МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ

ОЦЕНКИ СХОДИМОСТИ ДЛЯ МЕТОДА ФРАНКО-ВУЛЬФА С НЕТОЧНЫМ ГРАДИЕНТОМ

проект в рамках курса математического практикума
4 семестр

Проект выполняли студенты 2 курса МФТИ

Ширяев Дмитрий Ядров Платон Никитин Иван

Научный руководитель: Фёдор Стонякин

Долгопрудный

2024

Оглавление

Введение

В этой статье мы сфокусируемся на задаче $\min_{x \in P} f(x)$,
где P - некоторая выпуклая область (в дальнейшем будем считать её компактной), а f - целевая функция, удовлетворяющая некоторому свойству регулярности, например, гладкости и выпуклости.
Также необходимо указать, какого типа операции доступны методу. Обычно используется оракул первого порядка, позволяющий вычислять градиенты функции в текущей точке, а также значение функции. В этой статье мы будем рассматривать неточный оракул вычисления градиента в точке.

В дальнейшем $\|\cdot\|$ мы обозначаем l_2 норму, $\mathbf{x}^*$ обозначаем одно из оптимальных решений для $\min_{\mathbf{x} \in \mathcal{P}} f(\mathbf{x})$ и определим $f^* = f(\mathbf{x}^*)$.
Для компактного выпуклого множества P диаметр $D = max_{x,y \in P} \|x - y\|$.
Также предполагаем, что f является дифференцируемой. Более того, если не указано иное, мы рассматриваем функцию f: $P \to \mathbb{R}$.
Для дальнейшего анализа нам необходимы будут следующие понятия:

Опр.1 *Гладкость.* Пусть $f : P \to \mathbb{R}$ – дифференцируемая функция.
Тогда f L-гладкая, если: $f(y) - f(x) \leq \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2$, $\forall x, y \in P$

Опр.2 *Выпуклость.* f назвается выпуклой, если:
$f(y) - f(x) \geq \langle \nabla f(x), y - x \rangle$, $\forall x, y \in P$

Глава 1

Почему метод Франк-Вульфа?

Для гарантии того, что траектория будет лежать в выполнимой области P существует несколько подходов:

1. *Оператор проекции.* Доступ к оператору проекции $\Pi_{\mathcal{P}}$ области P, который для данной точки $x \in \mathbb{R}^n$ возвращает $\Pi_{\mathcal{P}}(x) = \arg\min_{y \in \mathcal{P}} \|x - y\|$.

2. *Барьерная функция.* Доступ к барьерной функции выполнимой области P, значение которой возрастает до бесконечности при приближении к границе P.

3. *Линейная минимизация.* Доступ к оракулу линейной минимизации (LMO), который получая на вход линейную целевую функцию $c \in \mathbb{R}^n$ возвращает $y \in \operatorname{argmin}_{\mathbf{x} \in \mathcal{P}} \langle c, x \rangle$.

В методе Франк-Вульфа используется третий подход - LMO.
Традиционно задачи оптимизации решаются с использованием вариантов методов на основе проекций. Например, для некоторой допустимой области P с проектором $\Pi_{\mathcal{P}}(x)$ и гладкой целевой функцией f метод проекцией (PGD) обычно имеет следующий вид:
$x_{t+\frac{1}{2}} \leftarrow x_t - \gamma_t \nabla f(x_t)$
$x_{t+1} \leftarrow \Pi_{\mathcal{P}}(x_{t+\frac{1}{2}})$
где γ_t - некоторый размер шага (например, $\gamma_t = \frac{1}{L}$, если f является L-гладкой. Происходит спуск без учёта ограничений, после чего происходит проекция обратно в допустимую область (рис. 1.1). Оптимальные методы и оценки скорости сходимости известны для большинства сценариев. Однако, когда допустимая область усложняется, операция проекции может стать ограничивающим фактором. Часто требуется решение вспомогательной оптимизационной задачи — известной как задача проекции — над той же допустимой областью для каждого шага спуска. Сложность этой задачи не позволяет использовать методы на основе проекций для многих значимых задач с ограничениями.
Методы внутренней точки, хотя и обладают привлекательными теоретическими гарантиями, обычно требуют задания барьерной функции, описывающей допу-

стимую область. Во многих критических сценариях краткое описание допустимой области либо неизвестно, либо доказано ее отсутствие. Например, политоп, соответствующий совпадению (matching polytope), не допускает подходящих линейных программ, ни точных (Rothvoss, 2017), ни приблизительных (Braun and Pokutta, 2015a;b; Sinha, 2018). Кроме того, достижение достаточной точности при обновлениях шагов метода внутренней точки часто требует информации второго порядка, что иногда может ограничивать его применимость.

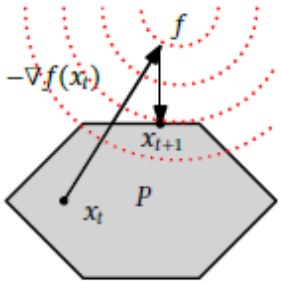

Рис. 1.1: Метод с проекции.[1]

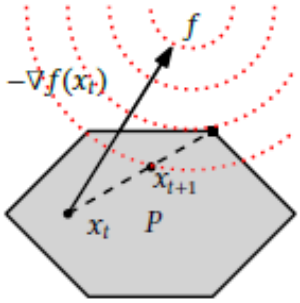

Рис. 1.2: Методы без проекций.[1]

Таким образом, при близком рассмотрении двух упомянутых ранее методов становится ясно, что оба существенно преобразуют ограниченную задачу (Opt) в неограниченную. Затем они либо корректируют обновления, нарушающие ограничения (как в методе градиентного спуска с проекцией), либо штрафуют приближение к нарушениям ограничений (как в методах внутренней точки). Тем не менее, существует еще одна категория техник, называемая методами без проекций, которая направлена непосредственно на решение задачи ограниченной оптимизации. В отличие от своих аналогов, эти методы обходят необходимость дорогостоящих проекций или стратегий штрафов и поддерживают выполнение ограничений на протяжении всего процесса. Самыми известными вариантами в этой категории являются методы Франк-Вульфа (Frank-Wolfe, FW) — вернемся к работам Франк и Вульфа (1956) — которые будут основной темой данной статьи и также известны как методы условного градиента (conditional gradient, CG) (Levitin and Polyak, 1966).

Исторически методы, подобные алгоритму Франк-Вульфа, привлекали ограниченное внимание из-за определенных недостатков, особенно из-за неоптимальных скоростей сходимости. Однако в начале 2013 года произошло заметное возрождение интереса. Этот рост интереса в значительной степени объясняется изменением требований и другими свойствами, которые теперь стали актуальными. В частности, эти методы отлично подходят для работы с сложными ограничениями и обладают низкой сложностью итераций. Это делает их очень эффективными в

контексте задач машинного обучения масштаба больших данных.

Вместо использования потенциально дорогих операций проекции, методы Франк-Вульфа используют так называемый "линейный минимизационный оракул" (Linear Minimization Oracle, LMO). Этот подпрограммный модуль заключается только в оптимизации линейной функции на допустимой области и часто оказывается более эффективным с точки зрения затрат по сравнению с традиционными проекциями (рис. 1.2).
Основные обновления в методах Франк-Вулфа часто основаны на правиле нахождения новой точки:
$\upsilon_t \leftarrow \arg\min_{\upsilon \in P} \langle \nabla f(x_t), \upsilon \rangle$
$x_{t+1} \leftarrow (1 - \gamma_t)x_t + \gamma_t \upsilon_t$,
где любое решение для argmin подходит. В качестве шага возьмём, например, $\gamma_t = \frac{2}{t+2}$. По сути, LMO определяет направление для спуска. Затем в пределах допустимой области формируются выпуклые комбинации точек для поддержания выполнения ограничений. С точки зрения теории сложности, методы Франк-Вульфа сводят оптимизацию выпуклой функции f на P к повторной оптимизации эволюционирующих линейных функций на P.
Схема самой базовой вариации алгоритма Франк-Вульфа приведена ниже:
1 $x_0 \in P$
2 $for\ t = 0\ to\ T - 1\ do:$
3 $\upsilon_t \leftarrow \arg\min_{\upsilon \in P} \langle \nabla f(x_t), \upsilon \rangle$
4 $x_{t+1} \leftarrow (1 - \gamma_t)x_t + \gamma_t \upsilon_t$
5 end for

Глава 2

Алгоритм Франк-Вульфа

Основная идея алгоритма Франк-Вульфа заключается в том, чтобы не следовать отрицательному градиенту, а следовать альтернативному направлению спуска, которое достаточно хорошо согласовано с отрицательным градиентом, обеспечивает достаточный первичный прогресс, и для которого мы можем легко обеспечить выполнимость путем вычисления выпуклых комбинаций. Это можно сделать с помощью вышеупомянутого оракула линейной минимизации (LMO), с помощью которого мы можем оптимизировать отрицательную часть градиента над выполнимой областью P, а затем взять полученную вершину для формирования альтернативного направления спуска.
LMO принимает на вход линейную функцию c, на выходе оракула мы имеем $v \in \arg\min_{x \in P} \langle c, x \rangle$. Необходимо обратить внимание, что v не обязательно является уникальным, и без потери общности мы предполагаем, что v – крайняя точка P. В контексте метода Франк-Вульфа такие точки часто называются *атомами*.
Общая схема процесса представлена на рисунке 2.1[1] ниже.

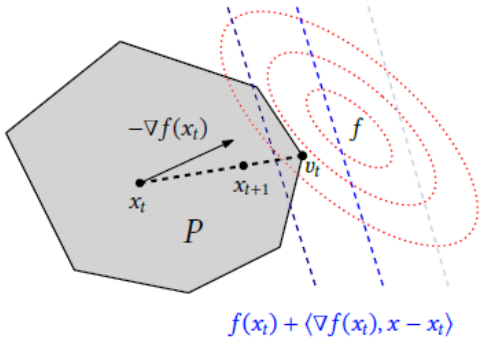

Рис. 2.1

Шаг Франк-Вульфа: для минимизации выпуклой функции f над политопом P строится линейная аппроксимация f в точке x_t, как $f(x_t) + \langle \nabla f(x_t), x - x_t \rangle$. Вершина Франк-Вульфа v_t минимизирует эту аппроксимацию. Шаг переходит от x_t

к x_{t+1}, двигаясь в сторону υ_t, определяемую правилом выбора шага. Красные линии контуров функции f, синие — линейная аппроксимация.

Опр Frank-Wolf gap:

- $f(x) - f(x^*)$ – primal gap
- $\langle \nabla f(x), x - x^* \rangle$ – dual gap
- $max_{\upsilon \in P} \langle \nabla f(x), x - \upsilon \rangle$ – FW gap

Лемма 1 Пусть f выпуклая функция, тогда для $\forall x \in P$
$f(x) - f(x^*) \leq \langle \nabla f(x), x - x^* \rangle \leq max_{\upsilon \in P} \langle \nabla f(x), x - \upsilon \rangle$
Доказательство непосредственно следует из выпуклости и определения υ.

Глава 3

Шаги в методе Франк-Вульфа

Шаг, который мы использовали выше $\gamma = \frac{2}{t+2}$ является базовой стратегией и носит название "открытый цикл"(**open loop**) или "агностический"размер шага. Выбор такого шага делает алгоритм независимым от параметров (не требует никаких параметров функции или их оценок), однако во многих важных случаях есть лучшие варианты.
Другим вариантом шага Франк-Вульфа является короткий шаг (**short step**).
Рассмотрим лемму:
Лемма 2. *Оценка изменения значения целевой функции для гладких задач.*
Пусть f – L-гладкая функция и $x_{t+1} = (1-\gamma)x_t + \gamma_t \upsilon_t$, где $x_t, \upsilon_t \in P$. Тогда:
$f(x_t) - f(x_{t+1}) \geq \gamma_t \langle \nabla f(x_t), x_t - \upsilon_t \rangle - \gamma_t^2 \frac{L}{2} \|x_t - \upsilon_t\|^2$.
Доказательство:
Из неравенства гладкости:
$f(y) - f(x) \leq \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2$
Возьмём $x = x_t, y = x_{t+1}$, подставим определение x_{t+1}. Это даст нам искомое неравенство. □
Теперь, вместо того, чтобы подставлять размер шага *открытого цикла*, мы можем рассматривать правую часть как выражение с одной переменной γ_t и максимизировать его. Это приводит к выбору:
$\gamma_t = \frac{\langle \nabla f(x_t), x_t - \upsilon_t \rangle}{L\|x_t - \upsilon_t\|^2}$ и $f(x_t) - f(_{t+1}) \geq \frac{\langle \nabla f(x), x_t - \upsilon_t \rangle}{2L\|x_t - \upsilon_t\|^2}$
Технически мы можем формировать выпуклые комбинации только если $\gamma_t \in [0, 1]$, поэтому мы должны ограничить $\gamma_t = min\{\frac{\langle \nabla f(x), x_t - \upsilon_t \rangle}{L\|x_t - \upsilon_t\|^2}, 1\}$. Заметим, что γ_t неотрицательный, т.к. FW gap $\langle \nabla f(x), x_t - \upsilon_t \rangle \geq 0$. Такой шаг в совокупности с *леммой 2* приводит к хорошей сходимости:
$f(x_t) - f(x_{t+1}) \geq \langle \nabla f(x), x_t - \upsilon_t \rangle / 2 \geq (f(x_t) - f(x^*))/2$,
то есть прогресс целевой функции составляет не менее половины от FW gap и, следовательно, не менее половины разницы между текущим значением функции и ее минимумом. Стратегия *коротких шагов* позволяет избежать накладных расходов на линейный поиск, однако, к сожалению, она требует знания константы гладкости L или, по крайней мере, достаточно точной ее верхней оценки. Конечно отчасти эту проблему решил Pedgerosa (2020) в своей статье, динамически аппрок-

симируя L, что приводит к лишь немного более медленной скорости сходимости с постоянным коэффициентом. Вкратце, в этой статье выполняется мультипликативный поиск L до тех пор, пока неравенство гладкости:
$f(x_t) - f(x_{t+1}) \geq \gamma_t \langle \nabla f(x_t), x_t - \upsilon_t \rangle - \gamma_t^2 \frac{M}{2} \|x_t - \upsilon_t\|^2$
не выполнится для аппроксимации M величины L и короткого шага γ_t.
Так чем же плох адаптивный шаг? К сожалению, на практике проверка адаптивного неравенства может быть очень сложной, поскольку мы смешиваем вычисления значений функции, вычисления градиента и квадратичные нормы. Вместо этого давайте рассмотрим новый вариант адаптивной стратегии размера шага, где мы полагаемся на другой тест для принятия оценки M величины L:
$\langle \nabla f(x_t), x_t - \upsilon_t \rangle \geq 0$, где
$x_{t+1} = (1-\gamma_t)x_t + \gamma_t \upsilon_t$ как и раньше с $\gamma_t = min\{\frac{\langle \nabla f(x), x_t - \upsilon_t \rangle}{M\|x_t - \upsilon_t\|^2}, 1\}$. То есть мы тестируем только скалярные произведения с градиентов в разных точках. Это приводит к адаптивной стратегии размера шага, приведённой в алгоритме ниже:
Adaptive step-size strategy Input: Objective function f, smoothness estimate $\tilde{L}$, feasible points x, v with $\langle \nabla f(x), x - v \rangle \geq 0$, progress parameters $\eta \leq 1 < \tau$
Output: Updated estimate $\tilde{L}^*$, step-size γ
1. $M \leftarrow \eta \tilde{L}$
2. **loop**
3. $\quad \gamma \leftarrow \min\{\langle \nabla f(x), x - v \rangle / (M\|x - v\|^2), 1\}$
4. $\quad$ **if** $\langle \nabla f(x + \gamma(v - x)), x - v \rangle \geq 0$ **then**
5. $\quad\quad \tilde{L}^* \leftarrow M$
6. $\quad\quad$ **return** $\tilde{L}^*, \gamma$
7. $\quad$ **end if**
8. $\quad M \leftarrow \tau M$
9. **end loop**

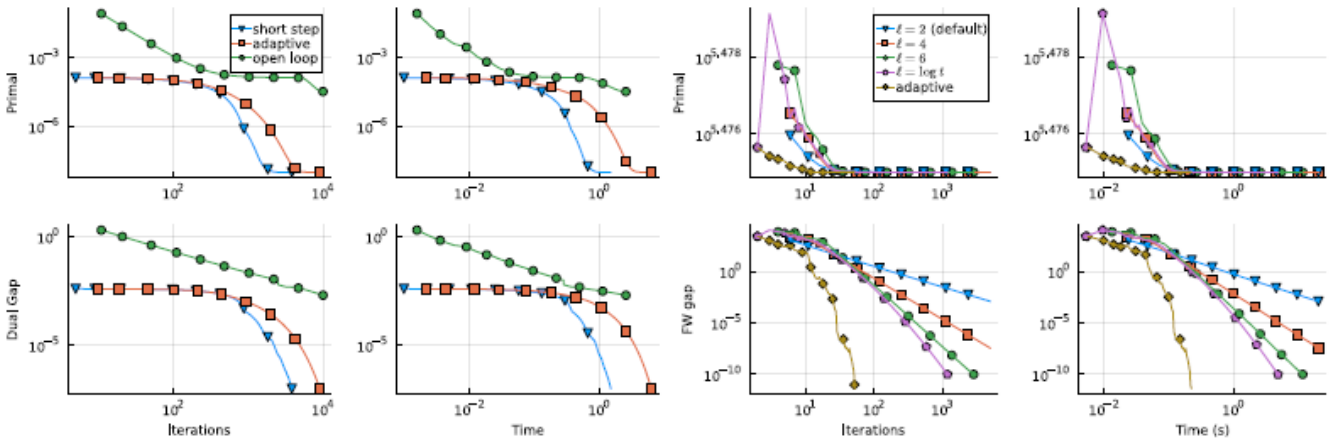

Рис. 3.1: Сравнение шагов.

На рис. 3.1[1] на графиках слева приведены скорости сходимости для простой

квадратичной функции над K-разреженным многогранником с тремя различными стратегиями размера шага. Справа – Скорость сходимости для простой квадратичной функции над K-разреженным многогранником для стратегий с открытым циклом вида $\gamma_t = \frac{l}{l+t}$. Мы видим, что (в зависимости от особенностей задачи) большие значения l достигают скорости сходимости более высокого порядка. Для сравнения также включена адаптивная стратегия размера шага. График построен в логарифмическом масштабе, так что порядок сходимости соответствует различным наклонам траекторий.
В дальнейшем мы будем использовать *short step* без динамической аппроксимации L для упрощения теоретических выкладок. Однако добавление этой стратегии не изменит полученного нами результата.

Теорема 1 Если f выпуклая и в качестве шага рассматривается *short step*, тогда $\forall t \geq 1$:
$f(x_t) - f^* \leq \frac{2LD^2}{t+2}$
Доказательство *Теоремы 1* можно найти в статье [1].

Глава 4

Неточный градиент и применение метода Франк-Вульфа с неточным градиентом

Заметим, что в некоторых случаях вычисление точного градиента целевой функции может быть затруднено или требовать больших вычислительных затрат. В таких ситуациях можно использовать метод Франк-Вулфа с неточным градиентом.
В случае неточного градиента мы предполагаем, что метод имеет доступ не к точному, а к приблизительному значению градиента $\widetilde{\nabla} f(x)$ в любой запрашиваемой точке x, что означает следующее:
$\nabla f(x) = \widetilde{\nabla} f(x) + \upsilon(x), \|\upsilon(x)\| \leq \Delta$

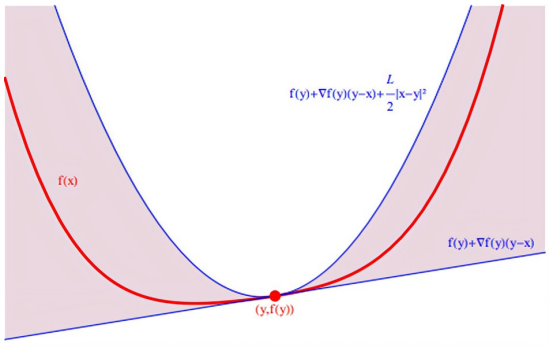

Рис. 4.1: точный оракул[3]

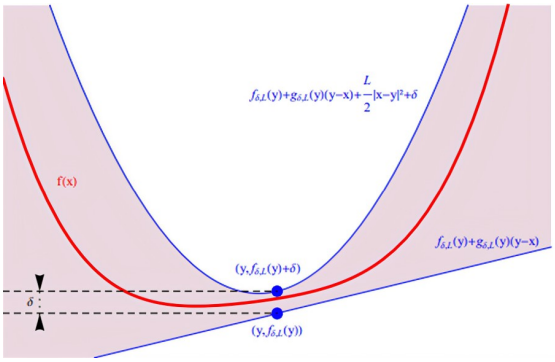

Рис. 4.2: неточный оракул[3]

Существует ряд причин, по которым использование неточного градиента может быть естественным:

- Зашумленные данные: В реальных задачах данные часто содержат шум, что затрудняет вычисление точного градиента.
- Сложные функции: Градиент сложных функций, таких как глубокие нейронные сети, может быть дорогостоящим для вычисления. Вычисление неточно-

го градиента с помощью методов стохастической оптимизации или аппроксимации может значительно ускорить процесс оптимизации.

- Аппроксимация функции: В некоторых случаях целевая функция может быть доступна только через аппроксимацию, например, через интерполяцию или стохастическую модель. В таких случаях вычисление точного градиента невозможно.

- Дифференциальная приватность[4]: Добавление шума к градиенту для защиты конфиденциальности данных делает его неточным. Алгоритм Франка-Вулфа с неточным градиентом может использоваться для решения задач оптимизации с дифференциальной приватностью.

- Распределённая оптимизация: В распределённых системах вычисление точного градиента может быть затруднено из-за обмена информацией между узлами. Использование неточных градиентов, вычисленных локально на каждом узле, может быть более эффективным.

- Оптимизация и метод Монте-Карло[5]: В задачах оптимизации Монте-Карло градиент оценивается с помощью случайных выборок, что неизбежно приводит к некоторой неточности

- Онлайн - оптимизация: Может не быть достаточно времени для точного вычисления градиента, поэтому приближение становится необходимым.

Глава 5

Полученные результаты

Прежде чем перейти к полученным результатам, напомним неравенство интерполяции:
Лемма 3 *Интерполяция.* аналог L-гладкости f для неточного градиента:
$f(y) - f(x) \leq \langle \widetilde{\nabla} f(x), y - x \rangle + L\|y - x\|^2 + \frac{\Delta^2}{2L}$, $\forall x, y \in P$
Доказательство *леммы 3* можно найти в статье [2].

Алгоритм Франк-Вульфа с использованием неточного оракула:
1 Выбираем $x_0 \in P$
2 for $k = 0$, ...
3 If x_k удовлетворяет некоторым условиям – останавливаемся
4 $\upsilon_k \in LMO_\upsilon(\widetilde{\nabla} f(x_k))$
5 $d_k^{FW} = \upsilon_k - x_k$
6 $x_{k+1} = x_k + \gamma_k d_k^{FW}$
7 End for

Где $\gamma_k = min\{\frac{\langle \widetilde{\nabla} f(x_k), x_k - \upsilon_k \rangle}{2L\|x_k - \upsilon_k\|^2}, 1\}$. Здесь мы подставили в *short step* неточный градиент и разделили *short step* пополам, если он меньше 1 (это понадобится нам для доказательства сходимости).
Далее будем работать в предположении $\gamma_k < 1$. Этого можно добиться за счёт конечности кол-ва итераций и увеличения L.
Ситуацию $\gamma_k = 1$ планируетм изучить дополнительно.
Рассмотрим неравенство интерполяции (*Лемма 3*) и подставим туда сначала $x_{k+1} = x_k + \gamma_k d_k$, а затем *short step* $\gamma_k = \frac{-\langle \widetilde{\nabla} f(x_k), d_k \rangle}{2L\|d_k\|^2}$, тогда:
$f(x_{k+1}) \leq f(x_k) + \langle \widetilde{\nabla} f(x_k), x_{k+1} - x_k \rangle + L\|x_{k+1} - x_k\|^2 + \frac{\Delta^2}{2L} \leq$
$\leq f(x_k) + \gamma_k \langle \widetilde{\nabla} f(x_k), d_k \rangle + L\gamma_t^2 \|d_k\|^2 + \frac{\Delta^2}{2L} =$
$= f(x_k) - \frac{\langle \widetilde{\nabla} f(x_k), d_k \rangle^2}{2L\|d_k\|^2} + \|d_k\|^2 \frac{\langle \widetilde{\nabla} f(x_k), d_k^2 \rangle}{4L\|d_k\|^4} + \frac{\Delta^2}{2L} = f(x_k) - \frac{\langle \widetilde{\nabla} f(x_k), d_k \rangle^2}{4L\|d_k\|^2} + \frac{\Delta^2}{2L}$
Таким образом, получили:
Лемма 4 *Базовое неравенство*
$f(x_{k+1}) - f(x_k) \leq -\frac{\langle \widetilde{\nabla} f(x_k), d_k \rangle^2}{4L\|d_k\|^2} + \frac{\Delta^2}{2L}$

Лемма 5 *Оценка на зазор двойственности для неточного градиента*
$-\langle\widetilde{\nabla} f(x_k), x^* - x_k\rangle \geq f(x_k) - f^* - D\Delta$
Доказательство:
$-\langle\widetilde{\nabla} f(x_k), x^* - x_k\rangle = -\langle\nabla f(x_k) - \upsilon(x), x^* - x_k\rangle =$
$= -\langle\nabla f(x_k), x^* - x_k\rangle + \langle\upsilon(x), x^* - x_k\rangle \geq -\langle\nabla f(x_k), x^* - x_k\rangle - \|\upsilon(x)\|\|x^* - x_k\| \geq$
$\geq f(x_k) - f^* - D\Delta$ $\square$

Лемма 6
$-\langle\widetilde{\nabla} f(x_k), \upsilon_k - x_k\rangle^2 \leq -\langle\widetilde{\nabla} f(x_k), x^* - x_k\rangle^2$
Доказательство:
$\langle\widetilde{\nabla} f(x_k), \upsilon_k - x_k\rangle = \langle\widetilde{\nabla} f(x_k), \upsilon_k + (x^* - x_k) - x^*\rangle =$
$= \langle\widetilde{\nabla} f(x_k), x^* - x_k\rangle + \langle\widetilde{\nabla} f(x_k), \upsilon_k - x^*\rangle,$
тогда после возведения в квадрат обоих частей неравенства получим:
$\langle\widetilde{\nabla} f(x_k), \upsilon_k - x_k\rangle^2 \geq \langle\widetilde{\nabla} f(x_k), x^* - x_k\rangle^2 + \langle\widetilde{\nabla} f(x_k), \upsilon_k - x^*\rangle^2$
$-\langle\widetilde{\nabla} f(x_k), \upsilon_k - x_k\rangle^2 \leq -\langle\widetilde{\nabla} f(x_k), x^* - x_k\rangle^2 - \langle\widetilde{\nabla} f(x_k), \upsilon_k - x^*\rangle^2 \leq$
$\leq -\langle\widetilde{\nabla} f(x_k), x^* - x_k\rangle^2$ $\square$

Лемма 7 *обобщённое неравенство для случая неточного градиента*
$f(x_{k+1}) - f^* \leq (f(x_k) - f^*)(1 - \frac{f(x_k) - f^*}{8LD^2}) + \frac{3\Delta^2}{4L}$
Доказательство:
Из *Леммы 4* и *Леммы 6*: $f(x_{k+1}) - f(x_k) \leq -\frac{\langle\widetilde{\nabla} f(x_k), x^* - x_k\rangle^2}{4L\|d_k\|^2} + \frac{\Delta^2}{2L}$
Тогда с помощью *Леммы 5* и $D \geq d_k$ получаем:
$f(x_{k+1}) - f(x_k) \leq -\frac{(f(x_k) - f^* - D\Delta)^2}{4LD^2} + \frac{\Delta^2}{2L}$ (1)
Заметим, что: $(a-b)^2 \geq \frac{a^2}{2} - b^2 \Leftrightarrow -(a-b)^2 \leq -\frac{a^2}{2} + b^2$. Применив к (1) получим:
$f(x_{k+1}) - f(x_k) \leq -\frac{(f(x_k) - f^*)^2}{8LD^2} + \frac{D^2\Delta^2}{4LD^2} + \frac{\Delta^2}{2L} = -\frac{(f(x_k) - f^*)^2}{8LD^2} + \frac{2\Delta^2}{4L}$
$f(x_{k+1}) - f^* \leq f(x_k) - f^* - \frac{(f(x_k) - f^*)^2}{8LD^2} + \frac{3\Delta^2}{4L}$
$f(x_{k+1}) - f^* \leq (f(x_k) - f^*)(1 - \frac{f(x_k) - f^*}{8LD^2}) + \frac{3\Delta^2}{4L}$ $\square$

С помощью *Леммы 7* докажем теорему 2:
Теорема 2 Если f выпуклая и L-гладкая, то для $\forall N \geq 1$:
$f(x_N) - f^* \leq \frac{8LD^2}{N+2} + \frac{3N\Delta^2}{4L}$
Доказательство:
Действительно, для $N = 1$: $f(x_1) - f^* \leq 2LD^2 + \frac{3\Delta^2}{4L} \leq \frac{8LD^2}{3} + \frac{3\Delta^2}{4L}$
Для $N > 1$ докажем по индукции.
Предположение. Для $N = k$: $f(x_k) - f^* \leq \frac{8LD^2}{k+2} + \frac{3k\Delta^2}{4L}$
Индукционный переход. $N = k + 1$. В сущности есть 2 случая:
1) если выполняется, что

$f(x_k) - f^* \leq \frac{8LD^2}{k+3} + \frac{3k\Delta^2}{4L}$
Заметим, что $1 - \frac{f(x_k)-f^*}{8LD^2} < 1$. Применив это к *Лемме 7* получим:
$f(x_{k+1} - f^* \leq f(x_k) - f^* + \frac{3\Delta^2}{4L} < \frac{8LD^2}{k+3} + \frac{3(k+1)\Delta^2}{4L}$ $\square$
2) если выполняется, что:
$f(x_k) - f^* > \frac{8LD^2}{k+3} + \frac{3k\Delta^2}{4L}$, тогда:
$-\frac{f(x_k)-f^*}{8LD^2} < -\frac{1}{k+3} - \frac{3k\Delta^2}{32L^2D^2}$
$1 - \frac{f(x_k)-f^*}{8LD^2} < 1 - \frac{1}{k+3} - \frac{3k\Delta^2}{32L^2D^2} < 1 - \frac{1}{k+3} = \frac{k+2}{k+3}$
Применив полученный выше результат к *Лемме 7*, получаем:
$f(x_{k+1}) - f^* \leq (f(x_k) - f^*)(1 - \frac{f(x_k)-f^*}{8LD^2} + \frac{3\Delta^2}{4L} < (\frac{8LD^2}{k+2} + \frac{3k\Delta^2}{4L})\frac{k+2}{k+3} + \frac{3\Delta^2}{4L}$

Из *Теоремы 2* следует сходимость последовательности x_k в некоторой окрестности, однако в этом способе происходит накопление в оценке скорости сходимости величины, зависящей от погрешности градиента.
Получим альтернативную оценку качества, выдаваемого методом приближённого решения задачи, сходимости для метода Франк-Вульфа с неточным градиентом:

Теорема 3 *Справедливо следующее неравенство*
$f(x_N) - f^* \leq \frac{\sqrt{4LD^2(f(x_1)-f^*)}}{\sqrt{N}} + \sqrt{2}D\Delta$
Доказательство:
Из *Леммы 4*:
$f(x_k) - f(x_{k+1}) \geq \frac{\langle\widetilde{\nabla}f(x_k),d_k\rangle^2}{4LD^2} - \frac{\Delta^2}{2L}$
Просуммируем k от 1 до N: $f(x_1) - f(x_N) \geq \sum_{k=1}^{N} \frac{\langle\widetilde{\nabla}f(x_k),d_k\rangle^2}{4LD^2} - \frac{N\Delta^2}{2L}$
$f(x_1) - f(x_N) \geq \frac{N}{4LD^2} min_{k\in\overline{1,N}}\langle\widetilde{\nabla}f(x_k), d_k\rangle^2 - \frac{N\Delta^2}{2L}$, отсюда получаем:
$min_{k\in\overline{1,N}}\langle\widetilde{\nabla}f(x_k), d_k\rangle^2 \leq \frac{4LD^2}{N}(f(x_1) - f(x_N)) + 2D^2\Delta^2 \leq$
$\leq \frac{4LD^2}{N}(f(x_1) - f^*) + 2D^2\Delta^2$
Прежде чем продолжить доказательство, дадим определение:
Невязка функции – разница между фактическим значением функции и её приближенным значением, полученным с помощью алгоритма.
Оценка невязки для выпуклых функций: $(f(x_N) - f^*)^2 \leq min_{k\in\overline{1,N}}\langle\widetilde{\nabla}f(x_k), d_k\rangle^2$
$f(x_N) - f^*)^2 \leq \frac{4LD^2}{N}(f(x_1) - f^*) + 2D^2\Delta^2$
т.к. $\forall a, b > 0$ выполнено: $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, то:
$f(x_N) - f^* \leq \sqrt{\frac{4LD^2}{N}(f(x_1) - f^*) + 2D^2\Delta^2} \leq \frac{\sqrt{4LD^2(f(x_1)-f^*)}}{\sqrt{N}} + \sqrt{2}D\Delta$ $\square$

Смотря на сходимость можно сказать, что вторая оценка дает результат хуже, чем первая. Однако, в первой оценке при больших N ошибка может накапливаться, поэтому первую оценку можно использовать при небольшом количестве итераций N, а вторую когда мы можем оценить $f(x_1) - f^*$ и эта разность не очень велика.

Выводы

В данной работе мы ознакомились с методом Франк-Вульфа для нахождения минимума функций на выпуклом множестве. Целью нашей работы было исследование сходимости метода Франк-Вульфа с неточным градиентом. Нами были получены две оценки (*Теорема 2*, *Теорема 3*, которые заметно отличаются друг от друга.

Первая оценка $f(x_N) - f^* \leq \frac{8LD^2}{N+2} + \frac{3N\Delta^2}{4L}$ очень похожа на оценку сходимости метода условного градиента (*Теорема 1*, но имеет существенный недостаток ввиде дополнительного слагаемого, включающего в себя произведение числа итераций алгоритма на квадрат ошибки, что при большом числе итераций может давать плохие результаты.

Вторая оценка $f(x_N) - f^* \leq \frac{\sqrt{4LD^2(f(x_1)-f^*)}}{\sqrt{N}} + \sqrt{2}D\Delta$ имеет худшую скорость сходимости, что говорит о не самых точных приближениях, используемых при ее получении . Так же разность $f(x_1) - f*$ в числителе может быть не очень информативным показателем.

Сравнивая две полученные выше оценки, можно сказать, что первая хороша, когда количество шагов алгоритма не очень большое, например, когда мы ограничиваем выполнение алгоритма конкретным числом итераций, в противном случае накопится большая ошибка. Вторая оценка имеет существенный недостаток из-за того, что имеет не линейную сходимость, но в то же время накопление ошибки не зависит от числа итераций, что может быть существенным плюсом. При выборе оценки нужно так же следить за константами L и D, а так же ориентироваться на исходную задачу и ее характерные черты.

Теорему 2 выводили Ширяев Дмитрий, Никитин Иван, *Теорему 3* – Ядров Платон.

Литература

[1] "The Frank-Wolfe algorithm: a short introduction" Sebastian Pokutta
https://arxiv.org/pdf/2311.05313.pdf

[2] "stopping rules for gradient methods for non-convex problems with additive noise in gradient" Fedor Stonyakin, Ilya Kuruzov, Boris Polyak
https://arxiv.org/pdf/2205.07544.pdf

[3] "first-order methods of smooth convex optimization with inexact oracle." Olivier Devolder, Fran¸cois Glineur, Yurii Nesterov
https://optimization-online.org/wp-content/uploads/2010/12/2865.pdf

[4] https://habr.com/ru/articles/395313/

[5] "Нелинейная стохастическая оптимизация методом Монте-Карло" Л. Сакалаускас
https://math.spbu.ru/user/gran/sb1/sakal.pdf