization methods with proven linear convergence rate.

Until recently, there were no methods for obtaining exact confidence sets under independent and otherwise arbitrary noises in the multidimensional case. The main theoretical contributions of this work are: development of a an exact confidence region construction method for linear model parameters in the context of independent and otherwise arbitrary additive noise, as well as analytical expression and conditions for the confidence interval boundaries consistency in the one-dimensional case. In practice, this method allows to determine the exact confidence region of the linear model parameter without a-priory information about the distribution of noise.

Approbation. The main results of the dissertation were reported at seminars at the Department of Mathematical Modelling of Energetic Systems of the Faculty of Applied Mathematics and Control Processes of St. Petersburg State University; at the Control of Complex Systems Laboratory of the Institute for Problems in Mechanical Engineering of the Russian Academy of Sciences; at the 11th Traditional Russian Youth Summer School “Control, Information and Optimization” (Solnechnogorsk, Moscow Region, June 14-20, 2015) and at the subsequent international conferences: American Control Conference (Portland, Oregon, USA, 2014), the 7th All-Russian Congress on Control Problems, VSPU-2014 (Moscow, Russia, 2014), the 16th IFAC Workshop on Control Applications of Optimization, CAO-2015 (Garmisch-Partenkirchen, Germany, 2015); the International Conference on Learning and Intelligent Optimization, LION-

2017 (Nizhny Novgorod, Russia, 2017); the 20th World Congress of the international Federation of Automatic Control, IFAC World Congress 2017 (Toulouse, France, 2017); the 3rd International Conference on Machine Learning, Optimization and Big Data, MOD 2017 (Volterra, Italy, 2017).

The results of the dissertation were partially used in the work on RFBR grants 13–07–00250, 17–51–53053 and RNF projects 16–19–00057.

Based on the materials of the work, a certificate of state registration of a computer program was obtained ORTA-2B.GRS [43].

Publications. The results obtained during the work on the dissertation are incorporated in 14 scientific papers [44–57]. The main results of the research are presented in the published scientific papers. Six papers [46; 48; 51–54] are published in journals indexed in the international scientometric databases Scopus and Web of Science. Works[44; 46; 47; 51; 53] were written in co-authorship. In work [51], A.D. Knysh was responsible for the prolem formulation, A.A. Boiarov was responsible for the classification algorithm development and training while A.A Senov was responsible for image preprocessing and segmentation algorithms development. In works [46; 47; 53], O.N. Granichin was responsible for the general formulation of the problems while A.A. Senov was in charge of for the implementation of the described methods, formulations and proofs of the theoretical results, development and approbation of the software.

In the **first Chapter** the necessary results from the theory of estimation and optimization are presented. The first section illustrates the relationship between the problem of estimating the exact confidence regions and the problem of convex optimization in high-dimensional spaces with an example of the pattern recognition problem. The second section contains description of the strictly convex function optimization problem, details of convex analysis, the gradient descent method, quasi-Newton methods, conjugate gradient methods and sequential subspace optimization. The third section outlines the methods for constructing confidence regions for the cases of normally distributed noise and symmetrically distributed noise, the confidence region construction problem for the linear model parameter under almost arbitrary noise is stated.

The main results of the work are presented in the **second Chapter**. In the first section the general scheme of methods of sequential subspace optimization is described; the lower and upper limits of convergence of SSO methods, the dependence of the convergence rate on the chosen subspaces and the quality of the subspace problem solution for quadratic and strictly convex cases have been proven. In the second section, methods for estimating the quasi-Newtonian direction and a method for constructing subspaces

are described. In the third section, on the basis of the obtained theoretical results, two sequential subspace optimization methods are formulated, and the linear rate of their convergence has been proven. The fourth section presents a modified sign–perturbed sums method, it is proved that the resulted confidence region for the linear model parameter is exact in the case of the independent and otherwise arbitrary additive noise. In the fifth section an analytical expression of the boundaries of the confidence interval for the MSPS method is obtained for the one-dimensional case, and their consistency with respect to the true parameter value is shown.

The **third Chapter** is devoted to illustration and empirical analysis of the proposed methods. In the first section, the properties of the confidence regions obtained by the method of modified sign–perturbed sums are illustrated in a number of examples with modelled data sets. The second section presents the results of the comparison of the proposed sequential subspace optimization method with several alternative methods in the cases of both synthetic and real-world optimization problems.

Structure and volume of the dissertation. The dissertation consists of an introduction, 3 chapters, conclusion and a list of references. The total volume of the dissertation amounts to 80 page, including 8 figures and 1 table. The bibliography comprises 115 titles. The figures and tables are numbered by chapters.

Chapter 1. Optimization in high-dimensional spaces and estimation under uncertainty

Two problems are considered in this Chapter: the convex optimization problem in high-dimensional spaces (Section 1.2) and the problem of estimating confidence regions under uncertainty and finite number of estimations (Section 1.3). The relationship between these two problems is illustrated at the example of the pattern recognition problem (Section 1.1), where a need to solve the corresponding optimization and confidence regions estimation problems arose.

1.1 Optimization and estimation in pattern recognition

Consider the practical formulation of the pattern recognition problem via the empirical risk minimization (the problem of pattern recognition is described in details in Chapter 2 of the book [4]):

$$f(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} \ell\left(y_i, g(\boldsymbol{\varphi}_i, \mathbf{x})\right) \to \min_{\mathbf{x}\in\mathbb{R}^{n_x}}, \tag{1.1}$$

where N — number of observations, $y_i \in \mathbb{R}$ — observed model outputs, $\boldsymbol{\varphi}_i \in \mathbb{R}^{n_\varphi}$ — observed model inputs, $g : \mathbb{R}^{n_\varphi} \times \mathbb{R}^{n_x} \to \mathbb{R}$ — *decision function* parameterized by the vector $\mathbf{x} \in \mathbb{R}^{n_x}$, and ℓ — *loss function*. In this case, observations are formed by the model $y_i = g(\boldsymbol{\varphi}_i, \mathbf{x}_\star) + \varepsilon_i$, where $\mathbf{x}_\star \in \mathbb{R}^{n_x}$ — true unknown parameter value, and $\varepsilon_i \in \mathbb{R}$ — additive noise. Note that such a formulation of the pattern recognition problem is neither generic nor unique, aspects empirical risk minimization are described in the works [4; 36; 58–60]. However, this statement is sufficient to demonstrate the following thesis: *the optimization problem and the confidence region estimation problem are complement to each other and can be applied to solve the same problem under different conditions*. In the following subsections, this thesis is illustrated by two examples: the case of a high dimension of the parameter space n_x and the case of a small number of dimensions N under arbitrary noise.

1.1.1 The case of a high-dimensional parameter space

Let the loss function ℓ and the decision function g be strictly convex by the argument $\mathbf{x} \in \mathbb{R}^{n_x}$. Then, the empirical risk function $f : \mathbb{R}^{n_x} \to \mathbb{R}$ is also strongly convex and the approaches described in 1.2 can be utilized to minimize it. In case of certain types of function g, function ℓ, and noise ε_i, the minimum point of the empirical risk $\operatorname{argmin} f(\mathbf{x})$ converges to the true value of the parameter $\mathbf{x}_\star$ with $N \to \infty$. Particular examples of a strictly convex loss functions are: quadratic $\ell(y, g(\boldsymbol{\varphi}, \mathbf{x})) = (y - g(\boldsymbol{\varphi}, \mathbf{x}))^2$, absolute with ℓ_2-regularization $\ell(y, g(\boldsymbol{\varphi}, \mathbf{x})) = |y - g(\boldsymbol{\varphi}, \mathbf{x})| + \lambda\|\mathbf{x} - \mathbf{x}_0\|^2$, logistic loss function $\ell(y, g(\boldsymbol{\varphi}, \mathbf{x})) = -y \log g(\boldsymbol{\varphi}, \mathbf{x}) - (1 - y)(1 - \log g(\boldsymbol{\varphi}, \mathbf{x}))$. In turn, the solving functions are linear by $\mathbf{x}$: $g(\boldsymbol{\varphi}, \mathbf{x}) = \sum\limits_{j=1}^{n_x} \mathbf{x}_j g_j(\boldsymbol{\varphi})$, and logistic: $g(\boldsymbol{\varphi}, \mathbf{x}) = (1 + \exp(-\sum\limits_{j=1}^{n_x} \mathbf{x}_j g_j(\boldsymbol{\varphi})))^{-1}$ functions. Systematic exposition of loss functions and solving functions is given in [4; 5; 60].

The key property of decision functions — machine learning models — is the generalization ability, which characterizes the complexity of the relationships that the corresponding model can describe (the concept of generalization ability is described in details in [61]). Despite choosing different decision function type, one of the main approaches to increase the generalization ability is to increase the number of parameters, that is, the dimension of the parameter space n_x. Thus, the strongly convex functions of of high dimension minimization problem is relevant in the context of the pattern recognition problem.

In the case of a quadratic loss function with a linear over $\mathbf{x}$ decision function, the equation (1.1) has the following form:

$$f_{MHK}(\mathbf{x}) = \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{n_x} \mathbf{x}_j g_j(\boldsymbol{\varphi}_i) \right)^2 \to \min_{\mathbf{x}}, \tag{1.2}$$

and the minimum point can be estimated by the least squares method: $\mathbf{x}_{MHK} = \left(\boldsymbol{\Phi}^\top \boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^\top \mathbf{y}$, where $\boldsymbol{\Phi}_{i,j} = g_j(\boldsymbol{\varphi}_i)$ and $\mathbf{y} = (y_1, \ldots, y_n)^\top$. Given the absence of noise and nondegeneration of the matrix $\boldsymbol{\Phi}^\top \boldsymbol{\Phi}$, the solution of the problem (1.2) exists, unique, and matches the true value of the parameter: $\hat{\mathbf{x}}_{MHK} = \operatorname{argmin} f_{MHK} = \mathbf{x}_\star$. If the noises ε_i are additive, centered, identically distributed, has finite variance, are independent of each other and with the inputs $\boldsymbol{\varphi}_j$, then the corresponding estimate is consistent: $\hat{\mathbf{x}}_{MHK} \xrightarrow[N \to \infty]{P} \mathbf{x}_\star$. Thus, under certain conditions on the nature of the noise,

the structure of the model and in the case of a sufficient number of observations, empirical risk minimization can lead to a satisfactory solution.

1.1.2 The case of uncertainties and small number of observations

In the previous section it is demonstrated that with a large number of measurements and certain restrictions on the nature of noise, the solution of the empirical risk minimizaiton probem can lead to a satisfactory solution. However, in the case of a small number of measurements and high uncertainty expressed in the unknown nature of potentially unlimited and systematic noises, the empirical risk minimization can lead to arbitrarily poor estimates and in this sense is of little use. Examples of noises sources are: output measurement errors, inadequacy of the selected model, as well as errors deliberately injected by the system opponent [3]. In such a situation, an alternative approach is to estimate the *confidence region* which contains the true parameter value with a given probability. Note that asymptotic confidence regions which contain the true parameter value with a given probability only with the number of observations tends to infinity. Thus, asymptotic methods are of little interest in case of small number of measurements. Therefore, the problem of constructing a *exact* confidence region which contain the true parameter value exactly with a given probability regardless of the number of observations are particularly relevant. Formally, the problem is posed as follows: for a finite set of observations $\{\boldsymbol{\varphi}_i, y_i\}_{i=1}^N$, constract a set $\mathcal{X}_\alpha$, which contains the true parameter value with a given probability $\alpha \in [0,1]$:

$$P(\mathbf{x}_\star \in \mathcal{X}_\alpha) = \alpha.$$

without significant restrictions on the noise distribution, despite its independency with themselves and inputs $\{\boldsymbol{\varphi}_i\}_1^N$.

1.2 Optimization in high-dimensional spaces

Consider the problem of unconditional optimization

$$f(\mathbf{x}) \to \min_{\mathbf{x}\in\mathbb{R}^n}, \tag{1.3}$$

where f belongs to the set $\mathcal{F}_{\mu,L}$ — μ of strongly convex doubly differentiable functions with Lipschitz gradient:

$$\mu\|\mathbf{x}-\mathbf{y}\|_2^2 \leqslant (\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^\top (\mathbf{x}-\mathbf{y}) \leqslant L\|\mathbf{x}-\mathbf{y}\|_2^2. \tag{1.4}$$

The concept of "high" dimension is defined as follows: the considered algorithms for solving the problem (1.3) must belong to the complexity class $\mathcal{O}(n)$ by the amount of memory used. Such a restriction, for example, makes it impossible to calculate and store the Hesse matrix as well as its full rank approximations.

Note that for functions of the class $\mathcal{F}_{\mu,L}$ the first inequality from the equation (1.4) can be refined.

Proposition 1. *Let $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^d)$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. Then*

$$(\mathbf{x}-\mathbf{y})^\top (\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})) \geqslant \frac{\mu L}{\mu + L}\|\mathbf{x}-\mathbf{y}\|^2 + \frac{1}{\mu+L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2.$$

Proof. See the proof of the Theorem 2.1.12 proof in [62]. □

It is known that the solution of the problem (1.3) exists and unique: $\forall\, f \in \mathcal{F}_{\mu,L}$: $\mathbf{x}_\star = \underset{\mathbf{x}\in\mathbb{R}^n}{\operatorname{argmin}} f(\mathbf{x})$. In cases when the minimum point $\mathbf{x}_\star$ cannot be expressed analytically, or the calculation of the analytical expression is time consuming, a wide range of iterative optimization methods can be used (see, for example [2; 63–65]), which construct a recurrent sequence of estimates $\mathbf{x}_0, \mathbf{x}_1, \ldots$ of the minimum point $\mathbf{x}_\star$.

In the case of a twice differentiable function f, the Newton–Raphson method can be used to solve the problem, defined by the recurrent relation $\mathbf{x}_{t+1} = \mathbf{x}_t - \left[\nabla^2 f(\mathbf{x}_t)\right]^{-1} \nabla f(\mathbf{x}_t)$ (the method and its history are analyzed in details in [66; 67]) and achieve quadratic convergence rate under the additional Lipschitz condition of the Hesse matrix(the convergence rate concept is discussed in Section 1.2.1). The method has a significant drawback: the need to calculate and store the inverse matrix of second derivatives, which makes it inapplicable in high-dimensional spaces. Another widely used method of solving the problem 1.3 is the gradient descent method (the authorship is commonly attributed to Cauchy [68], historical background together with a detailed description are given in [69]), that instead of the inverse Hesse matrix utilizes an experimenter-defined sequence of step sizes α_t: $\mathbf{x}_{t+1} = \mathbf{x}_t + \alpha_t \nabla f(\mathbf{x}_t)$. Despite its simplicity and computational efficiency, the disadvantage of the method is its low convergence rate both in practical and theoretical senses as well as sensitivity to step size selection. In the context of the task 1.3 methods that are in some sense between

the Newton-Raphson and gradient descent methods are of particular interest: superior to the latter in the terms of convergence rate and at the same time applicable in case of large space dimension n.

1.2.1 Optimization algorithms quality estimation

In this section the optimization algorithms quality assessment approaches are considered. A common way to measure an algorithm efficiency is to estimate the number of arithmetic operations required for completion. Owing to potential unboundedness of the number of required iterations and variativity of the complexity of the function and its gradient calculation, the estimate of the number of computational operations is rarely used to measure the quality of iterative optimization methods [62; 70]. Next, we list the most common methods for assessing the optimization method quality.

- *Oracle complexity* — the number of calls to the *Oracle* (an abstraction that, upon request, provides the values of the function and its gradient at a given point) required to achieve a given error value by function. For example, the algorithm $\mathcal{A}$ has complexity $\mathcal{O}\left(N_{\mathcal{A}}(\varepsilon)\right)$ means that to achieve the condition $f(\hat{\mathbf{x}}_\varepsilon) - f_\star < \varepsilon$ it is required to make an order of $\mathcal{O}\left(N_{\mathcal{A}}(\varepsilon)\right)$ calls to the Oracle.
- *Q-convergence* — a characteristic of the convergence rate that is based on the recurrent expression of sequence $f(\mathbf{x}_t) - f(\mathbf{x}_\star)$ or $\|\mathbf{x}_t - \mathbf{x}_\star\|$. A sequence $\|\mathbf{x}_t - \mathbf{x}_\star\| \underset{t}{\to} 0$ converges with
  - *Q-sublinear rate*, if $\lim \frac{\|\mathbf{x}_{t+1} - \mathbf{x}_\star\|}{\|\mathbf{x}_t - \mathbf{x}_\star\|} = 1$;
  - *Q-linear rate*, if $\exists r \in (0, 1) : \lim \frac{\|\mathbf{x}_{t+1} - \mathbf{x}_\star\|}{\|\mathbf{x}_t - \mathbf{x}_\star\|} \leqslant r$;
  - *Q-superlienar rate*, if $\lim \frac{\|\mathbf{x}_{t+1} - \mathbf{x}_\star\|}{\|\mathbf{x}_t - \mathbf{x}_\star\|} = 0$;
  - *Q-quadratic rate*, if $\exists r > 0 : \lim \frac{\|\mathbf{x}_{t+1} - \mathbf{x}_\star\|^2}{\|\mathbf{x}_t - \mathbf{x}_\star\|} \leqslant r$.
- *R-convergence* — a weaker analogue of Q-convergence which characterize the overall rate of convergence, instead of the rate of convergence at each iteration. Sequence $\|\mathbf{x}_t - \mathbf{x}_\star\| \underset{t}{\to} 0$ converges *R-linearly* if $\exists\, r_t$: $\|\mathbf{x}_t - \mathbf{x}_\star\| \leqslant r_t$ and r_t converges to zero Q-linearly. Other types of R-convergence are defined in a similar way.

Further, unless otherwise stated, the we wil use Q-convergence notation omitting ”Q” letter. For example, under certain conditions, Newton–Raphson method has

quadratic convergence, and the gradient descent method — has linear convergence rate (see theorem 3, §4, theorem 1, §5 Chapter 1 in [65], and theorems 1.2.4, 1.2.5 & 2.1.15 in [62]).

1.2.2 Quasi-Newton methods

Quasi-Newton methods iteratively construct the Hesse matrix approximation and for many problems demonstrate a faster convergence rate than the gradient descent method [70]. The step of the quasi-Newton method is $\mathbf{x}_{t+1} = \mathbf{x}_t - \alpha_t \mathbf{H}_t \nabla f(\mathbf{x}_t)$, where $\mathbf{H}_t$ is an approximation of the Hesse matrix, with $\mathbf{H}_{t+1}$ often computed by adding to the previous estimate $\mathbf{H}_t$ an update matrix of rank 1 or 2. The quasi-Newton methods family includes but not limited to: the Davidon–Fletcher–Powell method [14; 18], the Symmetric Rank 1 update method [19], and perhaps the most widely used method Broyden–Fletcher-Goldfarb–Shanno [15; 18] (BFGS). These methods share the disadvantage of Newton–Raphson method: they require calculation and storage of the Hesse matrix or its inverse in memory. Quasi-Newton memory–constrained methods, such as the L-BFGS method [17]) work around this difficulty by directly approximating the product of inverse Hesse matrix and the gradient vector $\left[\nabla^2 f(\mathbf{x})\right]^{-1} \nabla f(\mathbf{x})$ without reconstructing the Hessian itself.

Most quasi-Newton methods are based on *secant equation* — a common tool for finding roots of equations and constructing optimization methods. Consider a differentiable function $f : \mathbb{R}^n \to \mathbb{R}^m$, then its secant equation is the following

$$\nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \approx f(\mathbf{y}) - f(\mathbf{x}), \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

Note, that the secant equation follows from the Taylor decomposition of the function f up to the second elementat point $\mathbf{x}$: $f(\mathbf{y}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + o(\|\mathbf{y} - \mathbf{x}\|)$.

Chord equations play a key role in the construction of quasi-Newton optimization methods. Consider a twice differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, then the secant equation for the gradient ∇f will be:

$$\nabla^2 f(\mathbf{x})(\mathbf{y} - \mathbf{x}) \approx \nabla f(\mathbf{y}) - \nabla f(\mathbf{x}). \tag{1.5}$$

By substituting $\nabla^2 f(\mathbf{x})$ with an unknown matrix $\mathbf{B}_t$ and $\mathbf{x}$, $\mathbf{y}$ with successive estimates of the algorithm, we obtain a system of linear equations with respect to the matrix $\mathbf{B}_t$:

$$\mathbf{B}_t(\mathbf{x}_j - \mathbf{x}_{j-1}) = \nabla f(\mathbf{x}_j) - \nabla f(\mathbf{x}_{j-1}), \quad j = 0, \ldots, t. \tag{1.6}$$

To the best of our knowledge, for the first time the secant equations were used to obtain Hessian in the middle of 1950s (the corresponding work was published only in 1991 [14]). At the moment, many methods that Hessian estimates based on secant equation have been proposed: Davidon–Fletcher–Powell method [14; 18] method, SR-1 [19] method, Broyden–Fletcher–Goldfarb–Shanno [15; 18] (BFGS) method, BFGS method with truncated history [17] and others. Note that chord equations are used not only in quasi-Newtonian optimization methods (see, for example,The Barzilai-Borwein method [71; 72]). In addition, modifications of the secant equations are known which lead to better estimates of the Hesse matrix (see, for example [73; 74]).

Note that the error in the equation (1.5) allows an accurate estimate for Lipschitz gradient, as the following proposition states.

Proposition 2. *Let f bet twice differentiable function with L-Lipschitz Hessian. Then $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ the following holds true:*

$$\|\nabla^2 f(\mathbf{x})(\mathbf{x} - \mathbf{y}) - (\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))\| \leqslant \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2. \tag{1.7}$$

Proof. Directly follows from the formula of finite increments and Lipshitz propoerty of the Hesse matrix. See, for example, Chapter 3, Theorem 3.5, in [70]. □

Note that the system of equations (1.6) can be rewritten in matrix form and supplemented by the requirement of the matrix $\mathbf{B}_t$ symmetry. One of the ways of solving linear matrix equations under the condition of symmetry of the required matrix is proposed in [75]

Theorem 1 (Don, 1987)**.** *Consider known matrices $\mathbf{A} \in \mathbb{R}^{n\times m}$, $\mathbf{B} \in \mathbb{R}^{n\times m}$ and unknown matrix $\mathbf{X} \in \mathbb{R}^{m\times m}$. Then, the following system of linear equations with*

$$\begin{cases} \mathbf{A}\mathbf{X} = \mathbf{B} \\ \mathbf{X} = \mathbf{X}^\top \end{cases} \tag{1.8}$$

has solution with respect to $\mathbf{X}$ if and only if $\exists \mathbf{A}^{\sim} : \mathbf{A}\mathbf{A}^{\sim}\mathbf{B} = \mathbf{B}$ and $\mathbf{A}\mathbf{B}^\top = \mathbf{B}\mathbf{A}^\top$. Any solution can be expressed by the following formula:

$$\begin{aligned} \mathbf{X}_* = {} & \mathbf{A}^{\sim}\mathbf{B} + (\mathbf{I} - \mathbf{A}^{\sim}\mathbf{A})(\mathbf{A}^{\sim}\mathbf{B})^\top \\ & + (\mathbf{I} - \mathbf{A}\mathbf{A}^{\sim})\Theta(\mathbf{I} - \mathbf{A}\mathbf{A}^{\sim}), \end{aligned} \tag{1.9}$$

where $\Theta \in \mathbb{R}^{n\times n}$ *is an arbitrary matrix. Moreover, the minimum norm solution is achieved when* $\Theta = 0$.

Proof. See proof of the Theorem 2 in [75]. □

1.2.3 Conjugate gradients method

The conjugate gradient method proposed for solving systems of linear equations in [8] and extended to solving quadratic optimization problems by [12]. It uses a linear combination of the previous direction and the current gradient value as the step direction:

$$\begin{aligned} \mathbf{d}_t &= -\nabla f\left(\mathbf{x}_t\right) + \beta_t \mathbf{d}_{t-1}, \\ \mathbf{x}_{t+1} &= \mathbf{x}_t + \alpha_t \mathbf{d}_t, \end{aligned} \tag{1.10}$$

where the step size α_t is calculated using a linear search along the direction $\mathbf{d}_t$, and $\mathbf{d}_{-1}$ is assumed to be zero. One of the most common formulas for calculating the β_t coefficient is the Polyak–Ribiere–Polak formula, proposed independently in [10] and [11]:

$$\beta_t^{PR} = \frac{\nabla f\left(\mathbf{x}_t\right)^\top \left(\nabla f\left(\mathbf{x}_t\right) - \nabla f\left(\mathbf{x}_{t-1}\right)\right)}{\nabla f\left(\mathbf{x}_{t-1}\right)^\top \nabla f\left(\mathbf{x}_{t-1}\right)}. \tag{1.11}$$

Note that the formula (1.11) — is widely used but not the only way to calculate the β_t coefficient, initially the of Fletcher–Reeves formula [9] was used and subsequently many alternative variants of calculation of coefficient were proposed (the detailed list is given in work [76], sections 3 and 4). Despite the abundance of alternatives, the Polak–Ribiere–Polak formula is a standard way to calculate the coefficient β_t (see section 5.2 in [70]). In this regard, the formula (1.11) will be used to calculate the β_t coefficient unless otherwise specified.

Several important properties of the conjugate gradient method are stated in the following proposition

Proposition 3. *Let* $f \in \mathcal{F}_{\mu,L} : \mathbb{R}^d \to \mathbb{R}$ *— a quadratic function with Hesse matrix* $\nabla^2 f \equiv \mathbf{A}$. *Then, the conjugate gradient method converges to the optimum point in at*

most d steps with linear rate:

$$(\mathbf{x}_{t+1} - \mathbf{x}_\star)\mathbf{A}(\mathbf{x}_{t+1} - \mathbf{x}_\star) \leqslant \left(\frac{\frac{L}{\mu} - 1}{\frac{L}{\mu} + 1}\right)(\mathbf{x}_t - \mathbf{x}_\star)\mathbf{A}(\mathbf{x}_t - \mathbf{x}_\star),$$
$$\|\mathbf{x}_t - \mathbf{x}_\star\| = 0, \quad t \geqslant d.$$

Herewith,

- $\mathbf{x}_{t+1}$ *is the minimum point of* f *along a set* $\mathbf{x}_0 + \operatorname{span}\{\nabla f(\mathbf{x}_0), \ldots, \nabla f(\mathbf{x}_t)\}$*;*
- *directions* $\{\mathbf{d}_j\}$ *are mutualy conjugate with respect to matrix* $\mathbf{A}$*:* $\mathbf{d}_i\mathbf{A}\mathbf{d}_j = 0$*,* $\forall\, 0 \leqslant i < j < d$*.*

Proof. See proof of Theorems 5.4 & 5.5 in [70]. □

The later property of the $\{\mathbf{d}_j\}$ directions conjugacy is the key one, which is reflected in the name of the method. If the function f can not be described by a quadratic polynomial this property is violated together with the rest components of the Statement 3. A similar problem arises in the case of calculation errors unavoidable in practice: due to the construction of the method, errors tend to accumulate from iteration to iteration, thereby violating the conjugacy property.

A common way to deal with the violation of the conjugacy property is *restart technique*: restarting the algorithm by setting the next direction to the gradient direction every n iterations [9]. It is known that for a sufficiently wide class of functions, the use of restarts leads to n–quadratic convergence (see Theorem 1 in [13]), which is confirmed by practical results. This approach has several drawbacks: the restart in the direction of $-\nabla f(\mathbf{x}_t)$ does not take into account the accumulated information about the curvature of the function, it leads to a smaller *immediate* reduction of the function and, most importantly in the context of the task, the restart technique is not applicable in high-dimensional spaces, since the space dimension n exceeds the desired number of iterations.

Beale-Powell restart method [77] is an alternative approach which uses modified formula for direction $\mathbf{d}_t$ calculation:

$$\mathbf{d}_t = -\nabla f(\mathbf{x}_t) + \beta_t \mathbf{d}_{t-1} + \gamma_t \mathbf{d}_k. \tag{1.12}$$

The third summand $\gamma_t \mathbf{d}_k$ plays the role of a restart in the sense that the conjugacy property is preserved only for directions after the step k: $\{\mathbf{d}_j\}_{k+1 \leqslant j \leqslant t}$. Restart — updating

the k index to $t-1$ — is performed when one of the inequalities is executed:

$$\begin{gathered} \left|\nabla f\left(\mathbf{x}_t\right)^{\top} \nabla f\left(\mathbf{x}_{t-1}\right)\right| \geqslant c_1\|\nabla f\left(\mathbf{x}_t\right)\|^2, \\ c_2\|\nabla f\left(\mathbf{x}_t\right)\|^2 \leqslant -\mathbf{d}_t^{\top} \nabla f\left(\mathbf{x}_t\right) \leqslant c_3\|\nabla f\left(\mathbf{x}_t\right)\|^2. \end{gathered} \tag{1.13}$$

The first inequality corresponds to violation of the orthogonality property of subsequent gradients, and the second one corresponds to deviation of the direction $\mathbf{d}_t$ from the negative gradient direction of the function f. Note that the coefficients c_1, c_2, c_3 are usually set to 0.2, 0.8 and 1.2, respectively. Also note that for the Beale-Powell method there are convergence guarantees. However, it demonstrates good results in practice, including high-dimensional optimization problems [78].

1.2.4 Sequential subspace optimization

In the works [23; 24], the *sequential subspace optimization* (SSO, also known as iterative subspace minimization) was proposed, which is especially relevant in the context of high-dimensional problems. The idea of the method is to consequently apply two operations:

1. construction of subspace $\mathcal{D}_t \subset \mathbb{R}^n$ of small dimension: $|\mathcal{D}_t| = m_t << b$;
2. optimize the target function along the selected subspace: $\mathbf{x}_{t+1} = \underset{\mathbf{d} \in \mathcal{D}_t}{\operatorname{argmin}} f(\mathbf{x}_t + \mathbf{d})$.

Newton–Raphson method is used to solve the minimization problem along the selected subspace $\mathcal{D}_t$ — *subspace optimization problem*. As generators for the subspace $\mathcal{D}_t$ the authors propose to use: gradient values $\nabla f(\mathbf{x}_i)$, $i = t, \ldots, 0$, preceding directions $\mathbf{x}_i - \mathbf{x}_{i-1}$, as well as the following directions [79]:

$$\begin{gathered} \mathbf{d}_t^{(1)} = \mathbf{x}_t - \mathbf{x}_0, \qquad \mathbf{d}_t^{(2)} = \sum_{i=0}^{t} w_i \nabla f\left(\mathbf{x}_i\right), \\ w_i = \frac{1}{2} + \sqrt{\frac{1}{4} + w_{t-1}^2}, \qquad w_0 = 1. \end{gathered} \tag{1.14}$$

The method is supported by the guarantee of its sublinear convergence rate ($\mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}}\right)$ in Oracle complexity) for a class of smooth convex functions with Lipschitz gradient.

Similar ideas have been used in various preceding works on optimization: in the quadratic case, the conjugate gradient method implicitly finds the minimum of a function along the corresponding Krylov subspaces [80], in [81], an optimization method

was proposed in a gradually increasing "essential" subspace, in in [79], a method of sequential optimization along subspaces formed by a given gradient value and the so-called "Nemirovsky directions" (1.14), in [20; 21] a modifications to the gradient descent method is proposed that further minimize the target function along the preceding gradients. Moreover, many optimization methods can be interpreted in the context of the sequential subspace optimization approach if the minimization of the target function in the 2nd step of the SSO scheme is replaced by a surrogate minimization step. For example, gradient descent can be viewed as a SSO method with the subspace formed by the current gradient value $\mathcal{D}_t = \operatorname{span}\{\nabla f(\mathbf{x}_t)\}$ and the surrogate model $q_t(z) = z + \frac{1}{2\lambda_t} z^2$. The conjugate gradient method in the case of a quadratic objective function corresponds to the SSO method with subspaces of the form $\mathcal{D}_t = \operatorname{span}\{\mathbf{d}_t, \nabla f(\mathbf{x}_t)\}$ and an exact solution of the subspace optimization problem. Many quasi-Newtonian methods such as SR1 [19], PSB [82], and methods from the Broyden family [70] use the subspace $\mathcal{D}_t = \operatorname{span}\{\nabla f(\mathbf{x}_0), \ldots, \nabla f(\mathbf{x}_t)\}$ (see theorem 2.1 In [83]). However, it was in [23] that sequential subspace optimization was formulated explicitly and generically.

Worth highlighting works of Yuxian Yuan and co-authors on *subspace optimization methods*. Thus, in [84], authors proposed a method of sequential function minimization along the current gradient and the previous step direction thus generalizing the conjugate gradient method and demonstrated its convergence under sufficiently weak conditions with the use of the Wolfe conditions [85; 86] to select the step size. In [87], a general model of subspace optimization methods is formulated. A span of the current gradient value and the truncated history of the preceding directions $\{\nabla f(\mathbf{x}_t), \mathbf{s}_{t-1}, \ldots, \mathbf{s}_{t-m}\}$ are used as subspace $\mathcal{D}_t$. To minimize along the subspace $\mathcal{D}_t$ the quadratic surrogate $q_t(\mathbf{d}) = \nabla f(\mathbf{x}_t)^\top \mathbf{d} + \frac{1}{2}\mathbf{d}^\top \mathbf{B}_t \mathbf{d}$ is used, where the Hessian approximation $\mathbf{B}_t$ is estimated from the secant equations. A variation for the case constrained case is also considered. In [83], a quasi-Newton method using confidence regions is proposed. In [88] the proposed model is extend to the problem of solving a system of nonlinear equations, and in [25] the mentioned results are presented in a generalized and systematized form.

Sequential subspace optimization methods are actively applied in practice. In [26] an SSO method is proposed in context of the image restoration problem, where the subspace problem is solved by minimizing the quadratic majorant (surrogate) objective function specific to the particular image restoration task. In [27], the SSO method is used to solve a particular pattern recognition problem: estimating the support vector machine algorithm parameters [4]. The paper [30] discusses aspects of the application

of optimization methods to image processing problems, experimentally demonstrates the superiority of the sequential subspace optimization method over alternatives for image reconstruction and blur removal problems.

The sequential subspace optimization approach is actively used as a tool for methods construction. For example, in [28], the sequential subspace optimization approach is adapted to find an approximation of the posteriori distribution minimizing the Kulbak-Leibler divergence. In [29] in the context of solving the stochastic optimization problem, authors proposed to accelerate the chosen stochastic optimization method by regular iterations of the SSO method with a subspace formed from the previous values of gradients, Nemirovsky directions (1.14) and vectors between the current estimate and so-called *anchor points* — fixed previous point estimated. The effectiveness of the proposed modification is demonstrated on the problems of image classification: MNIST [89] and CIFAR-10 [90] with deep learning methods.

The projection-approximation-reconstruction approach in optimization is a special case of sequential subspace optimization. The main difference lies in the way the q_t surrogate is constructed: in the above mentioned works [22; 24; 25; 71; 83; 87] the surrogate is constructed analytically, while in the projection-approximation-reconstruction approach the surrogate coefficients are approximated based on the solution of the regression problem. The parameterized surrogate $q_t(\mathbf{z}) = q(\mathbf{z}|\theta_t)$ is considered and the parameter θ is estimated based on the following minimization problem:

$$\sum_{i=t}^{t-m+1} \left(f(\mathbf{x}_i) - q(\mathbf{D}_t\mathbf{x}_i|\theta)\right)^2 \to \min_{\theta}.$$

In [52; 53], the projection–approximation–reconstruction approach is used to accelerate gradient descent: after every m iterations of the gradient descent method, a parametrized surrogate $q(\cdot|\theta_t)$ is constructed based on the last m point estimates $\{\mathbf{x}_i\}_{t-m+1}^{t}$ and function values $\{f(\mathbf{x}_i)\}_{t-m+1}^{t}$. This surrogate and minimum is used as the next estimate $\mathbf{x}_{t+1}$. In[54] four algorithms based on sequential application of the following operations:

1. constructing projection matrix $\mathbf{D}_t \in \mathbb{R}^{n\times m}$;
2. approximation function f surrogate $q(\cdot|\theta_t)$ at the points $\{\mathbf{x}_i, f(\mathbf{x}_i)\}_{t-m+1}^{t}$;
3. reconstruct the function f minimum from the obtained surrogate q_t.

To construct the matrix $\mathbf{D}_t$, both the previous values of the gradient $\{\nabla f\,(\mathbf{x}_i)\}_{i=0}^{t}$ and and a random vector are used. The next estimate $\mathbf{x}_{t+1}$ is obtained directly from the minimum point $\operatorname{argmin} q_t$, where a quadratic function is used as the surrogate.

1.3 Estimation of confidence regions under conditions of uncertainty and a finite number of samples

Consider the problem of estimating the parameter of a linear input-output model from observations

$$y_i = \boldsymbol{\varphi}_i^T \mathbf{x}_\star + \varepsilon_i, \quad i = 1..N, \tag{1.15}$$

where N — number of measurements $y_i \in \mathbb{R}$ — observed system outputs, $\boldsymbol{\varphi}_i \in \mathbb{R}^n$ — observed system inputs, $\mathbf{x}_\star \in \mathbb{R}^n$ — unknown parameter vector, and $\varepsilon_i \in \mathbb{R}$ — unknown additive noise.

With a large number of observations and favorable conditions on noises ε_i, a satisfactory estimate $\hat{\mathbf{x}}$ of the true value of the parameter $\mathbf{x}_\star$ can be obtained by regression analysis. Many of the regression methods, in turn, can be formulated as an optimization problem. However, in the case of a small number of measurements, a low signal-to-noise ratio, and an unknown noise distribution, the resulting estimates of $\hat{\mathbf{x}}$ may differ significantly from the true value of the parameter and in this sense are unreliable. An alternative approach is to construct a *confidence region* containing the true parameter value with a given probability. Of particular interest is the problem of constructing a *exact* confidence set that contains the true value of the parameter exactly with a given probability regardless of the number of observations — as opposed to an asymptotic confidence set that contains the true value of the parameter with a given probability only when the number of observations tends to infinity. Note that the source of randomness can be due too both noises and inputs of the system (1.15).

Consider the following types of noise and corresponding approaches to determining confidence regions construction. For the case of independently and equally normally distributed noises, an exact formula for confidence region of parameter $\mathbf{x}$ is known [91; 92]. In the case of a large number of measurements and an noise distribution different from normal, similar procedures will result in a *asymptotic* confidence region. The methods of constructing confidence region based on bootstrapping technique [93–95] have similar disadvantages. In the case of uniformly bounded noise: $\exists C > 0$: $|\varepsilon| \leqslant C$ it is possible to construct a region contains the true value of the parameter *exactly* (so-called *set membership approach*) [96; 97], which can be considered as an exact confidence set containing the true parameter value with probability 1. For the case of i.i.d noises distributed symmetrically with respect to zero, there are several

methods for constructing exact confidence regions: the method of exclusion of sign-dominant correlation regions (LSCR, leave-out sign-dominant correlation regions [40; 98]) for the one-dimensional case and its generalization to the multidimensional case the sign-perturbed sums method (SPS,sign-perturbed sums [41; 99; 100]). Of particular importance is [42], where a modification of the LSCR method was proposed for the case of unknown noises but controlled inputs of a linear control plant.

1.3.1 Normally distributed noise

Under the assumption of normality, independence and the same noise distribution $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ with unknown variance σ^2 the exact confidence region of the parameter $\mathbf{x}_\star$ can be constructed using the Fisher distribution. Let $\alpha \in [0,1]$, then $P\left(\mathbf{x}_\star \in \mathcal{X}_\alpha^{\mathcal{N}}\right) = \alpha$, where the confidence region $\mathcal{X}_\alpha^{\mathcal{N}}$ is defined by the following formula (see Chapter 5 in [92]):

$$\mathcal{X}_\alpha^{\mathcal{N}} = \left\{ \mathbf{x} \in \mathbb{R}^N : \ \frac{1}{N-n}(\mathbf{x} - \hat{\mathbf{x}}_{MHK})^T \boldsymbol{\Xi}_{\mathbf{x}}^{-1} (\mathbf{x} - \hat{\mathbf{x}}_{MHK}) \leqslant F_\alpha(n, N-n) \right\}, \tag{1.16}$$

where $F_\alpha(n, N-n)$ — α-quantile of the Fisher distribution with n and $N-n$ degrees of freedom, $\hat{\mathbf{x}}_{MHK} = \left(\boldsymbol{\Phi}^\top \boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^\top \mathbf{y}$ — $\mathbf{x}_\star$ least squares estimate, $\boldsymbol{\Phi} = [\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_N]^\top$ — system inputs matrix, $\mathbf{y} = (y_1, \ldots, y_N)$ — system outputs vector, $\hat{\boldsymbol{\Xi}}_{\mathbf{x}} = \hat{\sigma}^2 \left(\Phi^\top \Phi\right)^{-1}$ — covariance matrix of the $\hat{\mathbf{x}}_{MHK}$ estimate, $\hat{\sigma}^2 = \frac{1}{N-n} \|\mathbf{y} - \boldsymbol{\Phi} \hat{\mathbf{x}}_{MHK}\|^2$ — unbiased estimate of σ^2.

1.3.2 Symmetrically distributed noise

Consider the following definition of *symmetric* multidimensional random variable.

Definition 1. *Let $(\Omega, \mathcal{F}, P)$ be a probability space, then multidimensional random variable $\xi : \Omega \to \mathbb{R}^n$ is called symmetric if:*

$$\forall\, A \in \mathcal{F} : \ P(\xi \in A) = P(-\xi \in A). \tag{1.17}$$

Suppose that the noises and inputs of the system $\{\varepsilon_1,\dots,\varepsilon_N,\boldsymbol{\varphi}_1,\dots,\boldsymbol{\varphi}_N\}$ are mutually independent, and additionally, noises are symmetrically distributed with respect to zero: $\varepsilon_i \sim -\varepsilon_i \,\forall\, i = 1..N$. In this case, the *sign–perturbed sums* (SPS) method [41] can be used to construct the exact confidence region of the parameter $\mathbf{x}_\star$:

$$\begin{gathered}\mathcal{X}_{M,q}^{SPS} = \left\{\mathbf{x} : |\{k \in \{1,\dots,M-1\} : \|\tilde{S}_0(\mathbf{x})\| < \|\tilde{S}_k(\mathbf{x})\|\}| \geqslant q\right\},\\ \tilde{S}_0(\mathbf{x}) = \sum_{i=1}^{N} \boldsymbol{\varphi}_i(y_i - \boldsymbol{\varphi}_i^T\mathbf{x}), \qquad \tilde{S}_k(\mathbf{x}) = \sum_{i=1}^{N} a_{k,i}\boldsymbol{\varphi}(y_i - \boldsymbol{\varphi}_i^T\mathbf{x}),\\ a_{k,i} = \begin{cases} 1 & \text{with probability } \frac{1}{2} \\ -1 & \text{with probability } \frac{1}{2} \end{cases}, \qquad k = 1,\dots,M-1.\end{gathered} \tag{1.18}$$

The following property holds true for the $\mathcal{X}_{M,q}^{SPS}$.

Proposition 4 (Csaji, Campi, Weyer, 2012)**.** *In the assumptions imposed above $\mathcal{X}_{M,q}^{SPS}$ is an exact confidence region for $\mathbf{x}_\star$:*

$$P\left(\mathbf{x}_\star \in \mathcal{X}_{M,q}^{\hat{CAN}}\right) = 1 - \frac{q}{M}.$$

Proof. See proof of Theorem 1 in [41]. □

Proposition 4 is substantially based on the following result

Proposition 5 (Csaji, Campi, Weyer [41])**.** *Let ξ be a symmetric multidimensional random variable and α — random sign (see equation (1.18)). Then random variables ξ and $\alpha\xi$ are independent*

Proof. See proof of Lemma 1 in [41]. □

For the next result, let us introduce the definition of *uniformly ordered* random variables.

Definition 2. *Random variables $\{\xi_i\}_1^n$ are called uniformly ordered, if for any permutation π of the set $\{1,\dots,n\}$ corresponding order $\{\xi_i\}_1^n$ is equiprobable:*

$$P(\xi_{\pi(1)} < \ldots < \xi_{\pi(n)}) = \frac{1}{n!}.$$

Proposition 6 (Csaji, Campi, Weyer [41])**.** *Let $\{\xi_i\}_1^n$ be i.i.d symmetric contnious random variables. Then they are uniformly bounded.*

Proof. See proof of Lemma 4 in [41]. □

Note that for $\mathcal{X}_{M,q}^{SPS}$ the following property always holds true: $\hat{\mathbf{x}}_{LS} \in \mathcal{X}_{M,q}^{SPS}$.

1.3.3 Arbitrary noise

Consider system (1.15), assume that $\{\boldsymbol{\varphi}_i\}_{i=1}^N$ are random vectors and together with noises $\{\varepsilon_i\}_{i=1}^N$ they satisfy the following assumptions and otherwise arbitrary:

– noises and system inputs $\{\varepsilon_1,\ldots,\varepsilon_N,\boldsymbol{\varphi}_1,\ldots,\boldsymbol{\varphi}_N\}$ are mutually independent;
– system inputs $\boldsymbol{\varphi}_1,\ldots,\boldsymbol{\varphi}_N$ are symmetrically and continiously distributed around the known mean vector $m_\varphi \in \mathbb{R}^n$.

Let us state the task of constructing an exact confidence region for the parameter $\mathbf{x}_\star$: for observations $\{\boldsymbol{\varphi}_i, y_i\}_1^N$ satisfying the above conditions and given $\alpha \in [0,1]$, construct a set $\mathcal{X}_\alpha \in \mathbb{R}^n$, such that:

$$P(\mathbf{x}_\star \in \mathcal{X}_\alpha) = \alpha. \tag{1.19}$$

It should be noted that some of the original results presented in the thesis have been further developed in the works of other authors. In example, the method described in Section 2.4 for obtaining the exact confidence region of a linear model parameter under almost arbitrary noise (1.19) have been generalized to the nonlinear case and applied to the problem of determining the confidence interval of the incubation time of material destruction in the Volkova M. V. thesis [101].

Chapter 2. Sequential subspace optimization and modified sign–perturbed sums methods

2.1 Properties of the SSO methods

2.1.1 General scheme of the SSO methods

Sequentional subspace optimization methods are to the sequential application of the following two operations:

1. susbspace formation $\mathcal{D}_t \subset \mathbb{R}^n$, dimension of which is significantly smaller than dimension of the original one $|\mathcal{D}_t| = m_t \ll n$;
2. approximations of the minimum of the target function along this subspace with respect to the current value $\mathbf{x}_t$:

$$\mathbf{x}_{t+1} \approx \underset{\mathbf{x} \in \{\mathbf{x}_t + \mathbf{d} \,:\, \mathbf{d} \in \mathcal{D}_t\}}{\operatorname{argmin}} f(\mathbf{x}). \tag{2.1}$$

In this case, both the subspace $\mathcal{D}_t$ and the method of finding the minimum f along the subspace can vary. Common vectors for the $\mathcal{D}_t$ subspace formation are: the current and previous values of the gradient $\nabla f(\mathbf{x}_k)$, changes in the gradient $\mathbf{y}_k = \nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}_{k-1})$ and the argument $\mathbf{s}_k = \mathbf{x}_k - \mathbf{x}_{K-1}$, as well as their combinations. As a method of solution of (2.1) a surrogate approach is oftenly used, in which a simpler function-surrogate q_t is minimized which in is close to the function f in the neighborhood of $\mathbf{x}_t$. One of the most common types of surrogate is the quadratic function $q_t(\mathbf{d}) = \nabla f(\mathbf{x}_t)^\top \mathbf{d} + \mathbf{d}^\top \mathbf{B}_t \mathbf{d}$, where matrix $\mathbf{B}_t$ is the Hesse matrix aproximation. It is worth noting that the step size along the found direction $\mathbf{d}_t$ are usually calculated with linear search methods.

In the following a general scheme of sequential subspace optimization methods is presented.

0. The target function f and the starting point $\mathbf{x}_0$, $t \leftarrow 0$ are given.
1. New subspace $\mathcal{D}_t$ is constructed.
2. The f function approximation along the subspace $\mathcal{D}_t$ in the $\mathbf{x}_t$ point a neighbourhood of is constructed: $q_t(\mathbf{d}) \sim f(\mathbf{x}_t + \mathbf{d})$, $\mathbf{d} \in \mathcal{D}_t$.

3. Subspace minimization problem is solved:

$$\mathbf{d}_t = \underset{\mathbf{d}\in\mathcal{D}_t}{\operatorname{argmin}}\, q_t(\mathbf{d}). \tag{2.2}$$

4. Line search is performed along the calculated direction: $\alpha_t = \underset{\alpha}{\operatorname{argmin}}\, f(\mathbf{x}_t + \alpha\mathbf{d}_t)$.
5. Next estimate is calculated: $\mathbf{x}_{t+1} = \mathbf{x}_t + \alpha_t\mathbf{d}_t$.
6. $t \leftarrow t+1$. Go to 1.

This scheme coincides with the one presented in [87] (see algorithm 3.1), except for the following: presence of a restriction on the norm of the step size, the target function decreasing monotony, and absence of the linear search step. Note that the linear search in step (4) can be performed in various ways: based on Wolfe's [85] conditions, numerical minimization, or based on a deterministic rule (for example, $\alpha_t \equiv 1$). Based on the above scheme, the following factors are determining for SSO methdos:

a. method for subspace construction $\{\mathcal{D}_t\}_{t\geqslant 0}$,
b. method for construction and minimization of the surrogate q_t.

It is worth noting that often the approximator q_t and the subspace $\mathcal{D}_t$ are not explicitly specified and the direction vector $\mathbf{d}_t$ is written out for reasons different from minimizing the surrogate along some subspace.

In practice, it is convenient to operate not with subspaces $\mathcal{D}_t$ but with sets of vectors forming them. Denote $\mathbf{D}_t = \left[\mathbf{d}_1^{(t)}, \ldots, \mathbf{d}_{m_t}^{(t)}\right] \in \mathbb{R}^{n\times m_t}$ — a matrix whose columns form the set $\mathcal{D}_t$. Then, the SSO scheme can be rewritten is rewritten in explicit projections form:

0. Set target fucntion f, initial point $\mathbf{x}_0$ and initial subspace matrix $\mathbf{D}_0$; $t \leftarrow 0$.
1. Construct the next subspace matrix $\mathbf{D}_t$.
2. Construct an approximation of the function f along the subspace $\mathcal{D}_t$ in the neighborhood of the point $\mathbf{x}_t$: $q_t(\mathbf{z}) \sim f(\mathbf{x}_t + \mathbf{D}_t\mathbf{z})$, $\mathbf{z} \in \mathbb{R}^{m_t}$.
3. Solve the subspace optimization problem:

$$\mathbf{z}_t = \underset{\mathbf{z}\in\mathbb{R}^{m_t}}{\operatorname{argmin}}\, q_t(\mathbf{z}). \tag{2.3}$$

4. Perform a linear search along the obtained direction: $\alpha_t = \underset{\alpha}{\operatorname{argmin}}\, f(\mathbf{x}_t + \alpha\mathbf{d}_t)$.
5. Calculate the next point: $\mathbf{x}_{t+1} = \mathbf{x}_t + \alpha_t\mathbf{d}_t = \mathbf{x}_t + \alpha_t\mathbf{D}_t\mathbf{z}_t$.
6. $t \leftarrow t+1$. Go to 1.

In a stricter form, this scheme is represented in the Algorithm 1. It is worth noting that the subspaceUpdate and subspaceSearch functions can store the state: previous values of points, gradients, their linear combinations, etc.

Algorithm 1 SSO_scheme($\mathbf{x}_0$, subspaceUpdate, subspaceSearch, lineSearch)

1: $\mathbf{D}_t \leftarrow$ subspaceUpdate($\mathbf{D}_{t-1}, \nabla f(\mathbf{x}_t)$)
2: $\mathbf{z}_t \leftarrow$ subspaceSearch($f, \mathbf{D}_t, \mathbf{x}_t$)
3: $\mathbf{d}_t \leftarrow \mathbf{D}_t\mathbf{z}_t$
4: $\alpha_t \leftarrow$ lineSearch($f, \mathbf{x}_t, \mathbf{d}_t$)
5: $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t + \alpha_t\mathbf{d}_t$
6: $t \leftarrow t + 1$; go to 1

The following proposition states that SSO schemes with explicit and implicit projections are equivalent.

Proposition 7. *Let* $\mathcal{D}_t = \operatorname{span}\left\{\mathbf{d}_t^{(1)}, \ldots, \mathbf{d}_t^{(m)}\right\}$, *and* $\mathbf{D}_t = \left[\mathbf{d}_t^{(1)}, \ldots, \mathbf{d}_t^{(m)}\right]$ — *matrix, whose columns form the* $\mathcal{D}_t$ *subspace. The the surrogate minimization problem* $\operatorname{argmin}_{\mathbf{d}\in\mathcal{D}_t} q_t(\mathbf{d})$ *is equivalent to the following one:* $\operatorname{argmin}_{\mathbf{z}\in\mathbb{R}^m} q_t(\mathbf{D}_t\mathbf{z})$.

As in the implicit projection scheme, various linear search procedures can be used in step (4). It is worth noting that the use of the matrix $\mathbf{D}_t$ instead of the subspace $\mathcal{D}_t$ not only simplifies the practical implementation of the algorithms, but allows an easier properties of the SSO methods properties and can also reduce the computational complexity of the algorithm. Indeed, the subspace optimization problem in the explicit statement (2.3) is intuitively easier task than the implicit one (2.2) becaouse of the problem statement: instead of the n-dimensional optimization problem with constraints, the m-dimensional optimization problem without constraints (depending on the type q_t) is used. The following remark quantifies that difference.

Remark 1. *Let* q_t *be a quadratic surrogate. Then for the implicit formulation* (2.2) *it is required to use an order of* $\mathcal{O}\left(n^2\right)$ *in memory and an order of* $\mathcal{O}\left(n^3\right)$ *the a number of operations, while for the explicit formulation* (2.3) *it is required to use only an order of* $\mathcal{O}\left(m^2\right)$ *in memory and an order of* $\mathcal{O}\left(m^3\right)$ *in a number of operations in the worst case:* $q_t(\mathbf{D}_t\mathbf{z}) = \nabla f(\mathbf{x}_t)^\top \mathbf{D}_t\mathbf{z} + \mathbf{z}\mathbf{D}_t^\top\mathbf{B}_t\mathbf{D}_t\mathbf{z} = \mathbf{r}_t^\top\mathbf{z} + \mathbf{z}^\top\mathbf{Q}_t\mathbf{z}$, $\mathbf{r}_t \in \mathbb{R}^m$, $\mathbf{Q}_t \in \mathbb{R}^{m\times m}$.

Note the following property common to all SSO methods: the accuracy of any sequential subspace optimization method is limited by the quality of subspace selection, regardless of the subspace optimization problem soltion and step size selection quality.

Remark 2. *Consider the following sequential subspace optimization process* $\mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{d}_t$*, where* $\mathbf{d}_t \in \mathcal{D}_t$*. Then*

$$\|\mathbf{x}_t - \mathbf{x}_\star\| \geqslant \|\mathbf{x}_0 - \mathbf{x}_\star\| - \|\mathcal{P}_{\cup_{0\leqslant j\leqslant t}\mathcal{D}_j}(\mathbf{x}_0 - \mathbf{x}_\star)\|,$$

where $\mathcal{P}_{\cup_{0\leqslant j\leqslant t}\mathcal{D}_j}(\mathbf{x})$ *— projection of a vector* $\mathbf{x}$ *onto set* $\cup_{0\leqslant j\leqslant t}\mathcal{D}_j$*. Moreover, if the* $\mathbf{D}_j$ *columns form an orthonormal basis of the* $\mathcal{D}_j$ *subspace* $\forall\ j$*, then:*

$$\|\mathbf{x}_t - \mathbf{x}_\star\| \geqslant \|\prod_{j=0}^{t}\left(\mathbf{I} - \mathbf{D}_j\mathbf{D}_j^\top\right)(\mathbf{x}_0 - \mathbf{x}_\star)\|.$$

Proof. Finally, it is sufficient to note that $\mathbf{x}_t - \mathbf{x}_0 \in \operatorname{span}\{\mathbf{d}_0, \ldots, \mathbf{d}_t\}$. □

Remark 2 describes a lower bound on the residual of the sequential subspace optimization process.

2.1.2 Quadratic case

In this section several properties of the sequential subspace optimization methods will be illustrated with a basic example. Consider a quadratic target function:

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top\mathbf{A}\mathbf{x} + \mathbf{b}^\top\mathbf{x} + c, \tag{2.4}$$

where $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{A} \succ 0$, $\mathbf{b} \in \mathbb{R}^n$, $c \in \mathbb{R}$, coefficients $\mathbf{A}$, $\mathbf{b}$, c are unknown and denote $\mathbf{x}_\star = \mathbf{A}^{-1}\mathbf{b} = \operatorname{argmin} f$.

Fix matrix $\mathbf{D}_t \in \mathbb{R}^{n\times m}$, $1 \leqslant m \leqslant n$ è $\mathbf{x}_t \in \mathbb{R}$. And consider the f minimization problem along the subspace $\mathcal{D}_t = \{\mathbf{x}_t + \mathbf{D}_t\mathbf{z} \ :\ \mathbf{z} \in \mathbb{R}^m\}$:

$$f\left(\mathbf{x}_t + \mathbf{D}_t\mathbf{z}\right) \to \min_{\mathbf{z}}. \tag{2.5}$$

Consider Taylor series expansion of the function $f(\mathbf{x}_t + \mathbf{D}_t\mathbf{z})$ in point $\mathbf{x}_t$:

$$\begin{aligned} f(\mathbf{x}_t + \mathbf{D}_t\mathbf{z}) = f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top\mathbf{D}_t\mathbf{z} \\ + \frac{1}{2}\mathbf{z}^\top\mathbf{D}_t^\top\nabla^2 f(\mathbf{x}_t)\mathbf{D}_t\mathbf{z} + o(\|\mathbf{D}_t\mathbf{z}\|_2^2). \end{aligned} \tag{2.6}$$

Assume that surrogate q_t equals to the Taylor's expansion to the second power inclusive:

$$q_t(\mathbf{z}) = f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top\mathbf{D}_t\mathbf{z} + \frac{1}{2}\mathbf{z}^\top\mathbf{D}_t^\top\nabla^2 f(\mathbf{x}_t)\mathbf{D}_t\mathbf{z}. \tag{2.7}$$

Then (assuming that the $\mathbf{D}_t$ matrix is non-degenerate) the surrogate minimum is achieve at the point: $\mathbf{z}_t = \operatorname{argmin}_{\mathbf{z}} q_t = -\left(\mathbf{D}_t^\top \mathbf{A}\mathbf{D}_t\right)^{-1} \mathbf{D}_t^\top \nabla f\left(\mathbf{x}_t\right)$. Thus,

$$\mathbf{H}_t := \mathbf{D}_t \left(\mathbf{D}_t^\top \mathbf{A}\mathbf{D}_t\right)^{-1} \mathbf{D}_t^\top. \tag{2.8}$$

— an approximation of the inverse to the Hesse matrix. Despite the fact that the question of the proximity of the obtained minimum point of the surrogate (2.7) and the solution of the problem (2.5) remains open, the optimality of the Hesse matrix inverse approximation $\mathbf{H}_t$ in the context of the chosen subspace $\mathcal{D}_t$ is demonstrated by the following remark.

Remark 3. *In the previous notations, let* $\operatorname{rank}\ \mathbf{D}_t = m$. *Then matrix* $\mathbf{H}_t\mathbf{A} = \mathbf{D}_t \left(\mathbf{D}_t^\top \mathbf{A}\mathbf{D}_t\right)^{-1} \mathbf{D}_t^\top \mathbf{A}$ *has an eigenvalue 1 of multiplicity* m *with corresponding eigen vectors — columns of the matrix* $\mathcal{D}_t$.

Proof. It is suffice to note that

$$\mathbf{H}_t\mathbf{A}\mathbf{D}_t = \mathbf{D}_t \left(\mathbf{D}_t^\top \mathbf{A}\mathbf{D}_t\right)^{-1} \mathbf{D}_t^\top \mathbf{A}\mathbf{D}_t = \mathbf{D}_t.$$

□

Matrix $\mathbf{D}_t \left(\mathbf{D}_t^\top \mathbf{A}\mathbf{D}_t\right)^{-1} \mathbf{D}_t^\top$ — is an approximation of the matrix $\mathbf{A}^{-1}$ in the directions of the columns of the matrix $\mathbf{D}_t$. Remark 3 demonstrates that this approximation along the corresponding directions is accurate.

Next, using the inverse approximation to the Hesse matrix (2.8), consider the following process of the function f optimization:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \mathbf{D}_t \left(\mathbf{D}_t^\top \mathbf{A}\mathbf{D}_t\right)^{-1} \mathbf{D}_t^\top \nabla f\left(\mathbf{x_t}\right) = \mathbf{x}_t - \mathbf{H}_t \nabla f\left(\mathbf{x_t}\right). \tag{2.9}$$

Remark 4. *In the case of the optimization process* (2.9)*, the following recurrent relations take place for the estimates of* $\mathbf{x}_t$ *and the corresponding gradient values:*

$$\begin{aligned} \mathbf{x}_{t+1} - \mathbf{x}_\star &= \left(\mathbf{I} - \mathbf{H}_t\mathbf{A}\right)\left(\mathbf{x}_t - \mathbf{x}_\star\right) \\ &= \left(\mathbf{I} - \mathbf{H}_t\mathbf{A}\right)\left(\mathbf{I} - \mathbf{D}_t\mathbf{D}_t^\top\right)\left(\mathbf{x}_t - \mathbf{x}_\star\right), \end{aligned} \tag{2.10}$$

$$\nabla f\left(\mathbf{x}_{t+1}\right) = \left(\mathbf{I} - \mathbf{A}\mathbf{H}_t\right) \nabla f\left(\mathbf{x}_t\right). \tag{2.11}$$

Proof. Both relations follow from the secant equation: $\nabla f\left(\mathbf{x}\right) = \nabla f\left(\mathbf{x}\right) - \nabla f\left(\mathbf{x}_\star\right) = \mathbf{A}\left(\mathbf{x} - \mathbf{x}_\star\right)$. The second equality in (2.10) additionally follows from the Remark 3: the matrix $\mathbf{H}_t\mathbf{A}$ acts as a unit matrix along the columns of the matrix $\mathbf{D}_t$. □

From the Remark 4 it follows that the step towards $-\mathbf{H}_t \nabla f(\mathbf{x}_t)$ with the inverse of the Hesse matrix obtained by the formula (2.8) has a significant drawback: from the residual $\mathbf{x}_t - \mathbf{x}_\star$, not the entire component of the subspace $\mathcal{D}_t$ is removed: in general, $\mathbf{D}_t^\top (\mathbf{x}_{t+1} - \mathbf{x}_\star) \neq \mathbf{0}$. This is a significant disadvantage, as it can lead to the problem of zigzag trajectories by analogy with the gradient descent method [102; 103].

Note that the sequential subspace optimization step belongs to a span of the columns of the matrix $\mathbf{D}_t$. Consider the following decomposition:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \mathbf{D}_t \mathbf{D}_t^\top \mathbf{A}^{-1} \nabla f(\mathbf{x}_t) + \mathbf{D}_t \mathbf{D}_t^\top \boldsymbol{\xi}_t,$$

where $\mathbf{D}_t \mathbf{D}_t^\top \mathbf{A}^{-1} \nabla f(\mathbf{x}_t) = \mathbf{D}_t \mathbf{D}_t^\top (\mathbf{x}_t - \mathbf{x}_\star)$ — is "optimal" in a sense that it levels the residual along the subspace $\mathcal{D}_t$, and $\mathbf{D}_t \mathbf{D}_t^\top \boldsymbol{\xi}_t$ is an error. The the residual on step $t+1$ can be decomposed into the following to components:

$$(\mathbf{x}_{t+1} - \mathbf{x}_\star) = (\mathbf{I} - \mathbf{H}_t \mathbf{A}_t)(\mathbf{x}_t - \mathbf{x}_\star) + \mathbf{H}_t \mathbf{A}_t \boldsymbol{\xi}_t,$$

— *projection error* $(\mathbf{I} - \mathbf{H}_t \mathbf{A}_t)(\mathbf{x}_t - \mathbf{x}_\star)$ and *approximation error* $\mathbf{H}_t \mathbf{A}_t \boldsymbol{\xi}_t$.

The following Lemma characterizes the contribution of both errors into the overall rate of convergence.

Lemma 1. *Consider α_t and γ_t, that characterize relative values of projection error and approximation error:*

$$\begin{aligned} \|(\mathbf{I} - \mathbf{H}_t \mathbf{A}_t)(\mathbf{x}_t - \mathbf{x}_\star)\|_2^2 &\leqslant (1 - \alpha_t) \|\mathbf{x}_t - \mathbf{x}_\star\|_2^2, \quad \alpha_t \in [0, 1], \\ \|\boldsymbol{\xi}_t\|_2^2 &\leqslant (1 - \gamma_t) \|\mathbf{x}_t - \mathbf{x}_\star\|_2^2, \quad \gamma_t \in [0, 1]. \end{aligned} \tag{2.12}$$

Then, depending on the values of α_t and γ_t, the following convergence rates are achieved:

1. *If $\alpha_t, \gamma_t > 0$ then convergence rate is sublinear;*
2. *If $\exists\, \varepsilon > 0$: $\alpha_t, \gamma_t \geqslant \varepsilon$ the convergence is linear;*
3. *Superlinear convergence is achieved if one of the following conditions is satisfied:*

   (a) $\begin{cases} \exists\, \alpha' > 0, t' < \infty : \forall t \geqslant t', \alpha_t \geqslant \alpha', \\ \gamma_t \to 1, \end{cases}$

   (b) $\begin{cases} \alpha_t \to 1, \\ \exists\, \gamma' > 0, t' < \infty : \forall t \geqslant t', \gamma_t \geqslant \gamma'. \end{cases}$

Proof. Consider the contribution of both projection and approximation errors into the overall rate of convergence

$$\begin{aligned}\|\mathbf{x}_{t+1} - \mathbf{x}_\star\|_2^2 &= \|\left(\mathbf{I} - \mathbf{D}_t\mathbf{D}_t^\top\right)(\mathbf{x}_t - \mathbf{x}_\star)\|_2^2 + \|\mathbf{D}_t\mathbf{D}_t^\top\boldsymbol{\xi}_t\|_2^2 \\ &\leqslant (1-\alpha_t)\|\mathbf{x}_t - \mathbf{x}_\star\|_2^2 + \alpha_t(1-\gamma_t)\|\mathbf{x}_t - \mathbf{x}_\star\|_2^2 \\ &= (1 - \alpha_t\gamma_t)\|\mathbf{x}_t - \mathbf{x}_\star\|_2^2. \end{aligned} \tag{2.13}$$

The Lemma statement directly follows from the equation (2.13). □

Lemma 1 characterizes the contribution of approximation error and projection error to the overall convergence rate of the sequential subspace optimization methods for the quadratic case. It is noteworthy that the role of both errors is comparable: if the projection error is large, the reduction of the approximation error will not qualitatively affect the overall convergence rate and vice versa.

2.1.3 Strongly convex case

Consider a class of doubly differentiable functions $f : \mathbb{R}^d \to \mathbb{R}$ with L–Lipschitz Hessian and μ–strong convexity $\mathcal{F}_{\mu,L}$ (1.4). Denote $\mathbf{x}_\star = \operatorname{argmin}_{\mathbf{x}} f$ — the minimum point of the function f, which exists and unique due to strong convexity.

In this section possible convergence rates for such a class of functions in the context of the sequential subspace optimization approach is investigated.

Lemma 2. *Let $f \in \mathcal{F}_{\mu,L}$. Consider optimization process $\mathbf{x}_{t+1} = \mathbf{x}_t - \alpha_t(\nabla f(\mathbf{x}_t) + \delta_t)$, where α_t — step size, and δ_t — deviation from the gradient vector, s.t. $\delta_t^\top \nabla f(\mathbf{x}_t) = 0$, $\beta_t := \frac{\|\delta_t\|}{\|\nabla f(\mathbf{x}_t)\|}$. Then, sequence $\|\mathbf{x}_t - \mathbf{x}_\star\|$ is bounded from below by the following recurrent relation:*

$$\begin{aligned}\|\mathbf{x}_{t+1} - \mathbf{x}_\star\|^2 &\geqslant \|\mathbf{x}_t - \mathbf{x}_\star\|^2\left(1 + \alpha_t^2(1+\beta_t^2)\mu^2 - 2\alpha_t L\right), \\ \|\mathbf{x}_{t+1} - \mathbf{x}_\star\|^2 &\leqslant \|\mathbf{x}_t - \mathbf{x}_\star\|^2\left(1 + \alpha_t^2(1+\beta_t^2)L^2 - 2\alpha_t\left(1 - \beta_t(L-\mu)\right)\right).\end{aligned}$$

Proof. Denote $\triangle_t = \mathbf{x}_t - \mathbf{x}_\star$ and consider $\|\triangle_{t+1}\|^2$:

$$\begin{aligned}\|\triangle_{t+1}\|^2 &= \|\triangle_t\|^2 - 2\alpha_t\triangle_t^\top\nabla f(\mathbf{x}_t) + \alpha_t^2\|\nabla f(\mathbf{x}_t)\|^2 - 2\alpha_t\triangle_t^\top\delta_t \\ &+ 2\alpha_t^2\delta_t^\top\nabla f(\mathbf{x}_t) + \alpha_t^2\|\delta_t\|^2 \\ &= \|\triangle_t\|^2 + \alpha_t^2(1+\beta_t^2)\|\nabla f(\mathbf{x}_t)\|^2 - 2\alpha_t\triangle_t^\top\nabla f(\mathbf{x}_t) \\ &- 2\alpha_t\triangle_t^\top\delta_t.\end{aligned} \tag{2.14}$$

Taking advantage of the L–Lipschitz gradient and μ–convexity of the function, as well as the fact that $\triangle_t^\top\delta_t \leqslant \beta_t\|\nabla f(\mathbf{x}_t)\|\|\triangle_t\| - \beta_t\triangle_t^\top\nabla f(\mathbf{x}_t)$ a bound from below can be obtained for the $\|\triangle_t\|$:

$$\begin{aligned}\|\triangle_{t+1}\|^2 &\geqslant \|\triangle_t\|^2 + \alpha_t^2(1+\beta_t^2)\|\nabla f(\mathbf{x}_t)\|^2 - 2\alpha_t(1-\beta_t)\triangle_t^\top\nabla f(\mathbf{x}_t) \\ &- 2\alpha_t\beta_t\|\triangle_t\|\|\nabla f(\mathbf{x}_t)\| \\ &\geqslant \|\triangle_t\|^2\left(1 + \alpha_t^2(1+\beta_t^2)\mu^2 - 2\alpha_t L\right).\end{aligned}$$

Next, consider the upper bound on $\|\triangle_t\|$. The main interest is the summand $\triangle_t^\top\delta_t$, which in can be bounded from below as $\beta_t\triangle_t^\top\nabla f(\mathbf{x}_t) - \|\triangle_t\|\|\nabla f(\mathbf{x}_t)\|$. Then

$$\begin{aligned}\|\triangle_{t+1}\|^2 &\leqslant \|\triangle_t\|^2 + \alpha_t^2(1+\beta_t^2)\|\nabla f(\mathbf{x}_t)\|^2 - 2\alpha_t(1+\beta_t)\triangle_t^\top\nabla f(\mathbf{x}_t) \\ &+ 2\alpha_t\beta_t\|\triangle_t\|\|\nabla f(\mathbf{x}_t)\| \\ &\leqslant \|\triangle_t\|^2\left(1 + \alpha_t^2(1+\beta_t^2)L^2 - 2\alpha_t\left(1 - \beta_t(L-\mu)\right)\right).\end{aligned}$$

□

Remark 5. *In the conditions of the Lemma 2, the convergence rate of the residual norm* $\|\triangle_t\|$ *cannot exceed linear if* $\exists\varepsilon > 0$*:* $\beta_t \geqslant \frac{L^2}{\mu^2} - 1 + \varepsilon$*, or* $\beta_t \leqslant \frac{L^2}{\mu^2} - 1$*:*

$$\alpha_t \in \left[\varepsilon, \frac{L - \sqrt{L^2 - (1+\beta_t^2)\mu^2}}{(1+\beta_t^2)\mu^2} - \varepsilon\right] \cup \left[\frac{L + \sqrt{L^2 - (1+\beta_t^2)\mu^2}}{(1+\beta_t^2)\mu^2} + \varepsilon, \infty\right].$$

Proof. It is sufficient to note that roots of an equation $1 + \alpha_t^2(1+\beta_t^2)\mu^2 - 2\alpha_t L = 0$ havethe following form:

$$\alpha_\pm = \frac{L \pm \sqrt{L^2 - (1+\beta_t^2)\mu^2}}{(1+\beta_t^2)\mu^2}.$$

□

Results obtained in Lemma 2 are of only theoretical interest because of the fact that Lipshitz and strict convexity constants L and μ are hard to estimate. The following theorem partially corrects this drawback by imposing stronger conditions on the step $\mathbf{d}_t$.

Lemma 3. *Consider the iterative minimization process of the function* $f \in \mathcal{F}_{\mu,L}$*:* $\mathbf{x}_{t+1} = \mathbf{x}_t - \alpha_t \mathbf{d}_t$*, where* $\mathbf{d}_t$ *— step direction and* α_t *— step size. Denote* $\delta_t = \mathbf{d}_t - \frac{\mathbf{d}_t^\top \nabla f(\mathbf{x}_t)}{\|\nabla f(\mathbf{x}_t)\|^2} \nabla f(\mathbf{x}_t)$*,* $\beta_t := \frac{\|\delta_t\|}{\|\nabla f(\mathbf{x}_t)\|}$ *and assume that* $(\mathbf{x}_t - \mathbf{x}_\star)^\top \delta_t \geqslant 0$*. Then, the sequence* $\|\mathbf{x}_t - \mathbf{x}_\star\|$ *is bounded from above by the following recurrent relations:*

– *in case* $\alpha_t \in \left(0, \frac{2}{(1+\beta_t^2)(\mu+L)}\right]$*:*

$$\|\triangle_{t+1}\|^2 \leqslant \|\triangle_t\|^2 \left(1 + \alpha_t^2(1+\beta_t^2)\mu^2 - 2\alpha_t\mu\right), \tag{2.15}$$

– *in case* $\alpha_t \in \left(\frac{2}{(1+\beta_t^2)(\mu+L)}, \frac{2}{(1+\beta_t^2)L}\right]$*:*

$$\|\triangle_{t+1}\|^2 \leqslant \|\triangle_t\|^2 \left(1 + \alpha_t^2(1+\beta_t^2)L^2 - 2\alpha_t L\right). \tag{2.16}$$

Proof. Denote $\delta_t = \mathbf{d}_t - \frac{\mathbf{d}_t^\top \nabla f(\mathbf{x}_t)}{\|\nabla f(\mathbf{x}_t)\|^2} \nabla f(\mathbf{x}_t)$, then $\delta_t^\top \nabla f(\mathbf{x}_t) = 0$ and similar to the equation (2.14) we have

$$\|\triangle_{t+1}\|^2 = \|\triangle_t\|^2 + \alpha_t^2(1+\beta_t^2)\|\nabla f(\mathbf{x}_t)\|^2 - 2\alpha_t \triangle_t^\top \nabla f(\mathbf{x}_t) - 2\alpha_t \triangle_t^\top \delta_t.$$

According to the theorem conditions $\delta_t^\top \triangle_t \geqslant 0$, hence

$$\begin{aligned}
\|\triangle_{t+1}\|^2 &\leqslant \|\triangle_t\|^2 \left(1 - \frac{2\alpha_t\mu L}{\mu+L}\right) + \alpha_t \|\nabla f(\mathbf{x}_t)\|^2 \left(\alpha_t(1+\beta_t^2) - \frac{2}{\mu+L}\right) \\
&\leqslant^1 \|\triangle_t\|^2 \left(1 - \frac{2\alpha_t\mu L}{\mu+L} + \alpha_t^2(1+\beta_t^2)\mu^2 - \frac{2\alpha_t\mu^2}{\mu+L}\right) \\
&= \|\triangle_t\|^2 \left(1 + \alpha_t^2(1+\beta_t^2)\mu^2 - 2\alpha_t\mu\right)
\end{aligned}$$

where the 1st inequality and linear convergence of $\|\triangle_t\|$ holds true when:

$$0 < \alpha_t \leqslant \frac{2}{(1+\beta_t^2)(\mu+L)}.$$

On the other hand, if $\alpha_t \geqslant \frac{2}{(1+\beta_t^2)(\mu+L)}$, then the following recurrent inequality can be obtained in a similar way:

$$\|\triangle_{t+1}\|^2 \leqslant \|\triangle_t\|^2 \left(1 + \alpha_t^2(1+\beta_t^2)L^2 - 2\alpha_t L\right),$$

from which it follows that monotonic convergence also takes place under the following α_t:

$$\frac{2}{(1+\beta_t^2)(\mu+L)} \leqslant \alpha_t < \frac{2}{(1+\beta_t^2)L}.$$

□

Compared to the Lemma 2, Lemma 3 represents a more practical result. To satisfy the constraint on α_t, it is sufficient to obtain an upper bound on L based on the physical meaning of the problem, or using approximate computational procedures (see, for example, the power method [104], § 53, or the iteration method with Rayleigh relations in [105], as well as in [106], § 4.6). The β_t values have no explicit restrictions, but implicitly they are limited the need to select the corresponding α_t. Next, consider the Lemma 2 from the sequential subspace optimization point of view.

Theorem 2. *Consider $f \in \mathcal{F}_{\mu,L}$, optimization process $\mathbf{x}_{t+1} = \mathbf{x}_t - \alpha_t \mathbf{d}_t$, where α_t — stpe size and $\mathbf{d}_t$ — step direction. Denote $\gamma_t := \frac{\mathbf{d}_t^\top \nabla f(\mathbf{x}_t)}{\nabla f(\mathbf{x}_t)^\top \nabla f(\mathbf{x}_t)}$, $\beta_t := \frac{\|\mathbf{d}_t/\gamma_t - \nabla f(\mathbf{x}_t)\|}{\|\nabla f(\mathbf{x}_t)\|}$ and assume that $\beta_t \leqslant \frac{\rho}{L}$, where $\rho \in (0,1)$ and $\alpha_t = \frac{1-\rho}{\gamma_t(L^2+\rho^2)}$. The sequence $\|\mathbf{x}_t - \mathbf{x}_\star\|$ tends to zero with linear convergence rate:*

$$\|\mathbf{x}_{t+1} - \mathbf{x}_\star\|^2 \leqslant \left(1 - \frac{1-\rho}{L^2+\rho^2}\right) \|\mathbf{x}_t - \mathbf{x}_\star\|^2. \tag{2.17}$$

Proof. According to Lemma 2 the following recurrent majorizing relation is satisfied for $\|\mathbf{x}_t - \mathbf{x}_\star\|$:

$$\|\mathbf{x}_{t+1} - \mathbf{x}_\star\|^2 \leqslant \|\mathbf{x}_t - \mathbf{x}_\star\|^2 \left(1 + \overline{\alpha}_t^2(1+\beta_t^2)L^2 - 2\overline{\alpha}_t\left(1 - \beta_t(L-\mu)\right)\right),$$

where $\overline{\alpha}_t = \alpha_t\gamma_t$.

Substituting the β_t upper bound:

$$\|\mathbf{x}_{t+1} - \mathbf{x}_\star\|^2 \leqslant \|\mathbf{x}_t - \mathbf{x}_\star\|^2 \left(1 + \overline{\alpha}_t^2(L^2+\rho^2)L^2 - 2\overline{\alpha}_t\left(1-\rho\right)\right).$$

One can note that $\overline{\alpha}_t^2(L^2+\rho^2)L^2 - 2\overline{\alpha}_t\left(1-\rho\right)$ minimum is achieved when $\overline{\alpha_t} = \frac{1-\rho}{L^2+\rho^2}$ and equals to $-\frac{1-\rho}{L^2+\rho^2}$. Hence $\alpha_t = \frac{1-\rho}{\gamma_t(L^2+\rho^2)}$.

□

Theorem 2 demonstrates that linear convergence for strongly convex functions with Lipschitz gradient, which is a known fact for gradient descent (see, for example, Theorem 2.1.15 in [62], or Theorem 2, § 4, Chapter 1 in [65]), is preserved when difference between step directions and the gradient directions are uniformly bounded. This difference is characterized by the coefficients β_t and γ_t. It is noteworthy that γ_t is included only in the step size of the algorithm, but is absent in the convergence rate upper bound estimate. Thus, the ℓ_2–norm of the difference in the size of the selected direction $\mathbf{d}_t$ and the gradient in is compensated by the step size. The convergence rate estimate includes only the parameter ρ, which limits the difference between the selected direction and the gradient direction.

Having obtained the conditions for the linear convergence rate, we proceed further to study the conditions necessary to ensure a superlinear rate of convergence. The following Lemma presents the upper bound on the quasi-Newtonian methods residual.

Lemma 4. *Let $f \in \mathcal{F}_{\mu,L}$ and $\nabla^2 f$ satisfy the Lipschitz condition with the constant L'. Consider the method with the following step: $\mathbf{x}_{t+1} = \mathbf{x}_t - \alpha_t \mathbf{H}_t \nabla f(\mathbf{x}_t)$. Then the residual norm satisfies the following recurrence relation:*

$$\|\mathbf{x}_{t+1} - \mathbf{x}_\star\| \leqslant \alpha_t \frac{L}{2} \|\mathbf{H}_t\| \|\mathbf{x}_{t+1} - \mathbf{x}_\star\|^2 + (1 - \alpha_t)\|\mathbf{x}_t - \mathbf{x}_\star\| + \alpha_t \|(\mathbf{I} - \mathbf{H}_t \nabla^2 f(\mathbf{x}_t))(\mathbf{x}_t - \mathbf{x}_\star)\|.$$

Proof.

$$\begin{aligned}
\|\mathbf{x}_{t+1} - \mathbf{x}_\star\| &= \|\mathbf{x}_t - \mathbf{x}_\star - \alpha_t \mathbf{H}_t \nabla^2 f(\mathbf{x}_t)(\mathbf{x}_t - \mathbf{x}_\star) + \alpha_t \mathbf{H}_t \nabla^2 f(\mathbf{x}_t)(\mathbf{x}_t - \alpha_t \mathbf{x}_\star) \\
&\quad - \alpha_t \mathbf{H}_t \nabla f(\mathbf{x}_t)\| \\
&\leqslant \alpha_t \|\left(\mathbf{I} - \mathbf{H}_t \nabla^2 f(\mathbf{x}_t)\right)(\mathbf{x}_t - \mathbf{x}_\star)\| + (1 - \alpha_t)\|\mathbf{x}_t - \mathbf{x}_\star\| \\
&\quad + \alpha_t \|\mathbf{H}_t \left(\nabla^2 f(\mathbf{x}_t)(\mathbf{x}_t - \mathbf{x}_\star) - \nabla f(\mathbf{x}_t)\right)\|, \\
&\text{according to the Statement 2:} \\
&\leqslant \alpha_t \|\left(\mathbf{I} - \mathbf{H}_t \nabla^2 f(\mathbf{x}_t)\right)(\mathbf{x}_t - \mathbf{x}_\star)\| + (1 - \alpha_t)\|\mathbf{x}_t - \mathbf{x}_\star\| \\
&\quad + \alpha_t \|\mathbf{H}_t\| \frac{L'}{2} \|\mathbf{x}_t - \mathbf{x}_\star\|^2.
\end{aligned}$$

□

In the following Theorem, which uses the results of the Lemma (4), a characteristic of sequential subspace optimization methods convergence rate with a quasi-Newtonian step of the form (2.9) is provided.

Theorem 3. *Let $f \in \mathcal{F}_{\mu,L}$ and $\nabla^2 f$ satisfy Lipschitz condition with constant L'. Consider method with the following step $\mathbf{x}_{t+1} = \mathbf{x}_t - \alpha_t \mathbf{H}_t \nabla f(\mathbf{x}_t)$, where $\mathbf{H}_t = \mathbf{D}_t \mathbf{Q}_t^{-1} \mathbf{D}_t^\top$. Assume that the following assumptions are satisfied*

1. *matrices $\mathbf{D}_t$ are chosen in a way that $\mathbf{D}_t \mathbf{D}_t^\top \nabla f(\mathbf{x}_t) = \nabla f(\mathbf{x}_t)$ and its columns are orthonormal $\mathbf{D}_t^\top \mathbf{D}_t = \mathbf{I}$;*
2. *step sizes α_t are chosen in a way that $1 - \|\nabla f(\mathbf{x}_t)\| \leqslant \alpha_t \leqslant 1$;*
3. *matrices $\mathbf{D}_t$ are chosen in a way that $\exists\, c_0, C_0 > 0$: $\alpha_t \|(\mathbf{I} - \mathbf{D}_t \mathbf{D}_t^\top) \left[\nabla^2 f(\mathbf{x}_t)\right]^{-1} \nabla f(\mathbf{x}_t)\| \leqslant C_0 \|\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1} \nabla f(\mathbf{x}_t)\|^{1+c_0}$;*
4. *subspace optimization problem solution accuracy increases at a superlinear rate relative to the gradient norm: $\exists\, c_1, C_1 > 0$: $\|\left[\mathbf{D}_t^\top \nabla^2 f(\mathbf{x}_t) \mathbf{D}_t\right]^{-1} \mathbf{D}_t^\top \nabla f(\mathbf{x}_t) - \mathbf{Q}_t^{-1} \mathbf{D}_t^\top \nabla f(\mathbf{x}_t)\| \leqslant C_1 \|\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1} \nabla f(\mathbf{x}_t)\|$*

Then the method converges with a superlinear rate:

$$\exists C > 0, c > 0: \quad \|\mathbf{x}_{t+1} - \mathbf{x}_\star\| \leqslant C\|\mathbf{x}_t - \mathbf{x}_\star\|^{1+c},$$

where $C = \min\left(C_0, 2C_1, \left(L'\|\mathbf{Q}_t^{-1}\| + \frac{L'^2}{2} + L\right)\right)$, $c = \min(c_0, c_1, 1)$.

Proof. Substitute the error bounds from the Lemma 4 and matrix $\mathbf{H}_t = \mathbf{D}_t\mathbf{Q}_t^{-1}\mathbf{D}_t^\top$

$$\begin{aligned}\|\mathbf{x}_{t+1} - \mathbf{x}_\star\| &\leqslant \alpha_t\|\left(\mathbf{I} - \mathbf{D}_t\mathbf{Q}_t^{-1}\mathbf{D}_t^\top\nabla^2 f(\mathbf{x}_t)\right)(\mathbf{x}_t - \mathbf{x}_\star)\| + (1-\alpha_t)\|\mathbf{x}_t - \mathbf{x}_\star\| \\ &\quad + \alpha_t\|\mathbf{D}_t\mathbf{Q}_t^{-1}\mathbf{D}_t^\top\|\frac{L'}{2}\|\mathbf{x}_t - \mathbf{x}_\star\|^2.\end{aligned}$$

Note that $\mathbf{D}_t\mathbf{Q}_t^{-1}\mathbf{D}_t^\top\| \leqslant \|\mathbf{Q}_t^{-1}\|$ and consider $\|\left(\mathbf{I} - \mathbf{H}_t\nabla^2 f(\mathbf{x}_t)\right)(\mathbf{x}_t - \mathbf{x}_\star)\|$

$$\begin{aligned}&\|\left(\mathbf{I} - \mathbf{H}_t\nabla^2 f(\mathbf{x}_t)\right)(\mathbf{x}_t - \mathbf{x}_\star)\| \\ &= \|(\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1} - \mathbf{D}_t\mathbf{Q}_t^{-1}\mathbf{D}_t^\top)\nabla^2 f(\mathbf{x}_t)(\mathbf{x}_t - \mathbf{x}_\star)\| \\ &\leqslant \|(\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1} - \mathbf{D}_t\mathbf{Q}_t^{-1}\mathbf{D}_t^\top)\nabla f(\mathbf{x}_t)\| \\ &\quad + \|\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1} - \mathbf{D}_t\mathbf{Q}_t^{-1}\mathbf{D}_t^\top\|\frac{L'}{2}\|\mathbf{x}_t - \mathbf{x}_\star\|^2 \\ &\leqslant \|(\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1} - \mathbf{D}_t\mathbf{Q}_t^{-1}\mathbf{D}_t^\top)\nabla f(\mathbf{x}_t)\| + \left(\frac{1}{\mu} + \|\mathbf{Q}_t^{-1}\|\right)\frac{L'}{2}\|\mathbf{x}_t - \mathbf{x}_\star\|^2.\end{aligned}$$

Consider $\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1}\nabla f(\mathbf{x}_t) - \mathbf{D}_t\mathbf{Q}_t^{-1}\mathbf{D}_t^\top\nabla f(\mathbf{x}_t)$:

$$\begin{aligned}&\|\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1}\nabla f(\mathbf{x}_t) - \mathbf{D}_t\mathbf{Q}_t^{-1}\mathbf{D}_t^\top\nabla f(\mathbf{x}_t)\| \\ &\qquad \leqslant \|\mathbf{D}_t\mathbf{D}_t^\top\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1}\nabla f(\mathbf{x}_t) - \mathbf{D}_t\mathbf{Q}_t^{-1}\mathbf{D}_t^\top\nabla f(\mathbf{x}_t)\| \\ &\qquad + \|(\mathbf{I} - \mathbf{D}_t\mathbf{D}_t^\top)\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1}\nabla f(\mathbf{x}_t)\| \\ &\qquad \leqslant \|\mathbf{D}_t^\top\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1}\nabla f(\mathbf{x}_t) - \mathbf{Q}_t^{-1}\mathbf{D}_t^\top\nabla f(\mathbf{x}_t)\| \\ &\qquad + \|(\mathbf{I} - \mathbf{D}_t\mathbf{D}_t^\top)\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1}\nabla f(\mathbf{x}_t)\| \\ &\qquad \leqslant \|\mathbf{D}_t^\top\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1}\nabla f(\mathbf{x}_t) - \left[\mathbf{D}_t^\top\nabla^2 f(\mathbf{x}_t)\mathbf{D}_t\right]^{-1}\mathbf{D}_t^\top\nabla f(\mathbf{x}_t)\| \\ &\qquad + \|\left[\mathbf{D}_t^\top\nabla^2 f(\mathbf{x}_t)\mathbf{D}_t\right]^{-1}\mathbf{D}_t^\top\nabla f(\mathbf{x}_t) - \mathbf{Q}_t^{-1}\mathbf{D}_t^\top\nabla f(\mathbf{x}_t)\| \\ &\qquad + \|(\mathbf{I} - \mathbf{D}_t\mathbf{D}_t^\top)\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1}\nabla f(\mathbf{x}_t)\| \\ &\qquad \leqslant \|\left[\mathbf{D}_t^\top\nabla^2 f(\mathbf{x}_t)\mathbf{D}_t\right]^{-1}\mathbf{D}_t^\top\nabla f(\mathbf{x}_t) - \mathbf{Q}_t^{-1}\mathbf{D}_t^\top\nabla f(\mathbf{x}_t)\| \\ &\qquad + \left(1 + \frac{L}{\mu}\right)\|(\mathbf{I} - \mathbf{D}_t\mathbf{D}_t^\top)\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1}\nabla f(\mathbf{x}_t)\|.\end{aligned}$$

The last inequality holds true due to the following:

$$\begin{aligned}
&\|\mathbf{D}_t^\top \left[\nabla^2 f(\mathbf{x}_t)\right]^{-1} \nabla f(\mathbf{x}_t) - \left[\mathbf{D}_t^\top \nabla^2 f(\mathbf{x}_t)\mathbf{D}_t\right]^{-1} \mathbf{D}_t^\top \nabla f(\mathbf{x}_t)\| \\
&\quad \leqslant \|\left[\mathbf{D}_t^\top \nabla^2 f(\mathbf{x}_t)\mathbf{D}_t\right]^{-1} \mathbf{D}_t \nabla^2 f(\mathbf{x}_t)(\mathbf{I} - \mathbf{D}_t\mathbf{D}_t^\top) \left[\nabla^2 f(\mathbf{x}_t)\right]^{-1} \nabla f(\mathbf{x}_t)\| \\
&\quad \leqslant \|\left[\mathbf{D}_t^\top \nabla^2 f(\mathbf{x}_t)\mathbf{D}_t\right]^{-1} \mathbf{D}_t \nabla^2 f(\mathbf{x}_t)\| \|(\mathbf{I} - \mathbf{D}_t\mathbf{D}_t^\top) \left[\nabla^2 f(\mathbf{x}_t)\right]^{-1} \nabla f(\mathbf{x}_t)\| \\
&\quad \leqslant \frac{L}{\mu}\|(\mathbf{I} - \mathbf{D}_t\mathbf{D}_t^\top) \left[\nabla^2 f(\mathbf{x}_t)\right]^{-1} \nabla f(\mathbf{x}_t)\|.
\end{aligned}$$

By substituting the obtained boundaries into the expression of the residuals from the Lemma 4 we get:

$$\begin{aligned}
\|\mathbf{x}_{t+1} - \mathbf{x}_\star\| &\leqslant \alpha_t \|\left(\mathbf{I} - \mathbf{D}_t\mathbf{Q}_t^{-1}\mathbf{D}_t^\top \nabla^2 f(\mathbf{x}_t)\right)(\mathbf{x}_t - \mathbf{x}_\star)\| + (1-\alpha_t)\|\mathbf{x}_t - \mathbf{x}_\star\| \\
&+ \alpha_t \|\mathbf{D}_t\mathbf{Q}_t^{-1}\mathbf{D}_t^\top\| \frac{L'}{2}\|\mathbf{x}_t - \mathbf{x}_\star\|^2, \\
&\text{due to limitations imposed on } \alpha_t \\
&\leqslant \alpha_t \|\left[\mathbf{D}_t^\top \nabla^2 f(\mathbf{x}_t)\mathbf{D}_t\right]^{-1} \mathbf{D}_t^\top \nabla f(\mathbf{x}_t) - \mathbf{Q}_t^{-1}\mathbf{D}_t^\top \nabla f(\mathbf{x}_t)\| \\
&+ 2\alpha_t \|(\mathbf{I} - \mathbf{D}_t\mathbf{D}_t^\top) \left[\nabla^2 f(\mathbf{x}_t)\right]^{-1} \nabla f(\mathbf{x}_t)\| \\
&+ \alpha_t \frac{L'}{2}\left(\frac{1}{\mu} + \|\mathbf{Q}_t^{-1}\|\right)\|\mathbf{x}_t - \mathbf{x}_\star\|^2 + L\|\mathbf{x}_t - \mathbf{x}_\star\|^2 \\
&+ \alpha_t \frac{L'}{2}\|\mathbf{Q}_t^{-1}\|\|\mathbf{x}_t - \mathbf{x}_\star\|^2, \\
&\text{substituting constraints from the Theorem conditions:} \\
&\leqslant \alpha_t C_0 \|\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1} \nabla f(\mathbf{x}_t)\|^{1+c_0} + 2\alpha_t C_1 \|\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1} \nabla f(\mathbf{x}_t)\|^{1+c_1} \\
&+ \left(\alpha_t L' \|\mathbf{Q}_t^{-1}\| + \alpha_t \frac{L'^2}{2} + L\right)\|\mathbf{x}_t - \mathbf{x}_\star\|^2 \\
&\leqslant \min(C_0, 2C_1)\|\mathbf{x}_t - \mathbf{x}_\star\|^{1+\min(c_0,c_1)} + \alpha_t C_0 \|\mathbf{x}_t - \mathbf{x}_\star\|^{2+2c_0} \\
&+ 2\alpha_t C_1 \|\mathbf{x}_t - \mathbf{x}_\star\|^{2+2c_1} + \left(\alpha_t L' \|\mathbf{Q}_t^{-1}\| + \alpha_t \frac{L'^2}{2} + L\right)\|\mathbf{x}_t - \mathbf{x}_\star\|^2.
\end{aligned}$$

□

Theorem 3 provides sufficient conditions for superlinear (and quadratic — depending on the c_0 and c_1 values) convergence. It demonstrates that the rate of convergence of sequential subspace optimization methods depends equally on the quality of the subspaces chosen and the subspace problem solution accuracy. We will discuss the conditions imposed in the Theorem in more detail. Conditions (1) and (2) appear to be the easiest to achieve. By selecting $\alpha_t = 1$ and $\mathbf{D}_t = [\nabla f(\mathbf{x}_t) / \|\nabla f(\mathbf{x}_t)\|]$ they are

obviously satisfied. The condition (3) characterizes an angle between $\mathcal{D}_t$ and the vector $\left[\nabla^2 f(\mathbf{x}_t)\right]^{-1} \nabla f(\mathbf{x}_t)$. The condition (4) characterizes the subspace optimization problem solution accuracy. These conditions (3) and (4) are of the greatest interest, and the question of the way to achieve them remains open. In the following sections we will consider several approaches to both the construction of $\mathbf{D}_t$ matrices and to the solution of the subspace problem (i.e. evaluation of the vector $\mathbf{Q}_t^{-1}\mathbf{D}_t^\top \nabla f(\mathbf{x}_t)$) in the context of proved Theorem 2 and Lemma 3.

2.2 Elements of sequential subspace optimization methods

As mentioned above, the formation of subspaces, as well as the way to solve the subspace optimization problem — the two main characteristics of SSO methods. This section describes approaches to each of them.

2.2.1 Subspace step estimation

Taking into account the Taylor expansion of the function f (2.6), it is natural to consider a quadratic surrogate of the form:

$$q_t(\mathbf{z}) = f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top \mathbf{D}_t\mathbf{z} + \frac{1}{2}\mathbf{z}^\top \mathbf{D}_t^\top \nabla^2 f(\mathbf{x}_t)\mathbf{D}_t\mathbf{z}. \tag{2.18}$$

finding the optimum point of which is equivalent to solving the following system of linear equations:

$$\mathbf{D}_t^\top \nabla^2 f(\mathbf{x}_t)\mathbf{D}_t\mathbf{z} = \mathbf{D}_t^\top \nabla f(\mathbf{x}_t), \tag{2.19}$$

wich if a inverse matrix $\left[\mathbf{D}_t^\top \nabla^2 f(\mathbf{x}_t)\mathbf{D}_t\right]^{-1}$ exists as achieved at the following point

$$\mathbf{z}_\star^{(t)} = \left[\mathbf{D}_t^\top \nabla^2 f(\mathbf{x}_t)\mathbf{D}_t\right]^{-1} \mathbf{D}_t^\top \nabla f(\mathbf{x}_t). \tag{2.20}$$

Unfortunately, computing the matrix $\mathbf{D}_t^\top \nabla^2 f(\mathbf{x}_t)\mathbf{D}_t$ generally requires computing the Hessian of the original function $\nabla^2 f(\mathbf{x}_t)$, which can be difficult due to memory constraints, computation time, or the lack of an explicit Hessian expression. For these

reasons, instead of optimal surrogate (2.18) (in terms of the Theorem 3) we consider a surrogate with an unknown matrix of quadratic coefficients $\mathbf{Q}_t$:

$$q_t(\mathbf{z}) = f(\mathbf{x}_t) + \nabla f\left(\mathbf{x}_t\right)^\top \mathbf{D}_t\mathbf{z} + \frac{1}{2}\mathbf{z}^\top\mathbf{Q}_t\mathbf{z}. \tag{2.21}$$

In turn, instead of the system (2.19) consider

$$\mathbf{Q}_t\mathbf{z} = \mathbf{D}_t^\top\nabla f\left(\mathbf{x}_t\right). \tag{2.22}$$

It can be solved in a way similar to (2.20) with the solution carried out in the following way:

$$\mathbf{z}^{(t)} = \mathbf{Q}_t^{-1}\mathbf{D}_t^\top\nabla f\left(\mathbf{x}_t\right), \tag{2.23}$$

where the only unknown element is the inverse Hesse matrix $\mathbf{Q}_t^{-1}$. Once $\mathbf{z}_t$ is defined, the next step direction of the algorithm is calculated in the obvious way:

$$\mathbf{d}_t = \mathbf{D}_t\mathbf{Q}_t^{-1}\mathbf{D}_t^\top\nabla f\left(\mathbf{x}_t\right). \tag{2.24}$$

In section 2.1 some properties of sequential subspace methods based on surrogates of this kind are demonstrated. Next, we'll look at several ways to construct the 2.23 directly or via evaluating the matrix $\mathbf{Q}_t^{-1}$. In this case, the goal is to obtain an estimate close to the optimal (2.20). We will use the history of the arguments $\{\mathbf{x}_t\}$, the corresponding gradient values $\nabla f\left(\mathbf{x}_t\right)$ and the values of the function $f(\mathbf{x}_t)$ as input to obtain those estimates. Note that the ultimate goal is to construct a new estimate of the vector (2.20) of the form (2.23) — explicit construction of matrix $\mathbf{Q}_t^{-1}$ is not necessary and methods obtaining evaluation (2.23) without explicit construction of the $\mathbf{Q}_t^{-1}$ are of no less interest to us.

Next, consider several approaches to constructing estimates of the vector $\mathbf{z}_\star^{(t)}$ (2.20) of the form (2.23) based on the use of *truncated history* surrogate values q_t, its arguments and gradients: $\{\mathbf{z}_j^{(t)}\}_{i=1}^K \subset \mathbb{R}^m$, $\{q_t(\mathbf{z}_j^{(t)})\}_{i=1}^K \subset \mathbb{R}^m$. The matrix $\mathbf{Q}_t$ is unknown to us, and therefore the values of the surrogate $\{q_t(\mathbf{z}_j^{(t)})\}_{j=1}^K \subset \mathbb{R}$ and its gradient $\{\nabla q_t(\mathbf{z}_j^{(t)})\}_{j=1}^K \subset \mathbb{R}^m$ are unknown too. Instead, we will use the following approximations:

$$\mathbf{z}_j^{(t)} = \mathbf{D}_t^\top(\mathbf{w}_j^{(t)} - \mathbf{x}_t), \tag{2.25}$$

$$q_t(\mathbf{z}_j^{(t)}) \approx f(\mathbf{w}_j), \tag{2.26}$$

$$\nabla q_t(\mathbf{z}_j^{(t)}) \approx \mathbf{D}_t^\top\nabla f(\mathbf{w}_j^{(t)}), \tag{2.27}$$

where instead of $\mathbf{w}_j^{(t)}$ one can choose, for example, the preceding vector $\mathbf{x}_{t-j}$, or a random vector from $\mathbb{R}^d$.

Remark 6. *Note that the approximations* (2.26) *and* (2.27) *are not exact even given the optimal estimate of the* $\mathbf{Q}_t$*. Thus, approximate inequalities become exact only if* $\mathbf{Q}_t = \mathbf{D}_t^\top \nabla^2 f(\mathbf{x}_t)\mathbf{D}_t$ *and the function* f *is quadratic. Indeed, let* $\mathbf{Q}_t = \mathbf{D}_t^\top \nabla^2 f(\mathbf{x}_t)\mathbf{D}_t$*, then*

$$\nabla q_t(\mathbf{z}_j^{(t)}) - \mathbf{D}_t^\top \nabla f(\mathbf{w}_j^{(t)}) = \mathbf{D}_t^\top \nabla^2 f(\mathbf{x}_t)(\mathbf{I} - \mathbf{D}_t\mathbf{D}_t^\top)(\mathbf{w}_j^{(t)} - \mathbf{x}_t) + o(\|\mathbf{w}_j^{(t)} - \mathbf{x}_t\|),$$
$$q_t(\mathbf{z}_j^{(t)}) - f(\mathbf{w}_j^{(t)}) = \frac{1}{2}(\mathbf{w}_j^{(t)} - \mathbf{x}_t)^\top \left(\nabla^2 f(\mathbf{x}_t) - \mathbf{D}_t\mathbf{D}_t^\top \nabla^2 f(\mathbf{x}_t)\mathbf{D}_t\mathbf{D}_t^\top\right)(\mathbf{w}_j^{(t)} - \mathbf{x}_t)$$
$$+\nabla f\left(\mathbf{x}_t\right)^\top (\mathbf{I} - \mathbf{D}_t\mathbf{D}_t^\top)(\mathbf{w}_j^{(t)} - \mathbf{x}_t) + o(\|\mathbf{x}_t - \mathbf{w}_j^{(t)}\|^2).$$

If additionally $\mathbf{D}_t\mathbf{D}_t^\top(\mathbf{w}_j^{(t)} - \mathbf{x}_t) = (\mathbf{w}_j^{(t)} - \mathbf{x}_t)$*, then*

$$\nabla q_t(\mathbf{z}_j^{(t)}) - \mathbf{D}_t^\top \nabla f(\mathbf{w}_j^{(t)}) = o(\|\mathbf{w}_j^{(t)} - \mathbf{x}_t\|),$$
$$q_t(\mathbf{z}_j^{(t)}) - f(\mathbf{w}_j^{(t)}) = o(\|\mathbf{w}_j^{(t)} - \mathbf{x}_t\|^2).$$

Finally, if $\mathbf{D}_t\mathbf{D}_t^\top(\mathbf{w}_j^{(t)} - \mathbf{x}_t) = (\mathbf{w}_j^{(t)} - \mathbf{x}_t)$ *and* f *is quadratic polynom, then*

$$\nabla q_t(\mathbf{z}_j^{(t)}) - \mathbf{D}_t^\top \nabla f(\mathbf{w}_j^{(t)}) = 0,$$
$$q_t(\mathbf{z}_j^{(t)}) - f(\mathbf{w}_j^{(t)}) = 0.$$

Proof. Let us clearly describe the difference between the gradient of the surrogate $\nabla q_t(\mathbf{z}_j^{(t)})$ and the projection $\mathbf{D}_t^\top \nabla f(\mathbf{w}_j^{(t)})$

$$\begin{aligned}\nabla q_t(\mathbf{z}_j^{(t)}) - \mathbf{D}_t^\top \nabla f(\mathbf{w}_j^{(t)}) &= \mathbf{D}_t^\top \nabla f\left(\mathbf{x}_t\right) + \mathbf{D}_t^\top \nabla^2 f(\mathbf{x}_t)\mathbf{D}_t\mathbf{z}_j^{(t)} + o(\|\mathbf{w}_j^{(t)} - \mathbf{x}_t\|)\\ &- \mathbf{D}_t^\top \nabla f\left(\mathbf{x}_t\right) - \mathbf{D}_t^\top \nabla^2 f(\mathbf{x}_t)(\mathbf{w}_j^{(t)} - \mathbf{x}_t)\\ &= \mathbf{D}_t^\top \nabla^2 f(\mathbf{x}_t)\mathbf{D}_t\mathbf{z}_j^{(t)} - \mathbf{D}_t^\top \nabla^2 f(\mathbf{x}_t)(\mathbf{w}_j^{(t)} - \mathbf{x}_t)\\ &+ o(\|\mathbf{w}_j^{(t)} - \mathbf{x}_t\|)\\ &= \mathbf{D}_t^\top \nabla^2 f(\mathbf{x}_t)(\mathbf{I} - \mathbf{D}_t\mathbf{D}_t^\top)(\mathbf{w}_j^{(t)} - \mathbf{x}_t) + o(\|\mathbf{w}_j^{(t)} - \mathbf{x}_t\|).\end{aligned}$$

Similarly, the difference between the value of the surrogate and the value of the function can be carried out:

$$\begin{aligned}q_t(\mathbf{z}_j^{(t)}) - f(\mathbf{w}_j^{(t)}) &= f(\mathbf{x}_t) + \nabla f\left(\mathbf{x}_t\right)^\top \mathbf{D}_t\mathbf{z}_j^{(t)} + \frac{1}{2}\mathbf{z}_j^{(t)\top}\mathbf{D}_t^\top \nabla^2 f(\mathbf{x}_t)\mathbf{D}_t\mathbf{z}_j^{(t)} - f(\mathbf{x}_t)\\ &- \nabla f\left(\mathbf{x}_t\right)^\top (\mathbf{w}_j^{(t)} - \mathbf{x}_t) - \frac{1}{2}(\mathbf{w}_j^{(t)} - \mathbf{x}_t)^\top \nabla^2 f(\mathbf{x}_t)(\mathbf{w}_j^{(t)} - \mathbf{x}_t)\\ &+ o(\|\mathbf{x}_t - \mathbf{w}_j^{(t)}\|^2)\\ &= \nabla f\left(\mathbf{x}_t\right)^\top (\mathbf{I} - \mathbf{D}_t\mathbf{D}_t^\top)(\mathbf{w}_j^{(t)} - \mathbf{x}_t) + o(\|\mathbf{x}_t - \mathbf{w}_j^{(t)}\|^2)\\ &+ \frac{1}{2}(\mathbf{w}_j^{(t)} - \mathbf{x}_t)^\top \left[\nabla^2 f(\mathbf{x}_t) - \mathbf{D}_t\mathbf{D}_t^\top \nabla^2 f(\mathbf{x}_t)\mathbf{D}_t\mathbf{D}_t^\top\right](\mathbf{w}_j^{(t)} - \mathbf{x}_t).\end{aligned}$$

□

2.2.2 Subspace step via the secant equation

Secant equations are often used to estimate the matrix of second derivatives. For example, the chord equations for the gradient is determinant to obtain an estimate of the (inverse) Hesse matrix in the Davidon–Fletcher–Powell [14; 18] method, SR-1 [19] method, Broyden–Fletcher–Goldfarb–Shanno [15; 18] method, L-BFGS [17] method, Barzilai–Borwein [71] method, etc. Consider the function $q : \mathbb{R}^m \to \mathbb{R}$ and two points $\mathbf{z}, \mathbf{z}_0 \in \mathbb{R}^m$. Then the corresponding secant equation of the gradient function ∇q is as follows:

$$\nabla^2 q(\mathbf{z}')(\mathbf{z} - \mathbf{z}_0) = \nabla q(\mathbf{z}) - \nabla q(\mathbf{z}_0),$$

where $\mathbf{z}' \in \operatorname{span}\{\mathbf{z},\mathbf{z}_0\}$. Note that it is easy to obtain an approximate version of the secant equation from the Taylor expansion: $\nabla^2 q(\mathbf{z})(\mathbf{z} - \mathbf{z}_0) = \nabla q(\mathbf{z}) - \nabla q(\mathbf{z}_0) + o(\|\mathbf{z} - \mathbf{z}_0\|)$. A more accurate estimate of the deviation is possible, for example, in case of L-Lipschitz gradient: $\|\nabla^2 q(\mathbf{z})(\mathbf{z} - \mathbf{z}_0) - (\nabla q(\mathbf{z}) - \nabla q(\mathbf{z}_0))\|_2^2 \leqslant \frac{L}{2}\|\mathbf{z} - \mathbf{z}_0\|^2$ (see Lemma 1.2.3 in [64]).

Consider a system of secant equations for q_t of the form (2.21)

$$\mathbf{Q}_t \left(\mathbf{z}_0^{(t)} - \mathbf{z}_j^{(t)}\right) = \nabla q_t(\mathbf{z}_0^{(t)}) - \nabla q_t(\mathbf{z}_j^{(t)}), \quad j = 1 \ldots K - 1, \tag{2.28}$$

where the corresponding values of the gradient arguments are calculated approximately by the formulas (2.26) and (2.27):

$$\begin{gathered} \mathbf{z}_j^{(t)} - \mathbf{z}_0^{(t)} = \mathbf{z}_j^{(t)} = \mathbf{D}_t^\top \mathbf{w}_j^{(t)} \\ \nabla q_t(\mathbf{z}_j^{(t)}) - \nabla q_t(\mathbf{z}_0^{(t)}) = \mathbf{D}_t^\top \left(\nabla f(\mathbf{w}_j^{(t)}) - \nabla f\left(\mathbf{x}_t\right)\right). \end{gathered}$$

Consider matrices $\mathbf{Z}_t$ and $\mathbf{G}_t$, whose strings are composed of differences in argument values and gradients, respectively:

$$\begin{gathered} \mathbf{Z}_t = \left[\mathbf{D}_t^\top \mathbf{w}_1^{(t)}, \ldots, \mathbf{D}_t^\top \mathbf{w}_K^{(t)}\right]^\top, \\ \mathbf{G}_t = \left[\mathbf{D}_t^\top \left(\nabla f(\mathbf{w}_1^{(t)}) - \nabla f\left(\mathbf{x}_t\right)\right), \ldots, \mathbf{D}_t^\top \left(\nabla f(\mathbf{w}_K^{(t)}) - \nabla f\left(\mathbf{x}_t\right)\right)\right]^\top. \end{gathered} \tag{2.29}$$

Then the secant equations written in matrix form: $\mathbf{Z}_t\mathbf{Q}_t = \mathbf{G}_t$ will serve as an approximation to the system of equations (2.28). Adding the natural symmetry condition of the matrix $\mathbf{Q}_t$ and multiplying the secant equation in a matrix form by $\mathbf{Q}_t^{-1}$ on the right,

we obtain the following system of equations with constraints:

$$\begin{cases} \mathbf{G}_t \mathbf{Q}_t^{-1} = \mathbf{Z}_t \\ \mathbf{Q}_t^{-1} = \left[\mathbf{Q}_t^{-1}\right]^\top. \end{cases} \tag{2.30}$$

The following Lemma suggests a way to solve the obtained system.

Lemma 5. *Solution of the system* (2.30) *is given by the following formula*

$$\begin{aligned} \hat{\mathbf{Q}}^{-1} = \mathbf{G}_t^+ \mathbf{Z}_t \\ + \left(\mathbf{I} - \mathbf{G}_t^+ \mathbf{G}_t\right) \left(\mathbf{G}_t^+ \mathbf{Z}_t\right)^\top, \end{aligned} \tag{2.31}$$

where $\mathbf{G}_t^+$ *— pseudoinverse of* $\mathbf{G}$*:* $\mathbf{G}_t^+ = \mathbf{G}_t^\top \left(\mathbf{G}_t \mathbf{G}_t^\top\right)^{-1}$.

Proof. (of Lemma 5) Begin with substituting matrix $\mathbf{G}_t^\top$ instead of $\mathbf{A}$ and $\mathbf{Z}_t$ instead of the matrix $\mathbf{B}$ in the conditions of the theorem 1 . Note that the equality $\mathbf{G}_t^\top \mathbf{Z}_t = \mathbf{Z}_t^\top \mathbf{G}_t$ is satisfied since q_t is quadratic, and the pseudoinverse $\mathbf{G}_t^+$ satisfies the condition $\mathbf{G}_t \mathbf{G}_t^+ = \mathbf{I}$ by definition. Thus, the conditions of the Theorem 1 are executed and the system (2.30) solution is given by the equation (2.31). □

Remark 7. *Computational complexity of the formula* (1.8) *is* $\mathcal{O}\left(m^2K + mK^2 + m^3\right)$ *in a number of operations and* $\mathcal{O}\left(K^2 + mK + m^2\right)$ *in occupied memory:*

- *to calculate* $\left(\mathbf{G}_t \mathbf{G}_t^\top\right)^{-1}$ *first one can get a complete singular decomposition of the matrix* $\mathbf{G}_t \in \mathbb{R}^{m \times K}$*,* $m \leqslant K$ *which requires* $\mathcal{O}\left(mK^2 + m^3\right)$ *operations and* $\mathcal{O}\left(K^2 + mK + m^2\right)$ *memory (see [107], summary table 8.6.1 on page 493), having obtained it, it is enough to square and reverse the diagonal matrix with singular numbers, which will require only* $\mathcal{O}(m)$ *operations;*
- *rest of other matrix calculus utilize* $\mathcal{O}\left(m^2K + mK^2\right)$ *in operations and* $\mathcal{O}\left(mK + m^2\right)$ *in memory.*

2.2.3 Subspace step via the quasi-Newton direction reconstruction

Consider a system similar to (2.30) but omitting the $\mathbf{Q}_t$ matrix symmetry requirement:

$$\mathbf{G}_t \mathbf{Q}_t^{-1} = \mathbf{Z}_t, \tag{2.32}$$

Note that the required value is not the matrix $\mathbf{Q}_t^{-1}$, but its product on the gradient projection: $\mathbf{Q}_t^{-1}\mathbf{D}_t^\top\nabla f(\mathbf{x}_t)$. Multiplying the equation (2.32) by the gradient projection from the right the equation takes form:

$$\mathbf{G}_t\mathbf{z}_t = \mathbf{Z}_t\mathbf{D}_t^\top\nabla f(\mathbf{x}_t), \tag{2.33}$$

— a system of linear equations with respect to the desired vector $\mathbf{z}_t = \mathbf{Q}_t^{-1}\mathbf{D}_t^\top\nabla f(\mathbf{x}_t)$. It can be solved, for example, using the least squares method:

$$\widehat{\mathbf{z}}_t = \left(\mathbf{G}_t^\top\mathbf{G}_t\right)^{-1}\mathbf{G}_t^\top\mathbf{Z}_t\mathbf{D}_t^\top\nabla f(\mathbf{x}_t). \tag{2.34}$$

Remark 8. *Computational complexity of the equation* (2.34) *is* $\mathcal{O}\left(m^3 + m^2K\right)$ *in number of operations, where* $\mathcal{O}\left(m^3\right)$ *is due ot inversion of the* $m \times m$ *matrix* $\mathbf{G}_t^\top\mathbf{G}_t$*, and* m^2K *is due to* $\mathbf{G}_t^\top\mathbf{G}_t$ *and* $\mathbf{G}_t^\top\mathbf{Z}_t$ *multiplication.*

2.2.4 Hesse matrix estimation via regression

Consider the following system of equations

$$\mathbf{z}_i^{(t)\top}\mathbf{Q}_t\mathbf{z}_i^{(t)} = f(\mathbf{x}_{t-i}) - \nabla f(\mathbf{x}_t)^\top\mathbf{D}_t\mathbf{z}_i^{(t)} - f(\mathbf{x}_t), \quad i = 0\ldots K-1. \tag{2.35}$$

Note that the expression (2.35) is equivalent to the approximation q_t (2.26) and that the matrix $\mathbf{Q}_t$ enters the equation as linear term. Such equations can be solved by the least squares method. For the quadratic regression model, we introduce an operator that establishes a bijection between the space of symmetric matrices of size $m \times m$ and the vector space of dimension $\frac{m(m+1)}{2}$: $\operatorname{vech} : \mathbb{R}^{m\times m} \to \mathbb{R}^{\frac{m(m+1)}{2}}$, and corresponding inverse $\operatorname{vech}^{-1} : \mathbb{R}^{\frac{m(m+1)}{2}} \to \mathbb{R}^{m\times m}$:

$$\begin{aligned}\operatorname{vech}(\mathbf{M}) = \big[&\mathbf{M}_{1,1},\ldots,\mathbf{M}_{m,1},\mathbf{M}_{2,2},\\ &\ldots,\mathbf{M}_{m,2},\ldots,\mathbf{M}_{m-1,m-1},\mathbf{M}_{m,m-1},\mathbf{M}_{m,m}\big]^\top\\ \operatorname{vech}^{-1}(\mathbf{v}) = \mathbf{M}:\ &\mathbf{M}_{i,j} = \mathbf{v}_k,\\ &k = m(\min(i,j)-1) + \max(i,j) - 1.\end{aligned}$$

Consider vector $\mathbf{q}_t = \operatorname{vech}(\mathbf{Q}_t)$, vectors $\psi_i^{(t)} = \operatorname{vech}\left(\mathbf{z}_i^{(t)}\mathbf{z}_i^{(t)\top}\right)$ and scalars $y_i^{(t)} = f(\mathbf{x}_{t-i}) - \nabla f(\mathbf{x}_t)^\top\mathbf{D}_t\mathbf{z}_i^{(t)} - f(\mathbf{x}_t)$, $i = 1..K$. Having matrix of independent variables $\Psi_t = \left[\psi_1^{(t)},\ldots,\psi_K^{(t)}\right]^\top$ and vector of dependent variable values

$\mathbf{y}_t = \left(y_1^{(t)}, \ldots, y_K^{(t)}\right)$ we'll get a linear relation: $\Psi_t \mathbf{q}_t = \mathbf{y}_t$. It can be solved with a linear least squares method: $\widehat{\mathbf{q}}_{tt} = (\Psi_t^\top \Psi_t)^{-1} \Psi_t^\top \mathbf{y}_t$. Corresponding matrix can be obtained by applying vech^{-1}:

$$\widehat{\mathbf{Q}}_t = \text{vech}^{-1}\left((\Psi_t^\top \Psi_t)^{-1} \Psi_t^\top \mathbf{y}_t\right). \tag{2.36}$$

In the case where the function f is a polynomial of the second degree of the form (2.4) and the differences $\mathbf{x}_t - \mathbf{x}_{t-i}$ belong to the linear span of the columns of the matrix $\mathbf{D}_t$, then the matrix $\mathbf{D}_t^\top \nabla^2 f(\mathbf{x}_t) \mathbf{D}_t$ is the solution of the equation (2.35) due to the following equality:

$$f(\mathbf{x}_{t-i}) - \nabla f\left(\mathbf{x}_t\right)^\top \mathbf{D}_t \mathbf{z}_i^{(t)} - f(\mathbf{x}_t) = \frac{1}{2} \mathbf{z}_i^{(t)\top} \mathbf{D}_t^\top \mathbf{A} \mathbf{D}_t \mathbf{z}_i^{(t)}.$$

Remark 9. *Computational complexity of the equation* (2.36) *and formation of matrix* $\mathbf{\Psi}_t$ *and vector* $\mathbf{y}_t$ *are* $\mathcal{O}\left(m^2 K + m^4\right)$ *in memory and* $\mathcal{O}\left(m^4 K + m^6\right)$ *in number of operations.*

2.2.5 Subspaces construction based on the gradients history

Consider the subspace $\mathcal{D}_t$ given by the linear span of T preceding gradient values

$$\mathcal{D}_t = \text{span}\left\{\nabla f\left(\mathbf{x}_{t-T}\right), \ldots, \nabla f\left(\mathbf{x}_t\right)\right\}. \tag{2.37}$$

The following Lemma demonstrates the linear convergence rate over subspaces that include the last gradient value $\nabla f\left(\mathbf{x}_t\right)$.

Lemma 6. *Consider function* $f \in \mathcal{F}_{\mu,L}$, $\mathbf{D}_t \in \mathbb{R}^{d \times m}$, $m \geqslant 1$: $\mathbf{D}_t^\top \mathbf{D}_t = \mathbf{I}$ *and* $\mathbf{D}_t \mathbf{D}_t^\top \nabla f\left(\mathbf{x}_t\right) = \nabla f\left(\mathbf{x}_t\right)$. *Then*

$$\|(\mathbf{I} - \mathbf{D}_t^\top \mathbf{D}_t)(\mathbf{x}_t - \mathbf{x}_\star)\| \leqslant \left(1 - \frac{\mu}{L}\right) \|\mathbf{x}_t - \mathbf{x}_\star\|.$$

Proof. Note, that

$$\begin{aligned}\|\mathbf{D}_t^\top \mathbf{D}_t(\mathbf{x}_t - \mathbf{x}_\star)\| &\geqslant \left\|\|\nabla f\left(\mathbf{x}_t\right)\|^{-2} \nabla f\left(\mathbf{x}_t\right) \nabla f\left(\mathbf{x}_t\right)^\top (\mathbf{x}_t - \mathbf{x}_\star)\right\| \\ &\geqslant \frac{\|\nabla f\left(\mathbf{x}_t\right) \mu \|\mathbf{x}_t - \mathbf{x}_\star\|^2\|}{\|\nabla f\left(\mathbf{x}_t\right)\|^2} = \frac{\mu \|\mathbf{x}_t - \mathbf{x}_\star\|^2}{\|\nabla f\left(\mathbf{x}_t\right)\|} \geqslant \frac{\mu}{L} \|\mathbf{x}_t - \mathbf{x}_\star\|.\end{aligned}$$

It remains only to use the orthonormality of the columns $\mathbf{D}_t$

$$\begin{aligned}\|(\mathbf{I} - \mathbf{D}_t^\top \mathbf{D}_t)(\mathbf{x}_t - \mathbf{x}_\star)\| &= \|\mathbf{x}_t - \mathbf{x}_\star\| - \|\mathbf{D}_t^\top \mathbf{D}_t(\mathbf{x}_t - \mathbf{x}_\star)\| \\ &\leqslant \left(1 - \frac{\mu}{L}\right) \|\mathbf{x}_t - \mathbf{x}_\star\|\end{aligned}$$

□

Thus the Lemma 6 demonstrates linear convergence for subspaces of the form (2.37) in terms of the Theorem 3 if the $\mathbf{D}_t$ matrix columns form an orthonormal basis of the space $\mathcal{D}_t$.

Consider the recurrent Algorithm 2 for constructing the $\mathbf{D}_t$ matrix columns of which form the set (2.37).

Algorithm 2 updateSubspace1($\mathbf{D}$, $\mathbf{g}$, m)

$\mathbf{m_0} \leftarrow$ # of columns in $\mathbf{D}$
$\mathbf{g} \leftarrow \frac{\mathbf{g}}{\|\mathbf{g}\|}$
for i from 1 to m_0 **do**
    $\mathbf{D}_{\cdot,i} \leftarrow \mathbf{D}_{\cdot,i} - \mathbf{D}_{\cdot,i}^\top \mathbf{g}$
    $\mathbf{D}_{\cdot,i} \leftarrow \frac{\mathbf{D}_{\cdot,i}}{\|\mathbf{D}_{\cdot,i}\|}$
if $m_0 < m$ **then**
    $\mathbf{D} \leftarrow [\mathbf{D}_{\cdot,1}, \ldots, \mathbf{D}_{\cdot,m_0}, \mathbf{g}]$
else
    $\mathbf{D} \leftarrow [\mathbf{D}_{\cdot,2}, \ldots, \mathbf{D}_{\cdot,m_0}, \mathbf{g}]$
return $\mathbf{D}$

Thanks to the condition 6 in the Algorithm 2 the number of columns of the matrix $\mathbf{D}$ will grow first $m - 1$ iterations until it reaches the value m. Note that if vectors $\mathbf{g}_1, \ldots, \mathbf{g}_m$ are passed successively to the algorithm, then the columns of the final matrix $\mathbf{D}$ will form an orthonormal basis of a linear span of the vectors $\{\mathbf{g}_i\}_1^m$.

Remark 10. *Computational complexity of the Algorithm 2 is* $\mathcal{O}(mn)$ *in memory and* $\mathcal{O}(mn)$ *in a number of operations.*

2.3 Sequential subspace optimization methods

Using the results obtained in the previous sections, we present specific implementations of two sequential subspace optimization methods, as well as explore their properties: computational complexity and convergence rate.

2.3.1 Corrective SSO method

Algorithm 3 SSO_correction($\mathbf{x}_0$, ρ, L, subspaceUpdate, subspaceSearch)

$\mathbf{D}_t \leftarrow$ subspaceUpdate($\mathbf{D}_{t-1}$, $\nabla f(\mathbf{x}_t)$)
$\mathbf{z}_t \leftarrow$ subspaceSearch(f, $\mathbf{D}_t$, $\mathbf{x}_t$)
$\mathbf{d}_t \leftarrow \mathbf{D}_t \mathbf{z}_t$
$\gamma_t \leftarrow \frac{\mathbf{d}_t^\top \nabla f(\mathbf{x}_t)}{\|\nabla f(\mathbf{x}_t)\|^2}$
if $\gamma_t = 0$ **then**
    $\mathbf{d}_t \leftarrow \nabla f(\mathbf{x}_t)$
    $\alpha_t \leftarrow \frac{1}{L}$
else
    $\beta_t \leftarrow \frac{\|\mathbf{d}_t/\gamma_t - \nabla f(\mathbf{x}_t)\|}{\|\nabla f(\mathbf{x}_t)\|}$
    if $\beta_t > \frac{\rho}{L}$ **then**
        $\mathbf{d}_t \leftarrow \gamma_t \nabla f(\mathbf{x}_t) + \frac{\rho}{L\beta_t}(\mathbf{d}_t - \gamma_t \nabla f(\mathbf{x}_t))$
    $\alpha_t \leftarrow \frac{1-\rho}{\gamma_t(L^2+\rho^2)}$
$\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \alpha_t \mathbf{d}_t$
$t \leftarrow t+1$; go to 1

In the Algorithm 1 a general scheme of sequential subspace optimization algorithms is presented. In such a broad formulation, it is difficult to guarantee any good behavior of the optimization process. However, if we modify the existing optimization process, at least a linear convergence rate can be achieved. Thus, Algorithm 3 presents a *corrective* SSO method is presented, that is based on the Theorem 2: if the difference between the SSO method step and the gradient is too large this difference is leveled (al-

gorithm lines 10–11). Moreover, if the SSO method step is orthogonal to gradient, then it will be skipped and a step in the antigradient direction will be taken instead (lines 5–7).

Remark 11. *Sequence* $\{\mathbf{x}_t\}_{t\geqslant 0}$ *generated by the Algorithm 3 satisfies the conditions of the Theorem 2 for any procedures subspaceUpdate and subspaceSearch.*

2.3.2 Quasi-Newton SSO method

In the Algorithm 4 a sequential subspace optimization method that combines the quasi-Newtonian step and the conjugate gradient method step is given. It is based on the subspace construction method (matrices $\mathbf{D}_t$) described in 2.2.5 and the quasi-Newton direction estimation method outlined in 2.2.3.

Algorithm 4 L-QNSSO_step(f, $\mathbf{x}_0$, $\mathbf{D}_{t-1}$, m)

1: $\mathbf{D}_t \leftarrow \text{subspaceUpdate1}\left(\mathbf{D}_{t-1}, \nabla f\left(\mathbf{x}_t\right), m\right)$
2: **if** $t = 0$ **then**
3: $\quad \mathbf{d}_t \leftarrow -\nabla f\left(\mathbf{x}_t\right)$; go to step 11
4: $T \leftarrow \min(t, m-1)$
5: $\mathbf{Z}_t \leftarrow [\mathbf{D}_t^\top(\mathbf{x}_t - \mathbf{x}_{t-1}), \ldots, \mathbf{D}_t^\top(\mathbf{x}_t - \mathbf{x}_{t-T})]^\top$
6: $\mathbf{G}_t \leftarrow [\mathbf{D}_t^\top(\nabla f_t - \nabla f_{t-1}), \ldots, \mathbf{D}_t^\top(\nabla f_t - \nabla f_{t-T})]^\top$
7: $\mathbf{d}_t^{QN} \leftarrow -\mathbf{D}_t\left(\mathbf{G}_t^\top\mathbf{G}_t\right)^{-1}\mathbf{G}_t^\top\mathbf{Z}_t\mathbf{D}_t^\top\nabla f\left(\mathbf{x}_t\right)$
8: $\beta_t \leftarrow \frac{\nabla f(\mathbf{x}_t)^\top(\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1}))}{\nabla f(\mathbf{x}_{t-1})^\top\nabla f(\mathbf{x}_{t-1})}$
9: $\mathbf{d}_t^{CG} \leftarrow \left(\mathbf{I} - \mathbf{D}_t\mathbf{G}_t^\top\left(\mathbf{G}_t\mathbf{G}_t^\top\right)^{-1}\mathbf{G}_t\mathbf{D}_t^\top\right)\left(-\nabla f\left(\mathbf{x}_t\right) + \beta_t\mathbf{d}_{t-1}\right)$
10: $\mathbf{d}_t \leftarrow \mathbf{d}_t^{QN} + \mathbf{d}_t^{CG}$
11: $\alpha_t \leftarrow \operatorname{argmin} f(\mathbf{x}_t + \alpha\mathbf{d}_t)$
12: $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t + \alpha_t\mathbf{d}_t$
13: $t \leftarrow t + 1$; go to 1

Remark 12. *On every iteration of the Algorithm 4 step direction vector* $\mathbf{d}_t$ *belongs to a linear span of the corresponding* $\mathbf{D}_t$ *matrix columns:* $\mathbf{d}_t \in \mathcal{D}_t = \text{span}\{\nabla f\left(\mathbf{x}_{t-\min(m-1,t)}\right), \ldots, \nabla f\left(\mathbf{x}_t\right)\}$.

Remark 13. *Let $f \in \mathcal{F}_{\mu,L}$ be a quadratic function. Then, the sequence of estimates $\mathbf{x}_0, \ldots, \mathbf{x}_t, \ldots$ generated by the Algorithm 4, converges to $\mathbf{x}_\star$ with a linear convergence rate in at most n steps.*

Proof. Note the first step of the Algorithm 4 equals to the first step of the conjugate gradient method: $\mathbf{d}_0 = \mathbf{d}_0^{CG} = -\nabla f(\mathbf{x}_0)$. Suppose that $\mathbf{d}_j = \mathbf{d}_j^{CG}\ \forall j = 0 \ldots t-1$ and using the fact that for the conjugate gradient method all previous conjugate directions are orthogonal to the current gradient value $\nabla f(\mathbf{x}_j)^\top \mathbf{d}_i^{CG} = 0\ \forall\ 0 \leqslant i < j$, we get $\mathbf{d}_t^{QN} = 0$, since

$$\begin{aligned}
&\mathbf{Z}_t \mathbf{D}_t^\top \nabla f(\mathbf{x}_t) = \\
&\quad = \left[(\mathbf{x}_t - \mathbf{x}_{t-1})\mathbf{D}_t \mathbf{D}_t^\top \nabla f(\mathbf{x}_t), \ldots, (\mathbf{x}_t - \mathbf{x}_{t-T})\mathbf{D}_t \mathbf{D}_t^\top \nabla f(\mathbf{x}_t)\right] \\
&\quad = \left[\mathbf{d}_{t-1}^\top \nabla f(\mathbf{x}_t), \ldots, \mathbf{d}_{t-T} \nabla f(\mathbf{x}_t)\right] = \mathbf{0}.
\end{aligned}$$

Due to conjugate directions property $\mathbf{d}_i^{CG\,\top} \nabla^2 f\, \mathbf{d}_j^{CG} = 0, \forall\ 0 \neq j$:

$$\begin{aligned}
&\mathbf{G}_t \mathbf{D}_t^\top \left(\nabla f(\mathbf{x}_t) + \beta_t \mathbf{d}_{t-1}\right) = \\
&\qquad = \sum_{j=0}^{t-1} c_j \mathbf{d}_j^{CG\,\top} \nabla^2 f\ \left(\nabla f(\mathbf{x}_t) + \beta_t \mathbf{d}_{t-1}\right) = 0.
\end{aligned}$$

For the quadratic case, directions generated by the 4 coincide with the directions of the conjugate gradient method, and therefore have the same properties, including convergence in at most n iterations with linear convergence rate according to the Proposition 3. □

Thus, for the Algorithm 4 finite convergence in the quadratic case is demonstrated. A corrective sequential subspace optimization method 3 can be used to ensure linear convergence in the strongly convex case.

Let's discuss the structure of the Algorithm 4 in more detail. The matrices $\mathbf{G}_t$ and $\mathbf{Z}_t$ accumulate information about the curvature of the function, thereby making a quasi-Newtonian step along $\mathcal{G}_t = \operatorname{span}\{\nabla f_t - \nabla f_{t-1}, \ldots, \nabla f_t - \nabla f_{t-T})\}$ — linear span of the truncated history of the gradient differences, and step by conjugate gradient method except contained in $\mathcal{G}_t$. Thus, a step in the quasi-Newtonian direction *can* neutralize the error accumulated by the conjugate gradient method contained in $\mathcal{G}_t$.

2.4 Modified sign–perturbed sums method

The sign–perturbed sums method proposed in [41], gives an exact confidence region under the condition of independent identically distributed noises symmetric with respect to zero. This restriction makes the SPS method inapplicable under conditions set in Section 1.3.3. In this section a modification of SPS called *modified sign–perturbed sums* (MSPS) method is proposed. In MSPS method symmetry of the distribution of the inputs of the system around a known mean is used instead of the noises symmetry and centerness. It is worth noting that similar ideas are used in some other randomized methods that work under almost arbitrary interference (see [36]).

Consider the problem of estimating the parameter of a linear input-output model with additive noise in the output observations:

$$y_i = \boldsymbol{\varphi}_i^T \mathbf{x}_\star + \varepsilon_i, \quad i = 1..N,$$

where N — number of observations, $y_i \in \mathbb{R}$ — observable system outputs, $\boldsymbol{\varphi}_i \in \mathbb{R}^n$ — observable system inputs, $\mathbf{x}_\star \in \mathbb{R}^n$ — unknown parameter vector and $\varepsilon_i \in \mathbb{R}$ — unknown additive noise. Suppose the input vectors $\boldsymbol{\varphi}_i$ are independent and identically distributed around the known mean $m_\varphi \in \mathbb{R}^n$. Consider the following notation:

$$S_k(\mathbf{x}) = \sum_{i=1}^{N} a_{i,k} \Delta_i (y_i - \boldsymbol{\varphi}_i \mathbf{x}), \tag{2.38}$$

$$S_0(\mathbf{x}) = \sum_{i=1}^{N} \Delta_i (y_i - \boldsymbol{\varphi}_i \mathbf{x}), \tag{2.39}$$

where $\Delta_i = \boldsymbol{\varphi}_i - m_\varphi$ — centerd input vectors, $a_{i,k}$ — so-called *random signs*, defined by the following distribution

$$a_{k,i} = \begin{cases} 1 & \text{ñ âåðîÿòíîñòüþ } \frac{1}{2}, \\ -1 & \text{ñ âåðîÿòíîñòüþ } \frac{1}{2}. \end{cases}$$

By analogy with the *sign–perturbed sums* introduced in [41], $S_k(\mathbf{x})$ are referred to as *modified sign–perturbed sums*. The modification is to introduce *randomized* multipliers Δ_i.

In the Algorithm 5 a method of modified sign–perturbed sums based on the sums (2.38) and (2.39) is stated in the form of pseudocode. It provides a tool for

Algorithm 5 MSPS($\{y_i\}_{i=1}^N$, $\{\varphi\}_{i=1}^N$, m_φ, M, q, $\mathbf{x}$)

for i in $1 \dots N$ **do**
  $n_i \leftarrow y_i - \varphi_i^T \cdot \mathbf{x}$
  $\Delta_i \leftarrow \varphi_i - m_\varphi$
for k in $1 \dots M-1$ **do**
  $S_k \leftarrow 0 \in \mathbb{R}^p$
  for i in $1 \dots N$ **do**
    $S_k \leftarrow S_k + \Delta_i \cdot n_i \cdot RANDOMSIGN()$
  $Z_k \leftarrow ||S_k||_2^2$
$S_0 \leftarrow 0 \in \mathbb{R}^p$
for i in $1 \dots N$ **do**
  $S_0 \leftarrow S_0 + \Delta_i \cdot n_i$
$Z_0 \leftarrow ||S_0||_2^2$
$r \leftarrow 0$
for k in $1 \dots M-1$ **do**
  if $Z_0 > Z_k$ **then**
    $r \leftarrow r + 1$
if $r >= q$ **then**
  return $FALSE$
else
  return $TRUE$
end procedure

constructing an exact confidence region of a linear model parameter (1.15) $\mathbf{x}^\star$ under assumptions 1.3.3. The main idea of the given Algorithm 5 is that from the independence and symmetry of the distribution of Δ_i follows the coincidence of the distributions of the sums $S_k(\mathbf{x}^\star) = \sum_{i=1}^N a_{k,i}\Delta_i\varepsilon_i$, $k = 0, \dots, M-1$. Thus, denoting $Z_k = ||S_k(\mathbf{x}^\star)|/$, the random variable Z_0 will take any position in the ascending sorted list $Z_{(0)}, \dots, Z_{(M-1)}$ with probability $\frac{1}{M}$. Therefore, Z_0 will *not* be among $q \leqslant M$ greatest values $\{Z_k\}_{k=0}^{M-1}$ with *exact* probability $1 - \frac{q}{M}$.

Formally, the above argument is proved in the following theorem. It states that the set of values $\mathbf{x}$ for which the Algorithm 5 returns $TRUE$ form an exact $(1 - \frac{q}{M})$-confidence region for the parameter $\mathbf{x}^\star$.

Theorem 4. *Consider a linear model with additive noise* (1.15). *Let the inputs* $\{\boldsymbol{\varphi}_i\}_1^N$ *and the noises* $\{\varepsilon_i\}_1^N$ *be mutually independent. Suppose that the inputs* $\{\boldsymbol{\varphi}_i\}_1^N$ *are identically and symmetrically distributed with respect to the known vector* $m_\varphi \in \mathbb{R}^n$. *Fix* $M \geqslant q \geqslant 0$ *— two natural numbers,* $\{S_k(\mathbf{x})\}_{k=0}^{M-1}$ *— M of modified sign-perturbed sums obtained via the formulas* (2.38) *and* (2.39). *Denote* $Z_k(\mathbf{x}) = ||S_k(\mathbf{x})||_2^2$ *and*

$$\mathcal{X}_{M,q} = \{\mathbf{x} \in \mathbb{R}^p : |\{k \in \{1, \ldots, M-1\} : Z_0(\mathbf{x}) < Z_k(\mathbf{x})\}| \geqslant 1\}. \tag{2.40}$$

Then $\mathcal{X}_{M,q}$ *is an exact* $(1 - \frac{q}{M})$*-confidence region for parameter* $\mathbf{x}^\star$

$$P_{\varepsilon_i, a_{k,i}}(\mathbf{x}^\star \in \mathcal{X}_{M,q}) = 1 - \frac{q}{M}.$$

Proof. It is sufficient to demonstrate that $\mathbf{x}^\star \in \mathcal{X}_{M,q}$ with probability $1 - \frac{q}{M}$. That is tantamount, $Z_0(\mathbf{x}^\star)$ is less than at least q different $Z_k(\mathbf{x}^\star)$ from $\{Z_k(\mathbf{x}^\star)\}_{k=1}^{M-1}$ with probability $1 - \frac{q}{M}$. In turn, it holds true if $\{Z_k(\mathbf{x}^\star)\}_{k=0}^{M-1}$ random variables are uniformly ordered. Note, that $Z_0(\mathbf{x}^\star)$ and $Z_k(\mathbf{x}^\star)$ can be expressed in the following way:

$$Z_k(\mathbf{x}^\star) = ||\sum_{i=1}^N a_{k,i}\Delta_i\varepsilon_i||, \quad Z_0(\mathbf{x}^\star) = ||\sum_{i=1}^N \Delta_i\varepsilon_i||.$$

According to the Statement 5 Δ_i and $a_{k,i}\Delta_i$ are i.i.d $\forall\, i = 1..N$, $k = 0..M-1$. Therefore, the values $\{Z_k(\mathbf{x}^\star)\}_{k=0}^{M-1}$ are independently and equally distributed. In addition, they are distributed continuously and therefore uniformly ordered by virtue of the Δ_i distribution continuity and the Statement 6. Thus, $Z_0(\mathbf{x}^\star)$ is not among the q largest values of $\{Z_k(\mathbf{x}^\star)\}_{k=1}^{M-1}$ with probability $1 - \frac{q}{M}$. □

The following remark characterizes the applicability of the Algorithm 5.

Remark 14. *The proof of the theorem 4 is substantially based on the random nature of the input vectors* $\{\boldsymbol{\varphi}_i\}_1^N$*: a probability that the true parameter value* $\mathbf{x}^\star$ *belongs to the confidence region* $\mathcal{X}_{M,q}$ *in the Theorem 4 is measured by random signs* $\{\alpha_{k,i}\}$ *and by the input vectors* $\{\boldsymbol{\varphi}_i\}_1^n$. *Thus, the scope of the Algorithm 5 refer to situations in which model inputs are randomly generated, their distribution is symmetric, and their mean is known — for example, generated by the experimenter.*

The Theorem 4 and the Algorithm 5 given in this section are substationally based on the sign-perturbed sums method and the corresponding theorem given in [41]. Despite this, the described modified sign-perturbed method uses an excellent idea of inputs

randomization, which allows to expand the scope of the algorithm, loosening restrictions on the nature of interference.

It is worth noting that the confidence region $\mathcal{X}_{M,q}$ obtained by the modified sign–perturbed sums method is the solution for the problem posed in the Section 1.3.3 with $\alpha = 1 - \frac{q}{M}$.

2.5 Properties of the confidence region of the MSPS method

This section demonstrates some properties of the confidence region $\mathcal{X}_{M,q}$ obtained by the Algorithm 5.

The explicit expression of a confidence region by the formula (2.40) is practically inconvenient, since it requires the use of approximate numerical procedures. In addition, the absence of an analytical expression of the boundaries of the confidence region $\mathcal{X}_{M,q}$ makes it difficult to study their theoretical properties.

Remark 15. *Under the conditions of the Theorem 4, region $\mathcal{X}_{M,q}$, defined by the formula (2.40), admits the following expression:*

$$\mathcal{X}_{M,q} = \bigcup_{\mathbb{I}\subset\{1,..,M-1\}:|\mathbb{I}|=q} \left(\bigcap_{k\in\mathbb{I}} \{\mathbf{x} : Z_0(\mathbf{x}) < Z_k(\mathbf{x})\} \right) \neq \varnothing$$

Remark 16. *It is sufficient to note that region $\mathcal{X}_{M,q}$ consists of points $\mathbf{x}$ for which $Z_0(\mathbf{x})$ is less than at least q of different $Z_k(\mathbf{x})$, i.e. at least q of different inequalities $Z_0(\mathbf{x}) < Z_k(\mathbf{x})$, $k = 1..M-1$ holds true*

The importance of the Remark 15 is that the computation of the region $\mathcal{X}_{M,q}$ can be reduced to the computation of simpler sets $\{\mathbf{x} : Z_0(\mathbf{x}) < Z_k(\mathbf{x})\}$ in the multidimensional case.

Consider the one dimensional case of a linear model (1.15)

$$y_i = \varphi_i x + \varepsilon, \qquad i = 1, \ldots, N, \tag{2.41}$$

where $y_i, \varphi_i, x, \varepsilon_i \in \mathbb{R}\ \forall\ i = 1, \ldots, N$. The following lemma provides an analytical expression of the sets $\{\mathbf{x} : Z_0(\mathbf{x}) < Z_k(\mathbf{x})\}$ for one dimensional case.

Lemma 7. *In the above expressions, let $d = 1$. Then set $\{x : Z_0(x) < Z_k(x)\}$ can be expressed in the following form:*

$$\{x : Z_0(x) < Z_k(x)\} = (B_k^{min}, B_k^{max}), \quad \textit{where}$$

$$B_k^{(1)} = \frac{\sum_{i=1}^N (1 - a_{k,i})\varphi_i y_i}{\sum_{i=1}^N (1 - a_{k,i})\varphi_i^2}, \qquad B_k^{(2)} = \frac{\sum_{i=1}^N (1 + a_{k,i})\varphi_i y_i}{\sum_{i=1}^N (1 + a_{k,i})\varphi_i^2},$$

$$B_k^{min} = \min(B_k^{(1)}, B_k^{(2)}), \qquad B_k^{max} = \max\left(B_k^{(1)}, B_k^{(2)}\right).$$

Proof. Consider inequality $Z_0(x^\star) < Z_k(x^\star)$

$$(\sum_{i=1}^N \Delta_i(y_i - \varphi_i x))^2 < (\sum_{i=1}^N a_{k,i}\Delta_i(y_i - \varphi_i x))^2$$

$$(\sum_{i=1}^N \Delta_i(y_i - \varphi_i x))^2 - (\sum_{i=1}^N a_{k,i}\varphi_i(y_i - \varphi_i x))^2 < 0,$$

$$(\sum_{i=1}^N (1 - a_{k,i})\Delta_i(y_i - \varphi_i x))(\sum_{i=1}^N (1 + a_{k,i})\Delta_i(y_i - \varphi_i x)) < 0,$$

$$\left(\frac{\sum_{i=1}^N (1 - a_{k,i})\Delta_i(\varepsilon_i + \Delta_i x^\star)}{\sum_{i=1}^N (1 - a_{k,i})\Delta_i^2} - x\right) \cdot$$

$$\cdot \left(\frac{\sum_{i=1}^N (1 + a_{k,i})\Delta_i(\varepsilon_i + \Delta_i x^\star)}{\sum_{i=1}^N (1 + a_{k,i})\Delta_i^2} - x\right) < 0.$$

□

It is worth noting that the analytic expression of set boundaries $\{x : Z_0(x) < Z_k(x)\}$ obtained by the Lemma (7) is significantly more efficient than pointwise computation via the Algorithm 5.

Lemma 8. *Let B_k be either $B_k^{(1)}$ or $B_k^{(2)}$ and in addition to 1.3.3 the following assumptions holds true:*

- *the second and fourth moments of the inputs $\{\varphi_i\}$ exists;*
- *$\{\varepsilon_i\}$ is either a random variable with a second non-centered moment bounded by the constant $C < \infty$, or a deterministic sequence uniformly bounded by the constant $\sqrt{C}$.*

$$E_{\varepsilon,\varphi,a}[B_k] = x^\star,$$

$$E_{\varepsilon,\varphi,a}[B_k - E_{\varepsilon,\varphi,a}[B_k]]^2 \xrightarrow[N\to\infty]{} 0,$$

where $E_{\varepsilon,\varphi,a}$ *— expected value of joint distribution of random values* $\{\varepsilon_i\}_{i=1}^N, \{\varphi_i\}_{i=1}^N$ *è* $\{a_i\}_{i=1}^N$.

Proof. Consider expected value of the boundaries

$$E_{\varepsilon,\varphi,a}B_{\{a_{k,i}\}} = E_{\varepsilon,\varphi,a}\left[\frac{\sum_{i=1}^N(1 \pm a_{k,i})\Delta_i(\varepsilon_i + \Delta_i x^\star)}{\sum_{i=1}^N(1 \pm a_{k,i})\Delta_i^2}\right] =$$

$$= x^\star + E_{\varepsilon,\varphi,a}\left[\frac{\sum_{i=1}^N(1 \pm a_{k,i})\Delta_i\varepsilon_i}{\sum_{i=1}^N(1 \pm a_{k,i})\Delta_i^2}\right].$$

Since Δ_i are independent with ε_i and $E\Delta_i = 0$:

$$E_{\varepsilon,\varphi,a}\left[\frac{\sum_{i=1}^N(1 \pm a_{k,i})\Delta_i\varepsilon_i}{\sum_{i=1}^N(1 \pm a_{k,i})\Delta_i^2}\right] = \sum_{i=1}^N|E_{\varepsilon,\varphi,a}\left[\frac{(1 \pm a_{k,i})\Delta_i\varepsilon_i}{\sum_{j=1}^N(1 \pm a_{k,j})\Delta_j^2}\right]| \leqslant$$

$$\leqslant \sum_{i=1}^N|E_{\varepsilon,\varphi,a}\left[\frac{(1 \pm a_{k,i})\Delta_i\varepsilon_i}{\sum_{j\neq i}^N(1 \pm a_{k,j})\Delta_j^2}\right]| = 0.$$

Thus the first part of the statement is proved. To prove the second part, note that $\forall\{x_i\}_1^N$ $2\sum_1^N x_i^2 \geqslant (\sum_1^N x_i)^2$ holds true. Hence,

$$E_{\varepsilon,\varphi,a}[B_{\{a_{k,i}\}} - E_{\varepsilon,\varphi,a}B_{\{a_{k,i}\}}]^2 \leqslant E_{\varepsilon,\varphi,a}\left[\frac{(\sum_{i=1}^N(1 \pm a_{k,i})\Delta_i\varepsilon_i)^2}{(\sum_{i=1}^N(1 \pm a_{k,i})\Delta_i^2)^2}\right] \leqslant$$

$$\leqslant CE_{\varepsilon,\varphi,a}\left[\frac{(\sum_{i=1}^N(1 \pm a_{k,i})\Delta_i\varepsilon_i)^2}{(\sum_{i=1}^N(1 \pm a_{k,i})\Delta_i)^4}\right] \xrightarrow[N\to\infty]{} 0.$$

Since ε_i are independent of Δ_i, the expectation on them can be taken separately. Thus, using the imposed restrictions:

$$E_{\varepsilon,\varphi,a}[B_{\{a_{k,i}\}} - E_{\varepsilon,\varphi,a}B_{\{a_{k,i}\}}]^2 \leqslant CE_{\varepsilon,\varphi,a}\left[\frac{(\sum_{i=1}^N(1 \pm a_{k,i})\Delta_i)^2}{(\sum_{i=1}^N(1 \pm a_{k,i})\Delta_i)^4}\right] =$$

$$= CE_{\varepsilon,\varphi,a}\left[\frac{1}{(\sum_{i=1}^N(1 \pm a_{k,i})\Delta_i)^2}\right] \xrightarrow[N\to\infty]{} 0.$$

That shows that the second statement is also true. □

Lemma 8 gives two important consequences. First, expected value of B_k — the interval $\mathcal{X}_{M,q}$ boundaries — equal to the true value of the parameter $x^\star$. Second, these bounds tend to $x^\star$ with increasing N. In summary, the confidence interval converges to $x^\star$ at $N \to \infty$. Thus, by analogy with parameter estimates, the confidence interval $\mathcal{X}_{M,q}$ is *consistent*.

Chapter 3. Comparative analysis of optimization and parameter estimation methods

3.1 MSPS method analysis using synthetic data

This section demonstrates several features of confidence regions obtained by the modified sign–perturbed sums method in comparison with the sign–perturbed sums method and the asymptotic confidence region obtained with the formula (1.16). Due to the fact that these methods have different areas of applicability: normally distributed noises, centered symmetrically distributed noises and arbitrary noises, — the results of comparison in each of these cases are predictable. Therefore, the following examples are purely illustrative.

3.1.1 Data modelling process description

In the conditions described in Section 1.3.3, the most interesting is the case of a small number of assumptions. For ease of visualization, consider the two-dimensional case $m = 2$. Put $\theta^\star = (-1, 2)^T$ and $\boldsymbol{\varphi}_i$ modelled in accordance to the following distribution:

$$\varphi_i \sim N(\mu_\varphi, \Sigma_\varphi), \quad \mu_\varphi = (1, -1)^T, \quad \Sigma_\varphi = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{pmatrix},$$

The number of observations N and the nature of the noises $\{\varepsilon_i\}$ remain unknown. Various combinations of which will be discussed in the following subsections.

3.1.2 Case of a large number of observations

Consider the following two types of noises:

1. *unbiased symmetric*: $\varepsilon_i \sim N(0,1)$,

2. *biased asymmetric*: $\varepsilon_i \sim Exp(\lambda = 1) + 4$.

Figure 3.1 (a) demonstrates the confidence regions for the case of the number of observations $N = 50$ and unbiased symmetric noise. All three confidence regions contain both the true value of the parameter and its least squares estimate. Note that the SPS and MSPS methods give confidence region much larger than region obtained by the formula (1.16). Figure 3.1 (b) demonstrates the confidence regions for the case of the number of observations $N = 50$ and the biased asymmetric noise. Despite the fact that the confidence region obtained by the MSPS method is much larger than the confidence regions obtained by other methods, it is the only one that contain the true parameter value. Confidence regions obtained by the SPS method and by the formula (1.16) are shifted along with the least squares estimate due to noise asymmetry and bias.

3.1.3 Case of a small number of observations

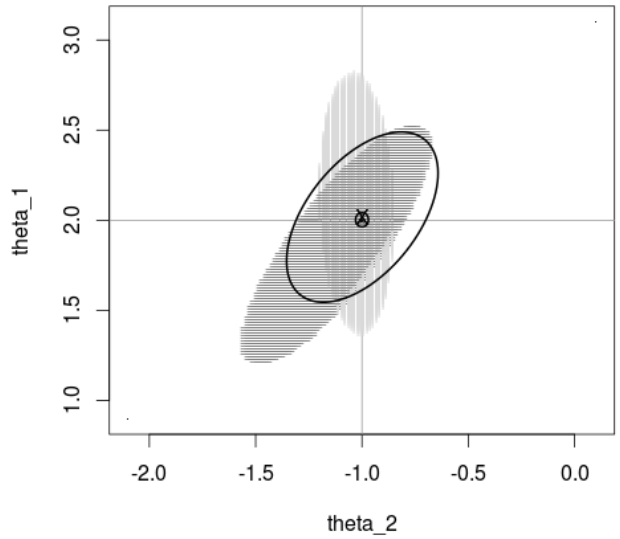
Consider the following two types of noises:

1. *unbiased asymmetric noise* $\varepsilon_i \sim Exp(\lambda = 1) - 1$,
2. *biased asymmetric noise* $\varepsilon_i \sim Exp(\lambda = 1) + 4$.

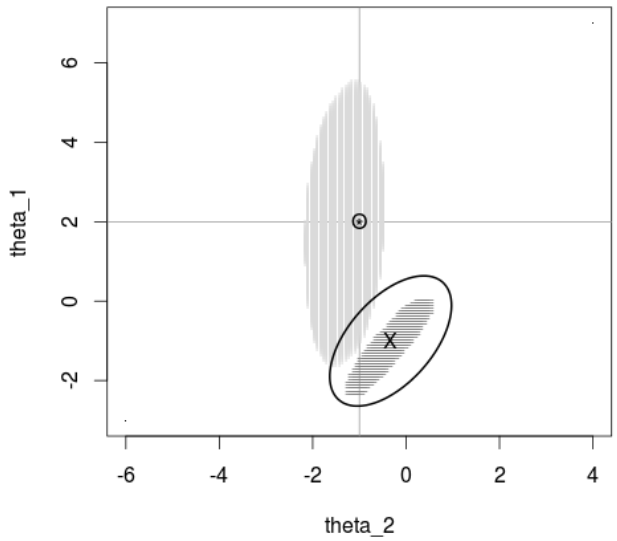
Figure 3.2 (a) demonstrates the confidence regions for the case of the number of observations $N = 15$ and unbiased asymmetric noise. It is noteworthy that the confidence region obtained by SPS method degenerates into an unbounded one. Figure 3.2 (b) demonstrates the confidence regions for the case of the number of observations $N = 15$ and the biased asymmetric noise. The confidence region of the MSPS method is the only one containing the true value of the parameter. In this case, the confidence region obtained by the SPS method is also unbounded.

3.1.4 Summary

In the previous sections, the confidence regions obtained by the SPS method MSPS method and by the formula (1.16) are illustrated for four cases: unbiased symmetric noise with a large number of observations, biased asymmetric noise with a small number of observations, unbiased asymmetric noise with a large number of observa-



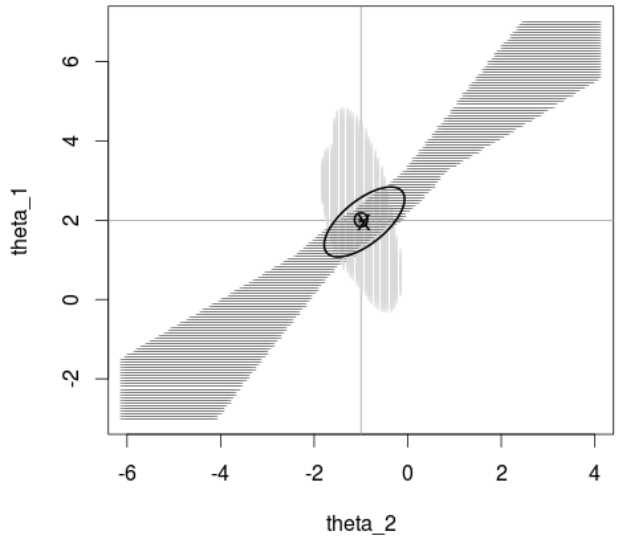
[Unbiased symmetric noise]



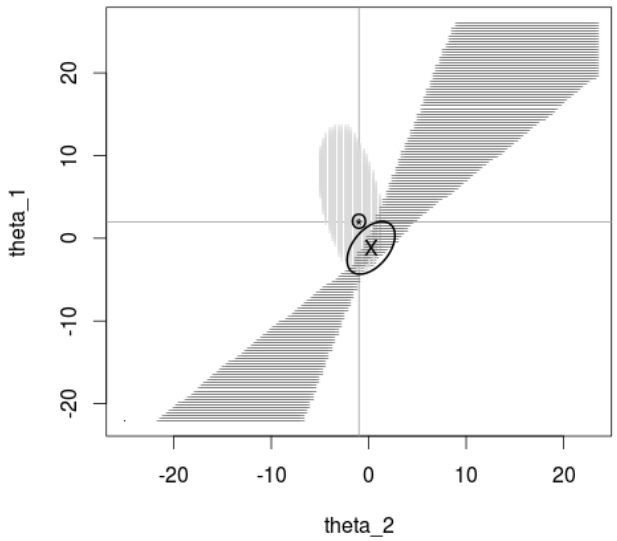
[Biased asymmetric noise]

Figure 3.1 – Confidence sets obtained by SPS (dark horizontal lines), MSPS (light vertical lines) and asymptotic (ellipsoid). The true parameter value is marked with a circle with a dot, the least squares estimate is marked with a cross

tions, and biased asymmetric noise with a small number of observations. The examples given are in agreement with the theoretical results obtained in Section 2.4: the modified sign–perturbed sums method yields an exact confidence region even under conditions of biased asymmetric noise and a small number of observations.



[Unbiased asymmetric noise] [Biased



asymmetric noise]

Figure 3.2 – Confidence sets obtained by SPS (dark horizontal lines), MSPS (light vertical lines) and asymptotic (ellipsoid). The true parameter value is marked with a circle with a dot, the least squares estimate is marked with a cross

3.2 Comparative analysis of optimization methods

This section provides a comparison of sequential subspace optimization methods and several common alternatives. The following optimization methods are considered

– CG: conjugate gradients method (1.10) with the β_t coefficient calculated via Polak–Ribiere–Polyak formula (1.11);

– BPCG: Beale–Powell conjugate gradient method (1.12), with the parameters c_1, c_2, c_3 set to 0.2, 0.8 and 1.2 respectively;
– L-BFGS(m): Broyden–Fletcher–Goldfarb–Shanno method, storing the history of the argument and gradient values for the last m steps only (see Algorithm 7.5 in [70]), with a unit matrix used as the initial approximation of the inverse Hesse matrix;
– L-QNSSO(m): Algorithm 4, storing the history of the argument and gradient values for the last m steps only.

To determine the step size for each of the methods, an approximate linear search was used, namely the Brent method [108]. All methods are implemented in Python programming language using linear algebra library NumPy [109] and scientific calculations library SciPy [110]. The source code is publicly available on the online GitHub service, repository address: https://github.com/obus/optimus,

3.2.1 Quadratic function

Consider the following functions

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} + \mathbf{b}^\top \mathbf{x}, \tag{3.1}$$

where $\mathbf{x} \in \mathbb{R}^n$, $n = 1000$, $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{b} = \mathbf{1} \in \mathbb{R}^n$ and the spectrum of the matrix $\mathbf{A}$ has the following form:

$$\rho(\mathbf{A}) = \left\{1 + \frac{i-1}{n-1}(\lambda_{max} - 1)\right\}_{i=1}^{n} \subset [1, \lambda_{max}]\,. \tag{3.2}$$

Consider the following evaluation method

1. pick initial point $\mathbf{x}_0$ from distribution $\mathcal{N}(\mathbf{x}_\star, \mathbf{I})$,
2. for each algorithm, run the optimization process, and calculate the minimum value of the function $f(\mathbf{x}_t)$ per thousand iterations,
3. repeat steps 1-2 a thousand times, thus obtaining a sample of $\min\limits_{0\leqslant t\leqslant n} f(\mathbf{x}_t)$ values for each algorithm,
4. repeat steps 1-3 for λ_{max} = 10, 100, 1000.

The distributions of the obtained samples are illustrated via boxplots on the Figures 3.3, 3.4 and 3.5 for the cases λ_{max} = 10, 100 and 1000 respectively. In all three

cases, the L-QNSSO(2) algorithm shows a similar or smaller error than the CG, BPCG and LBFGS analogs. Note that with the growth of λ_{max} L-QNSSO(2) is increasingly superior to the conjugate gradient method. This can be explained by the fact that ill-conditionality of the problem increases the error accumulated by the conjugate gradient method and deprives the corresponding directions of the conjugacy property. The L-QNSSO method eliminates this error due to an additional quasi-Newtonian step.

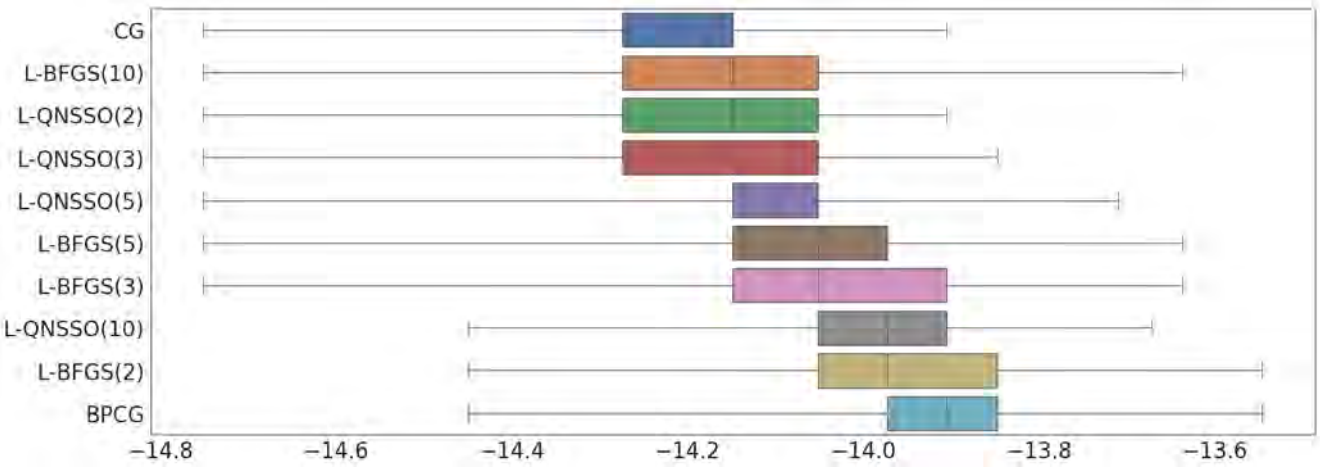


Figure 3.3 – Boxplots of $\min\limits_{0\leqslant t\leqslant n} f(\mathbf{x}_t)$ for the quadratic function (3.1) with $\Lambda_{max} = 10$. The value of the function in the logarithmic scale is marked along the abscissa axis. Left border, segment inside and right border of the box are 25%, 50% and 75% percentiles respectively, the whiskers correspond to the minimum and maximum values.

3.2.2 Rosenbrock function

Consider generalization of the Rosenbrock function [111] to the n–dimensinal case:

$$f(\mathbf{x}) = \sum_{i=1}^{n-1} \left[100(x_{i+1} - x_i^2)^2 + (1 - x_i)^2\right], \qquad -2.048 \leqslant x_i \leqslant 2.048 \quad \forall\, i = 1..n. \tag{3.3}$$

The function (3.3) global minimum is at $\mathbf{x}_\star = \mathbf{1}\ \forall n \geqslant 2$, and for $n \geqslant 4$ at least one local minimum is known to exist in the vicinity of $(-1, 1, \ldots, 1)$ [112].

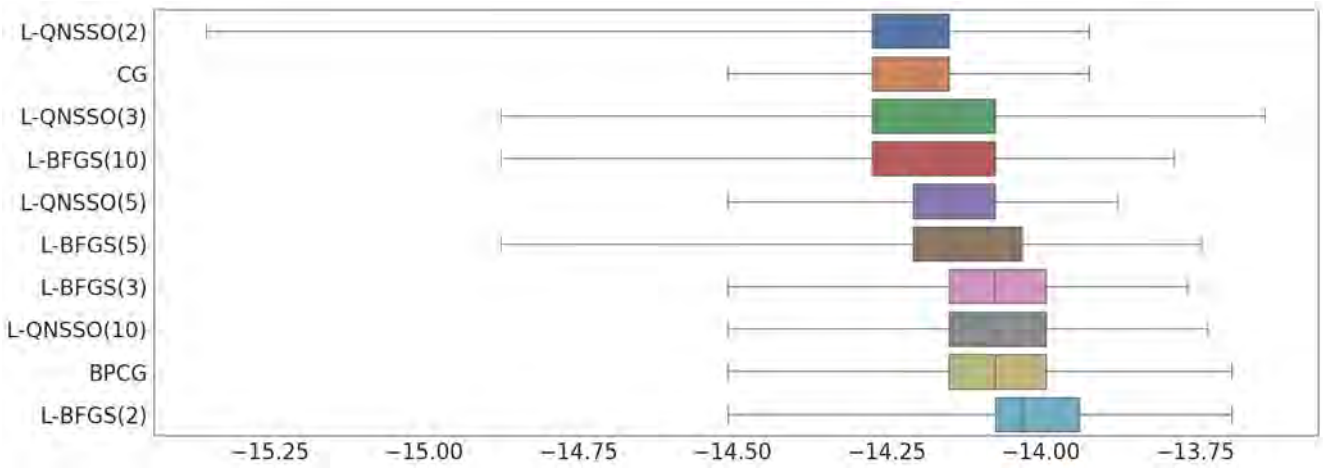


Figure 3.4 – Boxplots of $\min\limits_{0 \leqslant t \leqslant n} f(\mathbf{x}_t)$ for the quadratic function (3.1) with $\Lambda_{max} = 100$. The value of the function in the logarithmic scale is marked along the abscissa axis. Left border, segment inside and right border of the box are 25%, 50% and 75% percentiles respectively, the whiskers boundaries correspond to the minimum and maximum values.

In this case, it is assumed that the saddle point is in the vicinity of the point $(-0.555, 0.322, 0.115, 0.024, 0.011, 0.010, \ldots, 0.010, 0.0001)^\top$ [113]. In this regard, we will perform a comparison as follows:

1. sample initial point $\mathbf{x}_0$ from distribution

$$\begin{aligned} \mathbf{x}_0 &= (-0.555, |x_0^{(2)}|, \ldots, |x_0^{(n)}|)^\top, \\ x_0^{(i)} &\sim \mathcal{N}(0,1), \qquad \forall\, i = 2..n; \end{aligned} \tag{3.4}$$

2. for each algorithm, run the optimization process and calculate the minimum value of the function $f(\mathbf{x}_t)$ for T iterations;
3. repeat steps 1-2 a thousand times, thus obtaining a sample of $\min\limits_{0 \leqslant t \leqslant T} f(\mathbf{x}_t)$ values for each algorithm.

Distribution of the obtained samples are presented via boxplots at Figures 3.6 and 3.7 for a nubmer of iterations T equals50 and n respectively. The L-QNSSO(2) algorithm often achieves the best quality in the first fifty iterations. However, L-QNSSO method is inferior to the L-BFGS and BPCG methods in a larger number of iterations, although it shows results superior to the conjugate gradient method for T equals both 50 and n.

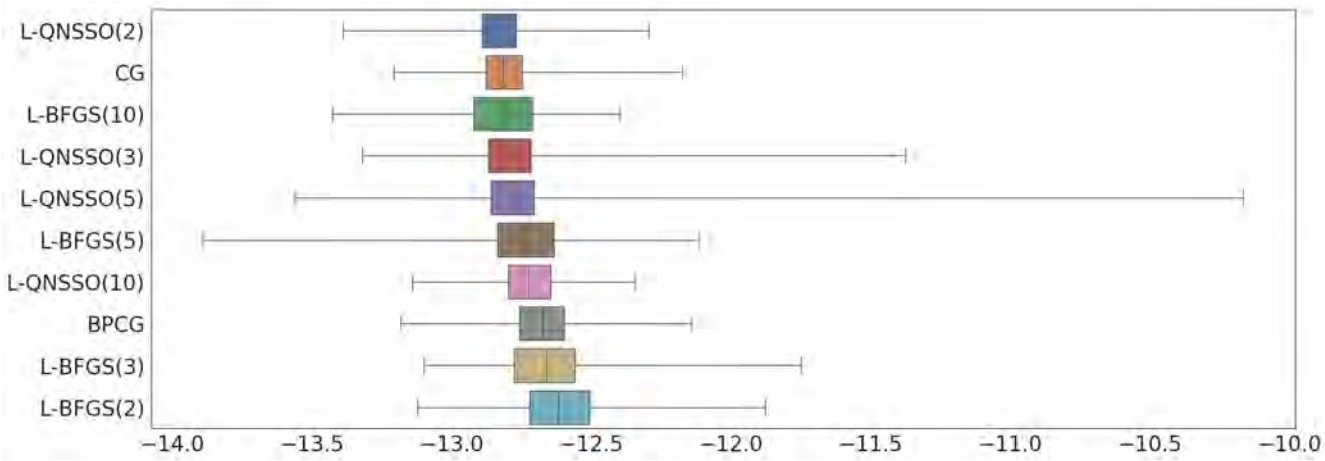


Figure 3.5 – Boxplots of $\min\limits_{0 \leqslant t \leqslant n} f(\mathbf{x}_t)$ for the quadratic function (3.1) with $\Lambda_{max} = 1000$. The value of the function in the logarithmic scale is marked along the abscissa axis. Left border, segment inside and right border of the box are 25%, 50% and 75% percentiles respectively, the whiskers boundaries correspond to the minimum and maximum values.

3.2.3 Linear regression with Tikhonov regularization

Consider a linear input-output model with additive noise similar to the model (1.15):

$$y_i = \boldsymbol{\varphi}_i^\top \mathbf{x}_\star + \varepsilon_i, \quad i = 1 \ldots N, \tag{3.5}$$

where $y_i \in \mathbb{R}$ — output scalars, $\boldsymbol{\varphi}_i \in \mathbb{R}^n$ — input vectors, $\mathbf{x} \in \mathbb{R}^n$ — unknown parameter vector, and noises are distributed in accordance to normal distribution: $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)\ \forall, i = 1 \ldots N$. Suppose that the number of observations is much smaller than the dimension of the parameter vector: $n \gg N$. Note that the model of the form (3.5) is relevant for many practical problems, including recommendation systems [45]. In such cases, the $\mathbf{x}_\star$ parameter estimation problem is often formulated as a linear regression problem with Tikhonov regularization, also known as ridge regression [60] — a special case of the pattern recognition problem (see Chapter 1.1).

$$f(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \boldsymbol{\varphi}_i^\top \mathbf{x})^2 + \lambda \|\mathbf{x}\|^2 \to \min_{\mathbf{x}}, \tag{3.6}$$

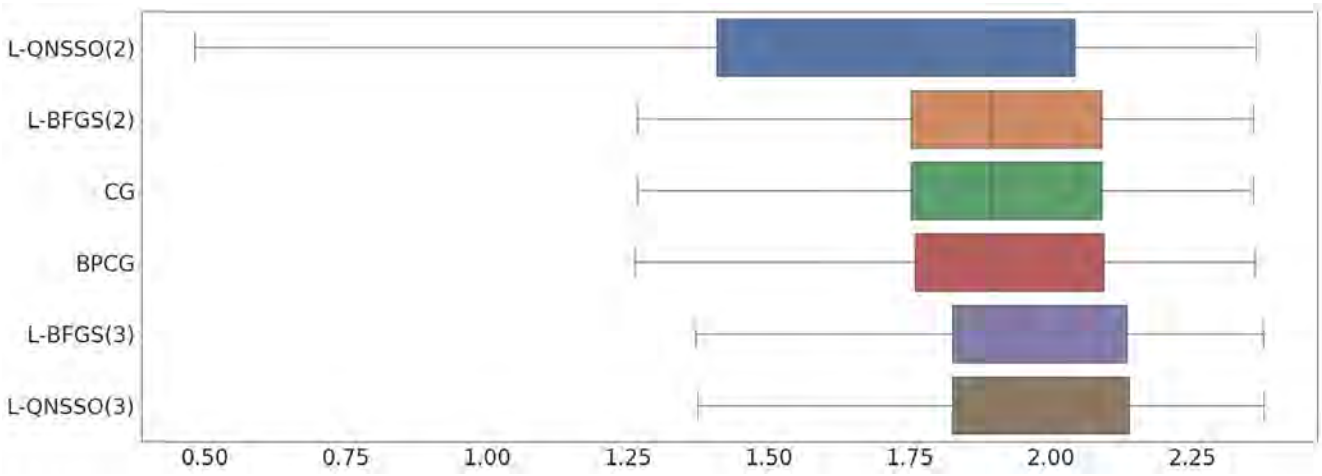


Figure 3.6 – Boxplots of $\min\limits_{0\leqslant t\leqslant 50} f(\mathbf{x}_t)$ for the Rosenbrock function (3.3). The value of the function in the logarithmic scale is marked along the abscissa axis. Left border, segment inside and right border of the box are 25%, 50% and 75% percentiles respectively, the whiskers boundaries correspond to the minimum and maximum values.

where $\lambda > 0$ — regularization coefficient. Such a problem formulation allows to obtain a unique solution (since function (3.6) strictly convex), and also helps to increase robustness to noise.

For optimization methods comparison on this problem, consider $n = 10^6$, $N = 10$, $\alpha = 10^{-6}$ the following procedure:

1. sample observations $\{y_i, \boldsymbol{\varphi}_i\}_1^N$ and initial point $\mathbf{x}_0$ in accordance to the following distributions:

$$\boldsymbol{\varphi}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \varepsilon_i \sim \mathcal{N}(\mathbf{0}, 10^{-2}), \quad \mathbf{x}_0 \sim \mathcal{N}(\mathbf{x}_\star, \mathbf{I});$$

2. for each algorithm $\mathcal{A}$ run the optimization process, and take the minimum function value for the first 100 steps $\min\limits_{0\leqslant t\leqslant 100} f(\mathbf{x}_t) =: f^{\mathcal{A}}_{min}$;
3. consider statistics $\Upsilon(\mathcal{A})$:

$$\Upsilon(\mathcal{A}) = |\{\mathcal{A}' : \ f^{\mathcal{A}'}_{min} < f^{\mathcal{A}}_{min}\}|; \tag{3.7}$$

4. repeat steps 1-3 a thousand times, thus obtaining a sample of $\Upsilon(\mathcal{A})$ values for each algorithm $\mathcal{A}$.

Let us explain the choice of $\Upsilon(\mathcal{A})$ as a metric for evaluating algorithms quality. Due to the fact that each repetition of steps (1) and (2) uses different functions f, absolute values of the function minimum are uninformative for comparison. Therefore, it is more

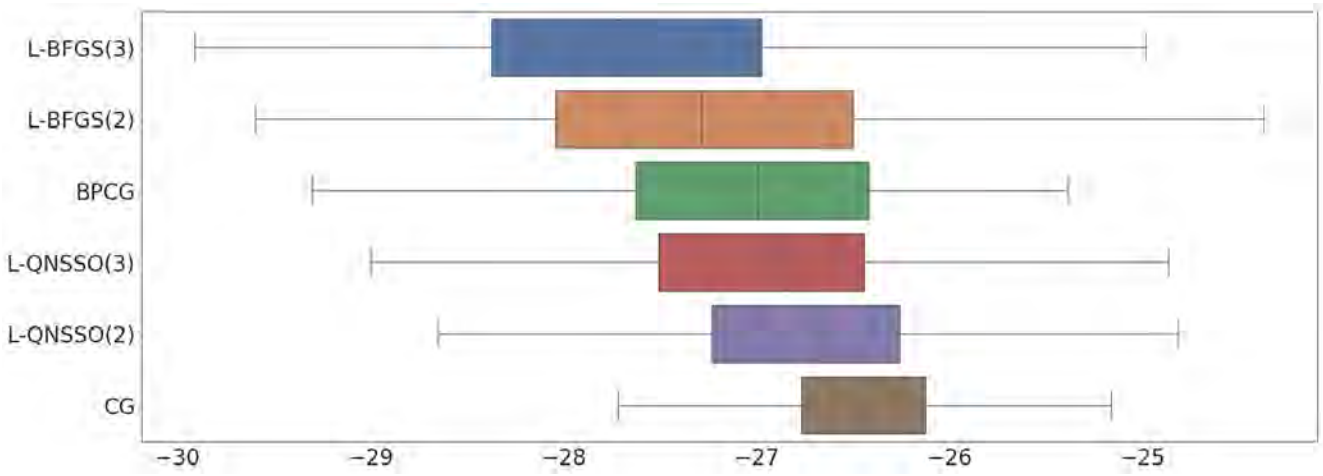


Figure 3.7 – Boxplots of $\min\limits_{0 \leqslant t \leqslant n} f(\mathbf{x}_t)$ for the Rosenbrock function (3.3). The value of the function in the logarithmic scale is marked along the abscissa axis. Left border, segment inside and right border of the box are 25%, 50% and 75% percentiles respectively, the whiskers boundaries correspond to the minimum and maximum values

convenient to consider the relative quality of algorithms — a number of algorithms whose result exceeds the result of $\mathcal{A}$, which is $\Upsilon(\mathcal{A})$.

Due to the high dimensionality of the problem, only the following methods are used for comparison: CG, BPCG, L-BFGS(2), L-BFGS(3), L-QNSSO(2), L-QNSSO(3). Result of the experiment — a sum of the values of the statistics $\Upsilon(\mathcal{A})$ for each of the algorithms — are presented in the Table 1. As one can see, L-QNSSO algorithms are significantly ahead of all competitors. In particular, the L-QNSSO(2) algorithm delivers the minimum value of the function in more than half of the experiments. It is also worth noting that in this problem, the dimension of the subspace plays a negative role for both the L-QNSSO method and the L-BFGS method.

3.2.4 Logistic regression for chemical compounds classification

Consider the DOROTHEA [114] dataset, which contains information about 1950 drugs (chemical compounds) divided into three sets: training (800 observations), validation (350 observations), and test (800 observations). Each drug is represented by a *feature vector* — a binary vector of dimension 10^5 (half of features were randomly

Table 1 – Sum of $\Upsilon(\mathcal{A})$ (3.7) values per 1000 experiments.

Algorithm	Sum of $\Upsilon(\mathcal{A})$
L-QNSSO(2)	425
L-QNSSO(3)	1498
CG	1896
L-BFGS(2)	3520
L-BFGS(3)	3808
BPCG	3853

generated to increase the problem complexity) and an activity flag characterizing the drug's binding to a thrombin. In this case the classification task is the following: on the basis of the training set to obtain the best accuracy of the prediction of drug activity on the validation and test samples.

Logistic regression — one of the methods of solving the binary classification problem — is a statistical model of probability prediction, often used for classification problems [115]. Basically, it is an application of a logistic function to a linear model: i.e. *decision function* has the following form::

$$h_i(\mathbf{x}) = h(\boldsymbol{\varphi}_i, \mathbf{x}) = \frac{1}{1 + \exp\left(-\boldsymbol{\varphi}_i^\top \mathbf{x}\right)},$$

where φ_i is the feature vector of the i–th object and $\mathbf{x}$ is the parameter vector of the model. Note that feature vector and the model parameter vector are of the same dimensionality. One of the most common ways to estimate logistic regression parameters is to maximize the likelihood function, a special case of the empirical risk functional minimization problem (1.1):

$$L(\mathbf{x}) = \sqrt[\frac{1}{N}]{\prod_{i=1}^{N} h_i(\mathbf{x})^{y_i}(1 - h_i(\mathbf{x}))^{(1-y_i)}} \to \max_{\mathbf{x}},$$

which is equivalent to minimizing its negative logarithm:

$$\begin{aligned} f(\mathbf{x}) &= -\log L(\mathbf{x}) \\ &= -\frac{1}{N}\sum_{i=1}^{N} \left[y_i \log h_i(\mathbf{x}) + (1 - y_i)\log(1 - h_i(\mathbf{x}))\right] \to \min_{\mathbf{x}}. \end{aligned} \tag{3.8}$$

Function (3.8) gradient by $\mathbf{x}$ has the following form

$$\nabla f(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}(h_i(\mathbf{x}) - y_i)\boldsymbol{\varphi}_i. \tag{3.9}$$

Thus, in the context of optimization, the primary goal is to minimize the value of the function (3.8) on the training set. Maximizing classification quality on a test dataset is a pattern recognition task and will not be considered further. Consider the following approach for optimization algorithms comparison:

1. sample an initial point in accordance to distribution $\mathbf{x}_0 \sim \mathcal{N}\left(\mathbf{0}, \sigma_x^2\mathbf{I}\right)$, $\sigma_x^2 = 10^{-6}$;
2. for each algorithm, run the optimization process of the function (3.8) on the training set, take the minimum value of the function for the first 100 iterations: $\min\limits_{0\leqslant t\leqslant 100} f(\mathbf{x}_t)$;
3. repeat steps 1-2 a thousand times, thus obtaining a sample of $\min\limits_{0\leqslant t\leqslant 100} f(\mathbf{x}_t)$ values for each algorithm.

Distribution of the obtained samples are presented via boxplots at Figure 3.8. It is easy to notice that the L-QNSSO and CG methods are significantly superior to the L-BFGS and BPCG methods. The best quality is achieved by L-QNSSO(5) algorithm in the sense of 50% and 25% percentiles, and in the sense of the minimum value. It is noteworthy that the relationship between the history size and the quality of the L-QNSSO method on this problem is not monotonic. Finally, in the context of the pattern recognition problem, it is common to restart the optimization process with different initial data in order to obtain the best result. In this way, the optimistic characteristics of the distribution of the obtained samples are relevant: 25% percentile and minimum value. In accordance to these metrics, L-QNSSO algorithms are superior to the alternatives

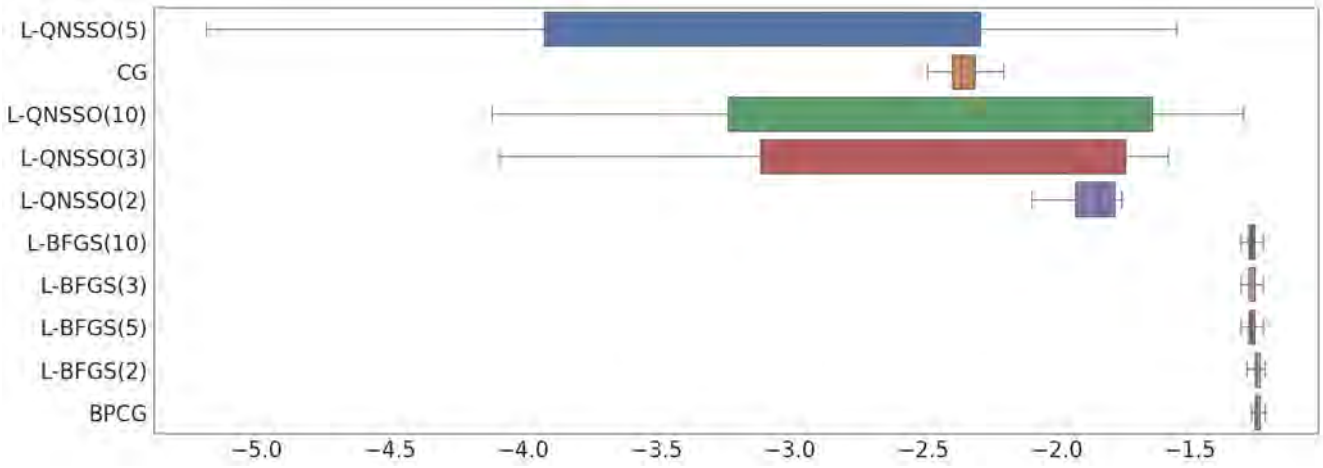


Figure 3.8 – Boxplots of $\min_{0 \leqslant t \leqslant 100} f(\mathbf{x}_t)$ for the logistic regression loss function function (3.8). The value of the function in the logarithmic scale is marked along the abscissa axis. Left border, segment inside and right border of the box are 25%, 50% and 75% percentiles respectively, the whiskers boundaries correspond to the minimum and maximum values

Conclusion

In conclusion we list the main scientific results of the work.

1. The characteristic of the sequential subspace optimization methods with quadratic surrogates convergence rate for the quadratic strictly convex case through the projection error and the approximation error is obtained (Section 2.1.2, Lemma 1) [57];
2. The criteria of sublinear, linear and superlinear convergence rates of sequential subspace methods with a quadratic surrogate are established for the case of a strictly convex objective function (Section 2.1.3, Theorem 3);
3. The corrective sequential subspace optimization method with linear convergence rate is developed (Section 2.3.1, Theorem 2), a sequential subspace optimization method with quasi-Newton is developed, that converges in a finite number of steps with linear convergence rate (Section 2.3.2, Remark 13) for a strictly convex quadratic function [57];
4. A modified sign–perturbed sums method is developed for constructing an exact confidence region of the linear model parameter for the case of noises independent with each other and with the model inputs, and otherwise arbitrary additive noises in observations (Section 2.4, Theorem 4), for the one dimensional case an analytical expression of the confidence interval boundaries and their consistency conditions are obtained (Section 2.5, Lemma 7 and Lemma 8) [57].

Bibliography

1. *Ljung*, *L.* System identification (2nd ed.): theory for the user / L. Ljung. – Upper Saddle River, NJ, USA : Prentice Hall PTR, 1999.
2. *Boyd*, *S.* Convex Optimization / S. Boyd, L. Vandenberghe. – Cambridge university press, 2004.
3. *Граничин*, *О. Н.* Рандомизированные алгоритмы оценивания и оптимизации при почти произвольных помехах / О. Н. Граничин, Б. Т. Поляк. – М.: Наука, 2003.
4. *Вапник*, *В. Н.* Теория распознавания образов: статистические проблемы обучения / В. Н. Вапник, А. Я. Червоненкис. – Наука. Гл. ред. физ.-мат. лит., 1974.
5. *Bishop*, *C. M.* Pattern Recognition and Machine Learning / C. M. Bishop. – Springer, 2006.
6. *Vapnik*, *V.* The Nature of Statistical Learning Theory / V. Vapnik. – Springer science & business media, 2013.
7. *Goodfellow*, *I.* Deep Learning / I. Goodfellow, Y. Bengio, A. Courville. – MIT press, 2016.
8. *Hestenes*, *M. R.* Methods of conjugate gradients for solving linear systems / M. R. Hestenes, E. Stiefel // Journal of Research of the National Bureau of Standards. – 1952. – Vol. 49, no. 6. – P. 409–436.
9. *Fletcher*, *R.* Function minimization by conjugate gradients / R. Fletcher, C. M. Reeves // The Computer Journal. – 1964. – Vol. 7, no. 2. – P. 149–154.
10. *Polak*, *E.* Note sur la convergence de méthodes de directions conjuguées / E. Polak, G. Ribiere // ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique. – 1969. – Vol. 3, R1. – P. 35–43.
11. *Polyak*, *B. T.* The conjugate gradient method in extremal problems / B. T. Polyak // USSR Computational Mathematics and Mathematical Physics. – 1969. – Vol. 9, no. 4. – P. 94–112.

12. *Golub*, *G. H.* Some history of the conjugate gradient and Lanczos algorithms: 1948–1976 / G. H. Golub, D. P. O'Leary // SIAM Review. – 1989. – Vol. 31, no. 1. – P. 50–102.

13. *Cohen*, *A. I.* Rate of convergence of several conjugate gradient algorithms / A. I. Cohen // SIAM Journal on Numerical Analysis. – 1972. – Vol. 9, no. 2. – P. 248–259.

14. *Davidon*, *W. C.* Variable metric method for minimization / W. C. Davidon // SIAM Journal on Optimization. – 1991. – Vol. 1, no. 1. – P. 1–17.

15. *Broyden*, *C. G.* The convergence of a class of double-rank minimization algorithms 1. general considerations / C. G. Broyden // IMA Journal of Applied Mathematics. – 1970. – Vol. 6, no. 1. – P. 76–90.

16. *Davidon*, *W. C.* New least-square algorithms / W. C. Davidon // Journal of Optimization Theory and Applications. – 1976. – Vol. 18, no. 2. – P. 187–197.

17. *Nocedal*, *J.* Updating quasi-Newton matrices with limited storage / J. Nocedal // Mathematics of Computation. – 1980. – Vol. 35, no. 151. – P. 773–782.

18. *Fletcher*, *R.* Practical Methods of Optimization (2nd ed.) / R. Fletcher. – New Your: John Wiley, 1987.

19. *Conn*, *A. R.* Convergence of quasi-Newton matrices generated by the symmetric rank one update / A. R. Conn, N. I. Gould, P. L. Toint // Mathematical Programming. – 1991. – Vol. 50, no. 1–3. – P. 177–195.

20. *Miele*, *A.* Study on a memory gradient method for the minimization of functions / A. Miele, J. Cantrell // Journal of Optimization Theory and Applications. – 1969. – Vol. 3, no. 6. – P. 459–470.

21. *Cragg*, *E.* Study on a supermemory gradient method for the minimization of functions / E. Cragg, A. Levy // Journal of Optimization Theory and Applications. – 1969. – Vol. 4, no. 3. – P. 191–205.

22. *Fletcher*, *R.* A limited memory steepest descent method / R. Fletcher // Mathematical Programming. – 2012. – Vol. 135, no. 1/2. – P. 413–436.

23. On iterated-subspace minimization methods for nonlinear optimization / A. R. Conn [et al.] // Linear and Nonlinear Conjugate-Gradient Related Methods. – SIAM, 1994. – P. 50–78.

24. *Narkiss*, *G.* Sequential Subspace Optimization Method for Large-Scale Unconstrained Optimization : tech. rep. / G. Narkiss, M. Zibulevsky ; Technion-IIT, Department of Electrical Engineering. – 2005. – P. 31.

25. *Yuan*, *Y.-X.* A Review on Subspace Methods for Nonlinear Optimization : tech. rep. / Y.-X. Yuan. – 2014.

26. *Chouzenoux*, *E.* A majorize–minimize strategy for subspace optimization applied to image restoration / E. Chouzenoux, J. Idier, S. Moussaoui // IEEE Transactions on Image Processing. – 2010. – Vol. 20, no. 6. – P. 1517–1528.

27. *Narkiss*, *G.* Support Vector Machine via Sequential Subspace Optimization / G. Narkiss, M. Zibulevsky. – Technion-IIT, Department of Electrical Engineering, 2005.

28. *Zheng*, *Y.* Efficient variational Bayesian approximation method based on subspace optimization / Y. Zheng, A. Fraysse, T. Rodet // IEEE Transactions on Image Processing. – 2014. – Vol. 24, no. 2. – P. 681–693.

29. SEBOOST-Boosting stochastic learning using subspace optimization techniques / E. Richardson [et al.] // Advances in Neural Information Processing Systems. – 2016. – P. 1534–1542.

30. *Zibulevsky*, *M.* L1-L2 optimization in signal and image processing / M. Zibulevsky, M. Elad // IEEE Signal Processing Magazine. – 2010. – Vol. 27, no. 3. – P. 76–88.

31. *Граничин*, *О. Н.* Оценивание параметров линейной регрессии при произвольных помехах / О. Н. Граничин // Автоматика и телемеханика. – 2002. – No. 1. – P. 30–41.

32. *Granichin*, *O.* Linear regression and filtering under nonstandard assumptions (Arbitrary noise) / O. Granichin // IEEE Transactions on Automatic Control. – 2004. – Vol. 49, no. 10. – P. 1830–1837.

33. *Goldenshluger*, *A. V.* Estimation of regression parameters with arbitrary noise / A. V. Goldenshluger, B. T. Polyak // Mathematical Methods of Statistics. – 1993. – Vol. 2, no. 1. – P. 18–29.

34. *Граничин*, *О. Н.* Алгоритм стохастической аппроксимации с возмущением на входе для идентификации статического нестационарного дискретного объекта / О. Н. Граничин // Вестник Санкт-Петербургского университета. Серия 1. Математика. Механика. Астрономия. – 1988. – No. 3. – P. 92–93.

35. *Граничин*, *О. Н.* Адаптивное управление с использованием пробных сигналов в канале обратной связи / О. Н. Граничин, В. Н. Фомин // Автоматика и телемеханика. – 1986. – No. 2. – P. 100–112.

36. *Граничин*, *О. Н.* Рандомизированные алгоритмы оценивания и оптимизации при почти произвольных помехах / О. Н. Граничин, Б. Т. Поляк. – Москва: Наука, 2003.

37. Two procedures with randomized controls for the parameters' confidence region of linear plant under external arbitrary noise / K. Amelin [et al.] // Proc. of the IEEE Int. Symposium on Intelligent Control (ISIC). – IEEE. 2012. – P. 1226–1231.

38. Combined procedure with randomized controls for the parameters' confidence region of linear plant under external arbitrary noise / K. Amelin [et al.] // Proc. of the 51st Conference on Decision and Control (CDC2012). – IEEE. 2013. – P. 2134–2139.

39. *Amelin*, *K.* Randomized controls for linear plants and confidence regions for parameters under external arbitrary noise / K. Amelin, O. Granichin // Proc. of the American Control Conference (ACC). – IEEE. 2012. – P. 0743–1619.

40. *Campi*, *M. C.* Guaranteed non-asymptotic confidence regions in system identification / M. C. Campi, E. Weyer // Automatica. – 2005. – Vol. 41, no. 10. – P. 1751–1764.

41. *Csaji*, *B. C.* Non-asymptotic confidence regions for the least-squares estimate / B. C. Csaji, M. C. Campi, E. Weyer // Proceedings of the 16th IFAC Symposium on System Identification (SYSID 2012). – 2012. – July. – P. 227–232.

42. *Granichin*, *O. N.* The nonasymptotic confidence set for parameters of a linear control object under an arbitrary external disturbance / O. N. Granichin // Automation and Remote Control. – 2012. – Vol. 73, no. 1. – P. 20–30.

43. Патент: Программа для оптического распознавания визуальной текстовой информации на арабском языке (ОРТА-2Б.ГРС) / О. А. Берникова [et al.]. – 2013.

44. Методы оптического распознавания текста на арабском языке / О. А. Берникова [et al.] // Стохастическая оптимизация в информатике. – 2013. – Vol. 9, no. 2. – P. 3–20.

45. *Сенов*, *А. А.* Доверительные множества при почти произвольных помехах в контексте линейных моделей рекомендательных систем / А. А. Сенов // Стохастическая оптимизация в информатике. – 2013. – Vol. 9, no. 1. – P. 68–86.

46. Exact confidence regions for linear regression parameter under external arbitrary noise / A. Senov [et al.] // American Control Conference (ACC), 2014. – IEEE. 2014. – P. 5097–5102.

47. *Сенов*, *А. А.* Идентификация параметров линейной регрессии при произвольных внешних помехах в наблюдениях / А. А. Сенов, О. Н. Граничин // XII всероссийское совещание по проблемам управления ВСПУ-2014. – 2014. – P. 2708–2719.

48. *Senov*, *A.* Improving distributed stochastic gradient descent estimate via loss function approximation / A. Senov // IFAC-PapersOnLine. – 2015. – Vol. 48, no. 25. – P. 292–297.

49. *Сенов*, *А. А.* Квадратичная проективная регрессия как метод обучения в разреженных пространствах высокой размерности / А. А. Сенов // Эвристические Алгоритмы и Распределенные Вычисления. – 2015. – Vol. 2, no. 4. – P. 73–92.

50. *Сенов*, *А. А.* Улучшение оценки распределенного стохастического градиентного спуска через аппроксимацию функции потерь / А. А. Сенов // Стохастическая оптимизация в информатике. – 2015. – Vol. 11, no. 1. – P. 103–126.

51. *Boiarov*, *A.* Arabic manuscript author verification using deep convolutional networks / A. Boiarov, A. Senov, A. Knysh // 2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR). – 2017. – P. 1–5.

52. *Senov*, *A.* Accelerating gradient descent with projective response surface methodology / A. Senov // International Conference on Learning and Intelligent Optimization. – Springer. 2017. – P. 376–382.

53. *Senov*, *A.* Projective approximation based gradient descent modification / A. Senov, O. Granichin // IFAC-PapersOnLine. – 2017. – Vol. 50, no. 1. – P. 3899–3904.

54. *Senov*, *A.* Projective approximation based quasi-Newton methods / A. Senov // International Workshop on Machine Learning, Optimization, and Big Data. – Springer. 2017. – P. 29–40.

55. *Сенов*, *А. А.* Глубокое обучение в задаче реконструкции суперразрешения изображений / А. А. Сенов // Стохастическая оптимизация в информатике. – 2017. – Vol. 13, no. 2. – P. 38–57.

56. *Сенов*, *А. А.* О методах последовательной подпространственной оптимизации / А. А. Сенов // Стохастическая оптимизация в информатике. – 2018. – Vol. 14, no. 2. – P. 40–61.

57. *Сенов*, *А. А.* Квазиньютоновские методы последовательной подпространственно оптимизации с квадратичным суррогатом ми- нимизации строго выпуклых функций / А. А. Сенов // Стохастическая оптимизация в информатике. – 2019. – Vol. 15, no. 1. – P. 20–68.

58. *Montgomery*, *D. C.* Introduction to Linear Regression Analysis / D. C. Montgomery, E. A. Peck, G. G. Vining. – Wiley-Interscience, 2007.

59. *Фомин*, *В. Н.* Рекуррентное оценивание и адаптивная фильтрация / В. Н. Фомин. – Москва: Наука, 1984.

60. *Hastie*, *T.* The Elements of Statistical Learning: Data Mining, Inference, and Prediction / T. Hastie, R. Tibshirani, J. H. Friedman. – New York, NY: Springer, 2009.

61. *Vapnik*, *V. N.* The nature of statistical learning theory / V. N. Vapnik. – Springer–Verlag, 1995.

62. *Нестеров*, *Ю. Е.* Введение в выпуклую оптимизацию / Ю. Е. Нестеров. – МЦНМО, 2010.

63. *Granichin*, *O. N.* Stochastic approximation search algorithms with randomization at the input / O. N. Granichin // Automation and Remote Control. – 2015. – Vol. 76, no. 5. – P. 762–775.

64. *Nesterov*, *Y.* Introductory Lectures on Convex Programming Volume I: Basic Course / Y. Nesterov. – Citeseer, 1998.

65. *Поляк*, *Б. Т.* Введение в оптимизацию / Б. Т. Поляк. – Наука. Гл. ред. физ.-мат. лит., 1983.

66. *Ypma*, *T. J.* Historical development of the Newton–Raphson method / T. J. Ypma // SIAM review. – 1995. – Vol. 37, no. 4. – P. 531–551.

67. *Polyak*, *B. T.* Newton's method and its use in optimization / B. T. Polyak // European Journal of Operational Research. – 2007. – Vol. 181, no. 3. – P. 1086–1096.

68. *Cauchy*, *A.-L.* Méthode générale pour la résolution des systemes d'équations simultanées / A.-L. Cauchy // Comp. Rend. Sci. Paris. – 1847. – Vol. 25, no. 1847. – P. 536–538.

69. *Crockett*, *J. B.* Gradient methods of maximization / J. B. Crockett, H. Chernoff // Pacific Journal of Mathematics. – 1955. – Vol. 5, no. 1. – P. 33–50.

70. *Nocedal*, *J.* Numerical Optimization / J. Nocedal, S. J. Wright. – Springer, 2006.

71. *Barzilai*, *J.* Two-point step size gradient methods / J. Barzilai, J. M. Borwein // IMA Journal of Numerical Analysis. – 1988. – Vol. 8, no. 1. – P. 141–148.

72. *Dai*, *Y.-H.* A new analysis on the Barzilai-Borwein gradient method / Y.-H. Dai // Journal of the Operations Research Society of China. – 2013. – Vol. 1, no. 2. – P. 187–198.

73. *Wei*, *Z.* New quasi-Newton methods for unconstrained optimization problems / Z. Wei, G. Li, L. Qi // Applied Mathematics and Computation. – 2006. – Vol. 175, no. 2. – P. 1156–1188.

74. *Zhang*, *J. Z.* New quasi-Newton equation and related methods for unconstrained optimization / J. Z. Zhang, N. Y. Deng, L. H. Chen // Journal of Optimization Theory and Applications. – 1999. – Vol. 102, no. 1. – P. 147–167.

75. *Don*, *F. J. H.* On the symmetric solutions of a linear matrix equation / F. J. H. Don // Linear Algebra and its Applications. – 1987. – Vol. 93. – P. 1–7.

76. *Dai*, *Y.-H.* Nonlinear conjugate gradient methods / Y.-H. Dai // Wiley Encyclopedia of Operations Research and Management Science. – 2010.

77. *Powell*, *M. J. D.* Restart procedures for the conjugate gradient method / M. J. D. Powell // Mathematical programming. – 1977. – Vol. 12, no. 1. – P. 241–254.

78. *Dai*, *Y.* Convergence properties of Beale-Powell restart algorithm / Y. Dai, Y. Yuan // Science in China Series A: Mathematics. – 1998. – Vol. 41, no. 11. – P. 1142–1150.

79. *Nemirovski*, *A.* Orth-method for smooth convex optimization / A. Nemirovski // Izvestia AN SSSR, Transl.: Eng. Cybern. Soviet J. Comput. Syst. Sci. – 1982. – Vol. 2. – P. 937–947.

80. *Shewchuk*, *J. R.* An introduction to the conjugate gradient method without the agonizing pain / J. R. Shewchuk. – 1994. – Carnegie-Mellon University. Department of Computer Science.

81. *Немировский*, *А. С.* Методы оптимизации, адаптивные к «существенной» размерности задачи / А. С. Немировский, Д. Б. Юдин // Автоматика и телемеханика. – 1977. – No. 4. – P. 75–87.

82. *Powell*, *M. J.* A new algorithm for unconstrained optimization / M. J. Powell // Nonlinear Programming. – Elsevier, 1970. – P. 31–65.

83. *Wang*, *Z.-H.* A subspace implementation of quasi-Newton trust region methods for unconstrained optimization / Z.-H. Wang, Y.-X. Yuan // Numerische Mathematik. – 2006. – Vol. 104, no. 2. – P. 241–269.

84. *Yuan*, *Y.-X.* A subspace study on conjugate gradient algorithms / Y.-X. Yuan, J. Stoer // ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik. – 1995. – Vol. 75, no. 1. – P. 69–77.

85. *Wolfe*, *P.* Convergence conditions for ascent methods / P. Wolfe // SIAM Review. – 1969. – Vol. 11, no. 2. – P. 226–235.

86. *Wolfe*, *P.* Convergence conditions for ascent methods. II: Some corrections / P. Wolfe // SIAM Review. – 1971. – Vol. 13, no. 2. – P. 185–188.

87. *Yuan*, *Y.-X.* Subspace techniques for nonlinear optimization / Y.-X. Yuan // Some Topics in Industrial and Applied Mathematics. – World Scientific, 2007. – P. 206–218.

88. *Yuan*, *Y.-X.* Subspace methods for large scale nonlinear equations and nonlinear least squares / Y.-X. Yuan // Optimization and Engineering. – 2009. – Vol. 10, no. 2. – P. 207–218.

89. *LeCun*, *Y.* The MNIST Dataset Of Handwritten Digits / Y. LeCun, C. Cortes, C. J. Burges. – 1999. – URL: http://yann.lecun.com/exdb/mnist.

90. *Krizhevsky*, *A.* Learning multiple layers of features from tiny images : tech. rep. / A. Krizhevsky, G. Hinton ; Citeseer. – 2009.

91. *Hartley*, *H. O.* Exact confidence regions for the parameters in non-linear regression laws / H. O. Hartley // Biometrika. – 1964. – Vol. 51, no. 3/4. – P. 347–353.

92. *Draper*, *N. R.* Applied Regression Analysis. Vol. 326 / N. R. Draper, H. Smith. – John Wiley & Sons, 1998.

93. *Davison*, *A. C.* Bootstrap Methods and Their Application / A. C. Davison, D. V. Hinkley. – Cambridge university press, 1997.

94. Bootstrapping regression models / D. A. Freedman [et al.] // The Annals of Statistics. – 1981. – Vol. 9, no. 6. – P. 1218–1228.

95. *Fox*, *J.* Bootstrapping regression models / J. Fox // An R and S-PLUS Companion to Applied Regression: A Web Appendix to the Book. Sage, Thousand Oaks, CA. URL: http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-bootstrapping.pdf. – 2002.

96. *Bai*, *E.-W.* Membership set estimators: size, optimal inputs, complexity and relations with least squares / E.-W. Bai, R. Tempo, H. Cho // IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications. – 1995. – Vol. 42, no. 5. – P. 266–277.

97. *Vicino*, *A.* Sequential approximation of feasible parameter sets for identification with set membership uncertainty / A. Vicino, G. Zappa // IEEE Transactions on Automatic Control. – 1996. – Vol. 41, no. 6. – P. 774–785.

98. *Dalai*, *M.* Parameter identification for nonlinear systems: guaranteed confidence regions through LSCR / M. Dalai, E. Weyer, M. C. Campi // Automatica. – 2007. – Vol. 43, no. 8. – P. 1418–1425.

99. *Csáji*, *B. C.* Sign-Perturbed Sums: A new system identification approach for constructing exact non-asymptotic confidence regions in linear regression models / B. C. Csáji, M. C. Campi, E. Weyer // IEEE Transactions on Signal Processing. – 2014. – Vol. 63, no. 1. – P. 169–181.

100. *Kieffer*, *M.* Guaranteed characterization of exact non-asymptotic confidence regions as defined by LSCR and SPS / M. Kieffer, E. Walter // Automatica. – 2014. – Vol. 50, no. 2. – P. 507–512.

101. *Волкова*, *М. В.* Рандомизированные алгоритмы оценивания параметров инкубационных процессов в условиях неопределённостей и конечного числа наблюдений: дис. …канд. физ.-мат. наук: 01.01.09 / М. В. Волкова. – 2018. – Санкт-Петербургский Государственный Университет, СПб.

102. *Akaike, H.* On a successive transformation of probability distribution and its application to the analysis of the optimum gradient method / H. Akaike // Annals of the Institute of Statistical Mathematics. – 1959. – Vol. 11, no. 1. – P. 1–16.

103. *Forsythe, G. E.* On the asymptotic directions of thes-dimensional optimum gradient method / G. E. Forsythe // Numerische Mathematik. – 1968. – Vol. 11, no. 1. – P. 57–76.

104. *Фаддеев, Д. К.* Вычислительные методы линейной алгебры. Vol. 1 / Д. К. Фаддеев, В. Н. Фаддеева. – Физматгиз Москва, 1960.

105. *O'Leary, D. P.* Estimating the largest eigenvalue of a positive definite matrix / D. P. O'Leary, G. Stewart, J. S. Vandergraft // Mathematics of Computation. – 1979. – Vol. 33, no. 148. – P. 1289–1292.

106. *Парлетт, Б. Н.* Симметричная проблема собственных значений: Численные методы / Б. Н. Парлетт, Х. Д. Икрамов, Ю. А. Кузнецов. – Мир, 1983.

107. *Golub, G. H.* Matrix Computations / G. H. Golub, C. F. Van Loan. – 4th. – JHU press, 2012.

108. *Brent, R. P.* An algorithm with guaranteed convergence for finding a zero of a function / R. P. Brent // The Computer Journal. – 1971. – Vol. 14, no. 4. – P. 422–425.

109. *Van Der Walt, S.* The NumPy array: a structure for efficient numerical computation / S. Van Der Walt, S. C. Colbert, G. Varoquaux // Computing in Science & Engineering. – 2011. – Vol. 13, no. 2. – P. 22.

110. SciPy: Open source scientific tools for Python / E. Jones, T. Oliphant, P. Peterson, [et al.]. – 2001–. – URL: http://www.scipy.org/ ; Viewed: 2019-09-01.

111. *Rosenbrock, H. H.* An automatic method for finding the greatest or least value of a function / H. H. Rosenbrock // The Computer Journal. – 1960. – Vol. 3, no. 3. – P. 175–184.

112. *Shang, Y.-W.* A note on the extended Rosenbrock function / Y.-W. Shang, Y.-H. Qiu // Evolutionary Computation. – 2006. – Vol. 14, no. 1. – P. 119–126.

113. *Quapp, W.* Searching minima of an n-dimensional surface: A robust valley following method / W. Quapp // Computers & Mathematics with Applications. – 2001. – Vol. 41, no. 3/4. – P. 407–414.

114. Result analysis of the NIPS 2003 feature selection challenge / I. Guyon [et al.] // Advances in Neural Information Processing Systems. – 2005. – P. 545–552.

115. *Cramer*, *J. S.* The early origins of the logit model / J. S. Cramer // Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences. – 2004. – Vol. 35, no. 4. – P. 613–626.