Аннотация

Градиентный слайдинг для седловых задач с композитами

Антышев Тихон Глебович

В данной работе рассматриваются седловые задачи с композитами вида $\min\limits_{x}\max\limits_{y} f(x) + F(x,y)$, где функция $F(x,y)$ - L_F-гладкая, μ-сильно выпуклая и сильно вогнутая, а функция $f(x)$ - выпуклая и соответственно L_f и L_g гладкие. В данной работе приводится расширение алгоритма слайдинга на подобную задачу, которое достигает $\mathcal{O}\left(\frac{L_f}{\mu}\log\frac{R_0^2}{\varepsilon}\right)$ оракульных вызовов градиента $\nabla F(x,y)$ и $\mathcal{O}\left(\frac{L_f}{\mu}\log\frac{R_0^2}{\varepsilon}\right)$ оракульных вызовов градиентов $\nabla f(x)$. Также приводятся различные варианты внутренних методов и оценки к ним.

Оглавление

Введение

Седловые задачи являются одним из классов задач оптимизации, которые возникают в различных областях, включая теорию игр, экономику и машинное обучение. Они характеризуются наличием седловой точки в целевой функции, где градиент равен нулю в одном направлении и ненулевой в другом направлении.

Эти задачи известны своей сложностью в решении, а также ограничениями существующих методов оптимизации, которые либо сходятся медленно к решению, либо являются нестабильными.

За последние годы возрос интерес к разработке новых методов оптимизации для решения седловых задач. Один из перспективных подходов заключается в использовании оптимизационных методов первого порядка, которые объединяют градиентный спуск с проксимальными операторами и двойственными переменными. Эти методы показали свою эффективность при решении выпуклых задач оптимизации и могут быть адаптированы для работы с седловыми задачами.

В данной работе основное внимание уделяется седловым задачам с композитной структурой. Мы предполагаем необходимость разделения оптимизации по каждой из целевых функций в сумме и достижение разделения оракульных сложностей.

Композитная оптимизация является гибкой структурой для решения задач оптимизации, связанных с сложными композитными целевыми функциями, широко распространеных во многих областях, например в распределённом машинном обучении.

Такие целевые функции обычно состоят из суммы нескольких компонентных функций, каждая из которых вносит свой вклад в отдельный аспект общей задачи оптимизации.

Основной целью этой диссертации является исследование и разработка метода градиентного слайдинга, специально адаптированного для седловых задач с композитными функциями, а также получение оценок его сходимости в зависимости от внутренних методов и параметров задачи.

Достоверность полученных результатов подтверждается оценками алгоритмов для обычных задач минимизации, полученными в других работах, перечисленных в списке литературы.

На основе указанных работ в списке литературы автор данной работы провел теоретическое исследование алгоритма слайдинга для седловых задач и получил оценки его сходимости.

Объем и структура работы. Диссертация состоит из введения, четырёх глав, заключения и двух приложений. Полный объём диссертации составляет 29 страниц с 0 рисунками и 0 таблицами. Список литературы содержит 12 наименований.

Глава 1. Слайдинг для сёдел

1.1 Предварительные предположения

Рассматривается следующая задача оптимизации

$$\min_{x\in\mathbb{R}^{d_x}} \max_{y\in\mathbb{R}^{d_y}} f(x) + F(x,y) \tag{1.1}$$

где функция $F(x, y) : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R}$ is L_F-гладкая, $f(x) : \mathbb{R}^{d_x} \to \mathbb{R}$ и - выпуклые и L_f-гладкая соответсвенно. Также мы предполагаем μ-сильную выпуклость и вогнутость седловой задачи. Композит данной задачи не обязательно являются проксимально-дружественными.

Определение 1 (L-глакость). Дифференцируемая функция $f(x) : \mathbb{R}^n \to \mathbb{R}$ называется L-гладкой с некоторой константой $L > 0$, если её градиент L-Липшицев, т.е. $\forall\, x, y \in \mathbb{R}^n$ выполняется

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \tag{1.2}$$

Определение 2 (μ-сильная выпуклость). Дифференцируемая функция $f(x) : \mathbb{R}^n \to \mathbb{R}$ называется μ-сильно выпуклой с некоторой константой $\mu > 0$, если е $\forall\, x, y \in \mathbb{R}^n$ выполняется

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2 \tag{1.3}$$

Определение 3 (ε-решение). Пара $(\hat{x}, \hat{y})$ называется ε-решением задачи (1.1), если

$$\|\hat{x} - x^*\|^2 + \|\hat{y} - y^*\|^2 \leq \varepsilon \tag{1.4}$$

Также для построения алгоритма слайдинга необходимо ввести проксимальной оператор $prox_{\eta f(\cdot)}(\hat{x})$.

Определение 4 (оператор $prox_{\eta f(\cdot)}(\hat{x})$). Проксимальным оператором называется

$$prox_{\eta f(\cdot)}(\hat{x}) = \arg\min_x \left(f(x) + \frac{1}{2\eta}\|x - \hat{x}\| \right) \tag{1.5}$$

Изначально вид данного оператора получается из интегрирования обратной схемы Эйлера. Таким образом, в данной работе под термином проксимально-дружественная функция понимается такая функция, что её проксимальный оператор может быть легко вычислен.

Примечание 1. В данной работе под обозначением $\|\cdot\|$ понимается Евклидова норма.

Таккже в данной работе используются стандартные обозначения из книг по методам оптимизации [1]. Так, например, (x^*, y^*) обозначает решение задачи (1.1), R_0 - расстояние от начальной точки алгоритма до решения, $\kappa = \frac{L}{\mu}$ - число обусловленности задачи.

Предполагается, что алгоритм решения оптимизационных задач может обращаться к градиентному ораклу всех функций (1.1) не более чем константное ($\mathcal{O}(1)$) число раз.

1.2 Градиентный слайдинг

Градиентный слайдинг, также известный как композитная оптимизация - это метод оптимизации первого порядка, который соединяет свойства градиентного спуска и проксимальных алгоритмов, впервые предложенный в работе [2]. Предлагается рассмотреть следующую задачу композитной оптимизации

$$\min_x p(x) + q(x) \tag{1.6}$$

где функции $p(x)$ и $q(x)$ являются соответсвенно L_p- и L_q-гладкими, также функция $q(x)$ является μ-сильно выпуклой. В самом общем виде подобная задача просто решается ускоренным методом Нестерова [3], данный алгоритм является модификацией метода тяжёлого шарика, предложенного в [4]. Для нахождения ε-решения алгоритму понадобиться $\mathcal{O}\left(\sqrt{\frac{L_p+L_q}{\mu}} \log \frac{1}{\varepsilon}\right)$ вызовов градиента функций.

Если дополнительно предположить, что функция $q(x)$ является проксимально-дружественной, т.е. у нас имеется точно вычислимый проксимальный оператор $prox_{\eta q(\cdot)}(\cdot)$, то задачу можно решить ускоренным проксимальным методом [5]. В таком случае алгоритму потребуется $\mathcal{O}\left(\sqrt{\frac{L_p}{\mu}}\log\frac{1}{\varepsilon}\right)$ вызовов градиента $p(x)$ для нахождения ε-решения задачи (1.6). Теперь необходимо рассмотреть случай, когда $q(x)$ является L_q-гладкой, но при этом не является проксимально-дружественной.

Задача при таких условиях усложняется: наличие проксимального оператора позволяло эффективно решать подзадачу алгоритма, однако в теперь возникающую подзадачу придётся решать приближённо численно.

И так как в подзадаче будет присутствовать только одна из функций задачи (1.6), то удаётся разделить исходную задачу на две. Мы рассмотрим оптимальный алгоритм композитной оптимизации приведённый в [6].

Algorithm 1 Оптимальная Композитная Оптимизация

Входные данные: Начальная точка $x^0 = x_f^0$
Параметры: $\tau \in (0,1)$, $\eta, \theta, \alpha > 0$, $N \in \{1,2,...\}$
for $k = 0,1,2\dots,N-1$ **do**
  $x_g^k = \tau x^k + (1-\tau)x_f^k$
  $x_f^{k+1} \approx \arg\min\limits_{x\in\mathbb{R}^d}\left[A_\theta^k(x) := p(x_g^k) + \langle\nabla p(x_g^k), x - x_g^k\rangle + \frac{1}{2\theta}\|x - x_g^k\|^2 + q(x)\right]$
  $x^{k+1} = x^k + \eta\alpha(x_f^{k+1} - x^k) - \eta\left(\nabla p(x_f^{k+1}) + \nabla q(x_f^{k+1})\right)$
end for
Возвращает: x^N

Для нахождения ε-решения данному алгоритму потребуется $\mathcal{O}\left(\sqrt{\frac{L_p}{\mu}}\log\frac{1}{\varepsilon}\right)$ вызовов градиента $p(x)$ и $\mathcal{O}\left(\sqrt{\frac{L_q}{\mu}}\log\frac{1}{\varepsilon}\right)$ вызовов градиента $q(x)$. Таким образом мы смогли разделить оракульные сложности для задачи (1.6). Именно этот приём и назвается слайдингом. Даваемое им расщепление может быть выгодно, например в тех случаях, когда вычисление градиента одного из композитов много больше чем у другого [7; 8].

1.3 Модификация алгоритма

Теперь проведём подификацию алгоритма предложенного алгоритма для решения задачи (1.1). В данном случае требуется пропустить вычисления градиентов по композиту $f(x)$, поэтому подзадача будет содержать градиент композита в какой-то точке и оптимизироваться относительно $F(x,y)$ и дополнительных членов.

Также при обновлении параметров мы используем моментный член и берём градиент функции $F(x,y)$ по переменной y с противоположным знаком, так как в задаче (1.1) происходит максимизация по y.

1.3.1 Случай сильной выпуклости-вогнутости, $\mu > 0$

Algorithm 2 Седловой Слайдинг

1: **Входные данные:** Начальные точки $x^0 \in \mathbb{R}^{d_x}$, $y^0 \in \mathbb{R}^{d_y}$
2: **Параметры:** α, θ, $\eta > 0$, $N \in \{1,2,...\}$
3: **for** $k = 0,1,2,\ldots,N-1$ **do**
4: $(x_f^k, y_f^k) \approx \arg\min_{x\in\mathbb{R}^{d_x}} \max_{y\in\mathbb{R}^{d_y}} A_\theta^k(x,y)$

$$A_\theta^k(x,y) := \langle \nabla f(x^k), x\rangle + F(x,y) + \frac{1}{2\theta}\|x - x^k\|^2 - \frac{1}{2\theta}\|y - y^k\|^2 \tag{1.7}$$

5: $x^{k+1} = x^k + \alpha\eta(x_f^{k+1} - x^k) - \eta\left(\nabla f(x_f^k) + \nabla_x F(x_f^{k+1}, y_f^{k+1})\right)$
6: $y^{k+1} = y^k + \alpha\eta(y_f^{k+1} - y^k) + \eta\nabla_y F(x_f^{k+1}, y_f^{k+1})$
7: **end for**
8: **Возвращает:** x^N, y^N

Заметим, что для подзадачи (1.7) существует эквивалентный вид, позволяющий рассмотреть её в градиентной форме

$$B_\theta^k(x,y) = \begin{pmatrix} \nabla f(x^k) \\ 0 \end{pmatrix} + \begin{pmatrix} \nabla_x F(x,y) \\ -\nabla_y F(x,y) \end{pmatrix} + \frac{1}{\theta}\begin{pmatrix} x - x^k \\ y - y^k \end{pmatrix} \tag{1.8}$$

Таким образом пара $(\overline{x}_f^k, \overline{y}_f^k)$ - решение уравнения $B_\theta^k(x,y) = 0$ тогда и только тогда, когда оно является и решением подзадачи (1.7).

1.3.2 Случай выпуклости-вогнутости, $\mu = 0$

Стоит заметить, что возможно модифицировать представленный Алгоритм **??** для задачи (1.1), в которой $\mu = 0$. Для этого достаточно убрать моментный член при обновлении переменных:

Algorithm 3 Седловой Слайдинг для $\mu = 0$

1: **Входные данные:** Начальные точки $x^0 \in \mathbb{R}^{d_x}$, $y^0 \in \mathbb{R}^{d_y}$

2: **Параметры:** θ, $\eta > 0$, $N \in \{1,2,...\}$

3: **for** $k = 0,1,2,\ldots,N-1$ **do**

4: $(x_f^k, y_f^k) \approx \arg\min_{x\in\mathbb{R}^{d_x}} \max_{y\in\mathbb{R}^{d_y}} A_\theta^k(x,y)$

$$A_\theta^k(x,y) := \langle \nabla f(x^k), x\rangle + F(x,y) + \frac{1}{2\theta}\|x - x^k\|^2 - \frac{1}{2\theta}\|y - y^k\|^2$$

5: $x^{k+1} = x^k - \eta\left(\nabla f(x_f^k) + \nabla_x F(x_f^{k+1}, y_f^{k+1})\right)$

6: $y^{k+1} = y^k + \eta \nabla_y F(x_f^{k+1}, y_f^{k+1})$

7: **end for**

8: **Возвращает:** x^N, y^N

Глава 2. Сходимость алгоритма

2.1 Общая теорема сходимости

Теорема 1. Если для Алгоритма 2 в условиях задачи (1.1) заданы следующие параметры:

$$\theta = \frac{1}{2L_f}, \quad \eta = \min\left[\frac{1}{4\mu}, \frac{1}{4L_f}\right], \quad \alpha = 2\mu \tag{2.1}$$

вспомогательная задача (1.7) решается с такой точностью, что выполнятеся:

$$\left\|B_\theta^k(x_f^k, y_f^k)\right\|^2 \leq \frac{L_f^2}{3}\left\|\begin{pmatrix} x^k - \overline{x}_f^k \\ y^k - \overline{y}_f^k \end{pmatrix}\right\|^2 \tag{2.2}$$

то для любого числа итераций такого, что

$$N \geq 2\max\left[1, \frac{L_f}{\mu}\log\frac{\left\|\begin{pmatrix} x^0 - x^* \\ y^0 - y^* \end{pmatrix}\right\|^2}{\varepsilon}\right] = 2\max\left[1, \frac{L_f}{\mu}\log\frac{R_0^2}{\varepsilon}\right] \tag{2.3}$$

выполняется следующая оценка:

$$\left\|\begin{pmatrix} x^N - x^* \\ y^N - x^* \end{pmatrix}\right\|^2 \leq \varepsilon \tag{2.4}$$

Доказательство: для доказательства необходимо несколько лемм, они приведены в приложении A вместе с доказательствами. Воспользуемся доказанным рекурсионным соотношением (A.2) для алгоритма 2:

$$\left\|\begin{pmatrix} x^N - x^* \\ y^N - x^* \end{pmatrix}\right\|^2 \leq (1-2\eta\mu)^N\left(R_0^2 + 2\frac{\eta}{\tau}\left(F(x^0,y^0) + f(x^0) - F(x^*, y^*) + f(x^*)\right)\right) =$$

$$= Const \cdot (1-\beta)^N$$

Соответсвенно при выборе N указанным способом получаем ε-решение. $\square$

2.2 Вспомогательная задача

В предыдущем пункте мы представили теорему, которая определяет число внешних итераций для получения ε-точного решения задачи (1.1). Однако приведённая теорема выполнятеся только при решении подзадачи (1.7), удовлетворяющей условию (2.2). В качестве внутренних методов предлагается рассмотреть 3: Алгоритм 4 (FOAM) из [9], Алгоритм 2 (GDAE) из [10] и Алгоритм EAG-V из [11].

2.2.1 FOAM

FOAM или First Optimal Algorithm for Minimax Optimization - алгоритм для решения сильно выпулкой, сильно вогнутой седловой оптимизации, который достигает нижней оценки, выведенной в [12]. Запишем теорему сходимости для данного алгоритма.

Теорема 2. Алгоритму FOAM из [9] для решения седловой задачи (1.7) требуется

$$= \mathcal{O}\left(\frac{L_F + 2L_f}{\sqrt{\mu(\mu + 2L_f)}} \log \frac{\sqrt{3}(2L_f + L_F)}{L_f} - \frac{L_F + 2L_f}{\sqrt{\mu(\mu + 2L_f)}} \log R_k\right) \tag{2.5}$$

вызовов градиентного оракула для достижения критерия остановки (2.2), где $R_k = \begin{pmatrix} x^k - \overline{x}_f^k \\ y^k - \overline{y}_f^k \end{pmatrix}$ - расстояния от начальной точки k-ой подзадачи до её решения.

Доказательство. Мы разобьём задачу оптимизации (1.7) следующим образом:

$$F'(x,y) = F(x,y) + \frac{1}{2\theta}\left\|x - x^k\right\|^2$$

В соответсвие с Теоремой 3 из [9] для нахождения ε-точного решения задачи (1.7) алгоритму потребуется

$$N = \mathcal{O}\left(\max\left[\frac{L}{\mu_x}, \frac{L}{\sqrt{\mu_x\mu_y}}\right]\log\frac{1}{\varepsilon}\right) = \mathcal{O}\left(\max\left[\frac{L_F+2L_f}{\mu_x}, \frac{L_F+2L_f}{\sqrt{\mu(\mu+2L_f)}}\right]\log\frac{1}{\varepsilon}\right) =$$

$$= \mathcal{O}\left(\max\left[\frac{L_F+2L_f}{\mu+2L_f}, \frac{L_F+2L_f}{\sqrt{\mu(\mu+2L_f)}}\right]\log\frac{1}{\varepsilon}\right) = \mathcal{O}\left(\frac{L_F+2L_f}{\sqrt{\mu(\mu+2L_f)}}\log\frac{1}{\varepsilon}\right) \tag{2.6}$$

обращений к градиентному оракулу. Тогда при обозначении Тогда Однако нам необходимо найти такое решение, что выполняется критерий остановки (2.2). То есть нужно перейти от оценки нормы к градиента к расстоянию до решения.

В соответствии с Леммой 3, доказанной в A нам известна оценка на необходимую точность оптимизационной задачи k-ой итерации. Тогда, подставляя в (2.6), получаем:

$$N^k = \mathcal{O}\left(\frac{L_F+2L_f}{\sqrt{\mu(\mu+2L_f)}}\log\frac{1}{\varepsilon_k}\right) = \mathcal{O}\left(\frac{L_F+2L_f}{\sqrt{\mu(\mu+2L_f)}}\log\frac{3(2L_f+L_F)^2}{L_f^2\left\|\begin{pmatrix} x^k - \overline{x}_f^k \\ y^k - \overline{y}_f^k\end{pmatrix}\right\|^2}\right) =$$

$$= \mathcal{O}\left(\frac{L_F+2L_f}{\sqrt{\mu(\mu+2L_f)}}\log\frac{3(2L_f+L_F)^2}{L_f^2} - \frac{L_F+2L_f}{\sqrt{\mu(\mu+2L_f)}}\log R_k^2\right)$$

$$= \mathcal{O}\left(\frac{L_F+2L_f}{\sqrt{\mu(\mu+2L_f)}}\log\frac{\sqrt{3}(2L_f+L_F)}{L_f} - \frac{L_F+2L_f}{\sqrt{\mu(\mu+2L_f)}}\log R_k\right)$$

Теорема доказана. $\square$

2.2.2 GDAE

GDAE или Gradient Descent-Ascent with Extrapolation - это линейно сходящийся алгоритм, который требует сильной выпуклости и вогнустости решаемой седловой задачи. Данный алгоритм много проще предыдущего, однако работает с седловой задачей в общем виде, т.е. без композитов.

Теорема 3. Алгоритму GDAE из [10] для решения седловой задачи (1.7) с параметрами $L_F \geq \mu > 2L_f$ требуется

$$\mathcal{O}\left(\min\left[T_a, T_b, T_c, T_d\right] \log \frac{C \cdot (2L_f + L_F)}{L_f R_k}\right) \tag{2.7}$$

вызовов градиентного оракула для достижения критерия остановки (2.2), где параметры T_a, T_b, T_c, T_d определяются как

$$\begin{aligned}
T_a &= \max\left\{\frac{L_x}{\mu_x}, \frac{L_y}{\mu_y}, \frac{L_{xy}}{\sqrt{\mu_x \mu_y}}\right\}, \\
T_b &= \max\left\{\frac{L_x}{\mu_x}, \frac{L_x L_y}{\mu_{xy}^2}, \frac{L_{xy}^2}{\mu_{xy}^2}\right\}, \\
T_c &= \max\left\{\frac{L_y}{\mu_y}, \frac{L_x L_y}{\mu_{yx}^2}, \frac{L_{xy}^2}{\mu_{yx}^2}\right\}, \\
T_d &= \max\left\{\frac{L_x L_y}{\mu_{xy}^2}, \frac{L_x L_y}{\mu_{yx}^2}, \frac{L_{xy}^2}{\mu_{xy}^2}, \frac{L_{xy}^2}{\mu_{yx}^2}\right\},
\end{aligned} \tag{2.8}$$

причём константа C не зависит от итерации и точности решения подзадачи, а указанные в (2.8) величины определяются как

$$L_x = L_F + 2L_f, \quad L_y = L_F - 2L_f \quad L_{xy} = L_F, \quad L_{yx} = L_F \tag{2.9}$$

$$\mu_x = \mu + 2L_f, \quad \mu_y = \mu - 2L_f \tag{2.10}$$

Причём существуют в смысле определения 2.12 μ_{xy} и μ_{yx}

Доказательство. Заметим сначала, что так как алгоритм GDAE работает с седловой задачей в общем виде необходимо узнать параметры вспомогательной задачи. Итак изначально:

$$A_\theta^k(x, y) = \langle \nabla f(x^k), x\rangle + F(x,y) + \frac{1}{2\theta}\|x - x^k\|^2 - \frac{1}{2\theta}\|y - y^k\|^2$$

Найдём градиент

$$\nabla A_\theta^k(x, y) = \begin{pmatrix} \nabla f(x^k) + \nabla_x F(x, y) + \frac{1}{\theta}(x - x^k) \\ \nabla_y F(x, y) - \frac{1}{\theta}(y - y^k) \end{pmatrix} =$$

$$= \begin{pmatrix} \nabla f(x^k) + \nabla_x F(x, y) + 2L_f(x - x^k) \\ \nabla_y F(x, y) - 2L_f(y - y^k) \end{pmatrix}$$

Тогда получим в обозначениях [10] из определениия гладкости и сильной выпуклости через гессиан или просто прямыми вычислениями необходимые параметры.

Сначала заметим, что

$$\nabla_x A_\theta^k(x,y) = \nabla f(x^k) + \nabla_x F(x, y) + 2L_f(x - x^k)$$

является суммой двух Липшецевых функций. Тогда получим, что $L_x = L_F + 2L_f$.

Определение 5(Assumption 6.2 [10]). $L_{xy} > 0$ - это такое число, что для любых $x, x_1, x_2 \in \mathbb{R}^{d_x}$ and $y, y_1, y_2 \in \mathbb{R}^{d_y}$, выполняются следующие неравенства

$$\begin{aligned} \|\nabla_x F(x, y_1) - \nabla_x F(x, y_2)\| &\le L_{xy}\|y_1 - y_2\|, \\ \|\nabla_y F(x_1, y) - \nabla_y F(x_2, y)\| &\le L_{xy}\|x_1 - x_2\|. \end{aligned} \tag{2.11}$$

Вселедствие данного определения получаем, что дополнительные члены в $\nabla_x A_\theta^k(x, y)$ и $\nabla_y A_\theta^k(x, y)$ зануляются и тогда параметр $L_{xy} = L_F$.

Определение 6(Assumption 6.3 [10]). μ_{xy}, $\mu_{yx} > 0$ - это такие числа, что для любых $x, x_1, x_2 \in \mathbb{R}^{d_x}$ and $y, y_1, y_2 \in \mathbb{R}^{d_y}$, выполняются следующие неравенства

$$\begin{aligned} \|\nabla_x F(x, y_1) - \nabla_x F(x, y_2)\| &\geq \mu_{xy} \|y_1 - y_2\|, \\ \|\nabla_y F(x_1, y) - \nabla_y F(x_2, y)\| &\geq \mu_{yx} \|x_1 - x_2\|. \end{aligned} \tag{2.12}$$

Из аналогичных рассуждений мы получаем перевёрнутое неравенство для градиента $F(x,y)$, однако данное неравенств о не следует ни из сильной выпуклости, ни из гладкости градиента, поэтому в данном случае мы будем брать μ_{xy}, μ_{yx} как существующие для данного случая.

Также стоит заметить, что придётся ввести предположение, что $L_F \geq \mu > 2L_f$, так как иначе задача не будет решаться алгоритмом GDAE вследствие потери задачей свойств сильной выпуклости. Тогда по Теореме 6.4 из [10] количество обращений к градиентному оракулу $F(x, y)$ будет равняться:

$$\mathcal{O}\left(\min\left[T_a, T_b, T_c, T_d\right] \log \frac{C}{\varepsilon_k}\right)$$

подставляя ε_k из 2.2, получим необходимое соотношение.

Теорема доказана. □

2.2.3 EAG-V

EAG-V или Extra Anchored Gradient with Varying step-size - это алгоритм для решения выпукло-вогнутых седловых задач, основанный на идее так называемых "якорных коэффициентов". В то время как моментные методы используют разность между переменными оптимизации текущей итерации и предыдущей, EAG-V прибавляет разность между нулевой (начальной) точкой и текущей. Причём её влияние убывает с каждой итерацией.

Ещё одно значительное отличие данного метода от остальных - наличие теоремы сходимости в терминах градиента, а не расстояния до решения, то есть в данном случае мы сможем записать теорему сходимости без использования Леммы 3.

Теорема 4(следствие 2 [11], Теорема 9 [6]). Алгоритму EAG-V из [11] для решения седловой задачи (1.7) требуется

$$N = \mathcal{O}(1) \tag{2.13}$$

вызовов градиентного оракула для достижения критерия остановки (2.2) *Доказательство.* Из указанных теорем можно получить, что для EAG-V решающего подзадачу, выполняется

$$\left\|B_\theta^k(x_f^k)\right\|^2 \leq \frac{C^2 \max\left[L_f^2, L_F^2\right] \cdot R_k^2}{N^2}$$

где N- число итераций, C - некоторая константа, независящая от параметров задачи. Для того, чтобы выполнялось условие остановки (2.2), необходимо выбрать

$$N = C \cdot \sqrt{3} = \mathcal{O}(1)$$

2.3 Итоговая сложность

Теперь когда мы получили сложности для внутреннего и внешнего методов можно получить итоговую оракульную сложность методов.

Теорема 5. Итоговая оракульная сложность алгоритма 2, при внутреннем методе FOAM из [9] при решении задачи (1.1)

$$\mathcal{O}\left(\frac{L_f(L_F + 2L_f)}{\mu\sqrt{\mu(\mu + 2L_f)}} \log \frac{\sqrt{3}R_0^2(2L_f + L_F)}{L_f} \log \frac{R_0^2}{\varepsilon}\right) \tag{2.14}$$

Доказательство. В 2.1 было доказано, что Алгоритм 2 требует

$$\mathcal{O}\left(\frac{L_f}{\mu} \log \frac{R_0^2}{\varepsilon}\right)$$

вызовов градиента для получения ε-точного решения, тогда итоговая оракульная сложность:

$$\mathcal{O}\left(\frac{L_F + 2L_f}{\sqrt{\mu(\mu + 2L_f)}} \log \frac{R_0^2}{\varepsilon}\right) \times$$

$$\times \mathcal{O}\left(\frac{L_F + 2L_f}{\sqrt{\mu(\mu+2L_f)}} \log \frac{\sqrt{3}(2L_f + L_F)}{L_f} - \frac{L_F + 2L_f}{\sqrt{\mu(\mu+2L_f)}} \log R_k\right) =$$

$$= \mathcal{O}\left(\frac{L_f(L_F + 2L_f)}{\mu\sqrt{\mu(\mu+2L_f)}} \log \frac{\sqrt{3}R_0^2(2L_f + L_F)}{L_f} \log \frac{R_0^2}{\varepsilon}\right)$$

Теорема доказана. □

Теорема 6. Итоговая оракульная сложность алгоритма 2, при внутреннем методе GDAE из [10]

$$\mathcal{O}\left(\frac{L_f}{\mu} \min\left[T_a, T_b, T_c, T_d\right] \log \frac{CR_0^2 \cdot (2L_f + L_F)}{L_f} \log \frac{R_0^2}{\varepsilon}\right)) \tag{2.15}$$

Доказательство. Аналогично предыдущей теореме итоговая оракульная сложность

$$\mathcal{O}\left(\frac{L_f}{\mu} \log \frac{R_0^2}{\varepsilon}\right) \times \mathcal{O}\left(\min\left[T_a, T_b, T_c, T_d\right] \log \frac{C \cdot (2L_f + L_F)}{L_f R_k}\right) =$$

$$= \mathcal{O}\left(\frac{L_f}{\mu} \min\left[T_a, T_b, T_c, T_d\right] \log \frac{CR_0^2 \cdot (2L_f + L_F)}{L_f} \log \frac{R_0^2}{\varepsilon}\right)$$

Теорема 7. Итоговая оракульная сложность алгоритма 2, при внутреннем методе EAG-V из [11]

$$\mathcal{O}\left(\frac{L_f}{\mu} \log \frac{R_0^2}{\varepsilon}\right) \times \mathcal{O}(1) = \mathcal{O}\left(\frac{L_f}{\mu} \log \frac{R_0^2}{\varepsilon}\right) \tag{2.16}$$

Заключение

Основные результаты работы заключаются в следующем. На основе анализа разаработанного алгоритма градиентного слайдинга для седловых задач были получены теоремы сходимости для внешнего метода вместе с критерием остановки для внутреннего.

Также был выведен эквивалентный переход между сходимсотью в смысле нормы градиента и расстояния до решения, что позволяет варьировать внутренний метод и получать оценки сходимости, не выводя сходимость по норме градиента для каждой его новой версии.

Стоит заметить, что был рассмотрен неускоренный метод градиентного слайднига и полученные в данной работе оценки не достигают оптимальных и могут быть улучшены.

Список литературы

1. *Polyak B.* Introduction to Optimization. — 07.2020.
2. *Juditsky A.*, *Nemirovskii A. S.*, *Tauvel C.* Solving variational inequalities with Stochastic Mirror-Prox algorithm. — 2011. — arXiv: 0809.0815 [math.OC].
3. *NESTEROV Y.* A method for unconstrained convex minimization problem with the rate of convergence o(1/k^2) // Doklady AN USSR. — 1983. — т. 269. — с. 543—547. — URL: https://cir.nii.ac.jp/crid/1570572699326076416.
4. *Polyak B.* Some methods of speeding up the convergence of iteration methods // USSR Computational Mathematics and Mathematical Physics. — 1964. — т. 4, № 5. — с. 1—17. — DOI: https://doi.org/10.1016/0041-5553(64)90137-5. — URL: https://www.sciencedirect.com/science/article/pii/0041555364901375.
5. *Güler O.* On the Convergence of the Proximal Point Algorithm for Convex Minimization // SIAM Journal on Control and Optimization. — 1991. — т. 29, № 2. — с. 403—419. — DOI: 10.1137/0329022. — eprint: https://doi.org/10.1137/0329022. — URL: https://doi.org/10.1137/0329022.
6. Optimal Gradient Sliding and its Application to Distributed Optimization Under Similarity / D. Kovalev [и др.]. — 2022. — arXiv: 2205.15136 [math.OC].
7. Ускоренный метаалгоритм для задач выпуклой оптимизации / А. В. Гасников [и др.] //. — 2021.
8. *Дж.М. Д.* Теория максмина //. — 1970.
9. *Kovalev D.*, *Gasnikov A.* The First Optimal Algorithm for Smooth and Strongly-Convex-Strongly-Concave Minimax Optimization. — 2022. — arXiv: 2205.05653 [math.OC].
10. *Kovalev D.*, *Gasnikov A.*, *Richtárik P.* Accelerated Primal-Dual Gradient Method for Smooth and Convex-Concave Saddle-Point Problems with Bilinear Coupling. — 2022. — arXiv: 2112.15199 [math.OC].
11. *Yoon T.*, *Ryu E. K.* Accelerated Algorithms for Smooth Convex-Concave Minimax Problems with $\mathcal{O}(1/k^2)$ Rate on Squared Gradient Norm. — 2021. — arXiv: 2102.07922 [math.OC].

12. Linear Lower Bounds and Conditioning of Differentiable Games / A. Ibrahim [и др.]. — 2020. — arXiv: 1906.07300 [cs.LG].

Приложение А

Доказательство сходимости

Лемма 1. Для алгоритма 2 в условиях задачи (1.1) при выборе параметра $\theta = \frac{1}{2L_f}$ выполняется следующее неравенство

$$2\left\langle \begin{pmatrix} x^* - x^k \\ y^* - y^k \end{pmatrix}, \begin{pmatrix} \nabla f(x_f^k) + \nabla_x F(x_f^k, y_f^k) \\ -\nabla_y F(x_f^k, y_f^k) \end{pmatrix} \right\rangle \leq$$

$$\leq -2\mu \left\| \begin{pmatrix} x^* - x_f^k \\ y^* - y_f^k \end{pmatrix} \right\|^2 - \theta \left\| \begin{pmatrix} \nabla f(x_f^k) + \nabla_x F(x_f^k, y_f^k) \\ -\nabla_y F(x_f^k, y_f^k) \end{pmatrix} \right\|^2 +$$

$$+3\theta \left(\|B_\theta^k(x_f^k, y_f^k)\|^2 - \frac{L_f^2}{3} \left\| \begin{pmatrix} x^k - \overline{x}_f^k \\ y^k - \overline{y}_f^k \end{pmatrix} \right\|^2 \right)$$

где $(\overline{x}_f^k, \overline{y}_f^k)$ - точное решение подзадачи на k-ой итерации Алгоритма 2.

Доказательство (приведено из Леммы 4 [6]): В точке (x^*, y^*) выполняется

$$\begin{pmatrix} \nabla f(x^*) + \nabla_x F(x^*, y^*) \\ -\nabla_y F(x^*, y^*) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

тогда используя сильную выпуклость-вогнусть задачи (1.1) получаем

$$2\left\langle \begin{pmatrix} x^* - x^k \\ y^* - y^k \end{pmatrix}, \begin{pmatrix} \nabla f(x_f^k) + \nabla_x F(x_f^k, y_f^k) \\ -\nabla_y F(x_f^k, y_f^k) \end{pmatrix} \right\rangle =$$

$$= 2\left\langle \begin{pmatrix} x^* - x_f^k \\ y^* - y_f^k \end{pmatrix}, \begin{pmatrix} \nabla f(x_f^k) + \nabla_x F(x_f^k, y_f^k) \\ -\nabla_y F(x_f^k, y_f^k) \end{pmatrix} \right\rangle +$$

$$+2\left\langle \begin{pmatrix} x_f^k - x^k \\ y_f^k - y^k \end{pmatrix}, \begin{pmatrix} \nabla f(x_f^k) + \nabla_x F(x_f^k, y_f^k) \\ -\nabla_y F(x_f^k, y_f^k) \end{pmatrix} \right\rangle \leq$$

$$\le 2\left\langle \begin{pmatrix} x^*-x_f^k \\ y^*-y_f^k \end{pmatrix}, \begin{pmatrix} \nabla f(x_f^k)+\nabla_x F(x_f^k,y_f^k) \\ -\nabla_y F(x_f^k,y_f^k) \end{pmatrix} - \begin{pmatrix} \nabla f(x^*)+\nabla_x F(x^*,y^*) \\ -\nabla_y F(x^*,y^*) \end{pmatrix} \right\rangle +$$

$$+\,2\left\langle \begin{pmatrix} x_f^k-x^k \\ y_f^k-y^k \end{pmatrix}, \begin{pmatrix} \nabla f(x_f^k)+\nabla_x F(x_f^k,y_f^k) \\ -\nabla_y F(x_f^k,y_f^k) \end{pmatrix} \right\rangle \le$$

$$\le -2\mu \left\| \begin{pmatrix} x^*-x_f^k \\ y^*-y_f^k \end{pmatrix} \right\|^2 + 2\left\langle \begin{pmatrix} x_f^k-x^k \\ y_f^k-y^k \end{pmatrix}, \begin{pmatrix} \nabla f(x_f^k)+\nabla_x F(x_f^k,y_f^k) \\ -\nabla_y F(x_f^k,y_f^k) \end{pmatrix} \right\rangle =$$

$$= -2\mu \left\| \begin{pmatrix} x^*-x_f^k \\ y^*-y_f^k \end{pmatrix} \right\|^2 + 2\theta\left\langle \theta^{-1}\begin{pmatrix} x_f^k-x^k \\ y_f^k-y^k \end{pmatrix}, \begin{pmatrix} \nabla f(x_f^k)+\nabla_x F(x_f^k,y_f^k) \\ -\nabla_y F(x_f^k,y_f^k) \end{pmatrix} \right\rangle =$$

$$= -2\mu \left\| \begin{pmatrix} x_f^k-x^* \\ y_f^k-y^* \end{pmatrix} \right\|^2 - \frac{1}{\theta}\left\| \begin{pmatrix} x_f^k-x^k \\ y_f^k-y^k \end{pmatrix} \right\|^2 - \theta \left\| \begin{pmatrix} \nabla f(x_f^k)+\nabla_x F(x_f^k,y_f^k) \\ -\nabla_y F(x_f^k,y_f^k) \end{pmatrix} \right\|^2 +$$

$$+\,\theta\left\| \theta^{-1}\begin{pmatrix} x_f^k-x^k \\ y_f^k-y^k \end{pmatrix} + \begin{pmatrix} \nabla f(x_f^k)+\nabla_x F(x_f^k,y_f^k) \\ -\nabla_y F(x_f^k,y_f^k) \end{pmatrix} \right\|^2$$

Так как функция $f(x)$ является L_f-гладкой, то получаем

$$2\left\langle \begin{pmatrix} x^*-x^k \\ y^*-y^k \end{pmatrix}, \begin{pmatrix} \nabla f(x_f^k)+\nabla_x F(x_f^k,y_f^k) \\ -\nabla_y F(x_f^k,y_f^k) \end{pmatrix} \right\rangle \le$$

$$\le -2\mu \left\| \begin{pmatrix} x_f^k-x^* \\ y_f^k-y^* \end{pmatrix} \right\|^2 - \frac{1}{\theta}\left\| \begin{pmatrix} x_f^k-x^k \\ y_f^k-y^k \end{pmatrix} \right\|^2 - \theta \left\| \begin{pmatrix} \nabla f(x_f^k)+\nabla_x F(x_f^k,y_f^k) \\ -\nabla_y F(x_f^k,y_f^k) \end{pmatrix} \right\|^2 +$$

$$+\theta\left\| B_\theta^k(x_f^k,y_f^k) + \nabla_{(x,y)} f(x_f^k) - \nabla_{(x,y)} f(x^k)\right\|^2 \le$$

$$\le -2\mu \left\| \begin{pmatrix} x_f^k-x^* \\ y_f^k-y^* \end{pmatrix} \right\|^2 - \frac{1}{\theta}\left\| \begin{pmatrix} x_f^k-x^k \\ y_f^k-y^k \end{pmatrix} \right\|^2 - \theta \left\| \begin{pmatrix} \nabla f(x_f^k)+\nabla_x F(x_f^k,y_f^k) \\ -\nabla_y F(x_f^k,y_f^k) \end{pmatrix} \right\|^2 +$$

$$+2\theta\left\| B_\theta^k(x_f^k,y_f^k)\right\|^2 + 2\theta L_f^2 \left\| \begin{pmatrix} x_f^k-x^k \\ y_f^k-y^k \end{pmatrix} \right\|^2 =$$

$$-2\mu \left\| \begin{pmatrix} x_f^k-x^* \\ y_f^k-y^* \end{pmatrix} \right\|^2 - \left(\frac{1}{\theta}-2\theta L_f^2\right)\left\| \begin{pmatrix} x_f^k-x^k \\ y_f^k-y^k \end{pmatrix} \right\|^2 + 2\theta\left\| B_\theta^k(x_f^k,y_f^k)\right\|^2 -$$

$$-\theta \left\| \begin{pmatrix} \nabla f(x_f^k)+\nabla_x F(x_f^k,y_f^k) \\ -\nabla_y F(x_f^k,y_f^k) \end{pmatrix} \right\|^2$$

Теперь подставляя изначально указанное значение $\theta = \frac{1}{2L_f}$, получим

$$2\left\langle \begin{pmatrix} x^* - x^k \\ y^* - y^k \end{pmatrix}, \begin{pmatrix} \nabla f(x_f^k) + \nabla_x F(x_f^k, y_f^k) \\ -\nabla_y F(x_f^k, y_f^k) \end{pmatrix} \right\rangle \le$$

$$\le -2\mu \left\| \begin{pmatrix} x_f^k - x^* \\ y_f^k - y^* \end{pmatrix} \right\|^2 - \frac{1}{2\theta} \left\| \begin{pmatrix} x_f^k - x^k \\ y_f^k - y^k \end{pmatrix} \right\|^2 + 2\theta \left\| B_\theta^k(x_f^k, y_f^k) \right\|^2 -$$

$$-\theta \left\| \begin{pmatrix} \nabla f(x_f^k) + \nabla_x F(x_f^k, y_f^k) \\ -\nabla_y F(x_f^k, y_f^k) \end{pmatrix} \right\|^2 =$$

$$= -2\mu \left\| \begin{pmatrix} x_f^k - x^* \\ y_f^k - y^* \end{pmatrix} \right\|^2 - \frac{1}{4\theta} \left\| \begin{pmatrix} x^k - \overline{x}_f^k \\ y^k - \overline{y}_f^k \end{pmatrix} \right\|^2 + 2\theta \left\| B_\theta^k(x_f^k, y_f^k) \right\|^2 -$$

$$-\theta \left\| \begin{pmatrix} \nabla f(x_f^k) + \nabla_x F(x_f^k, y_f^k) \\ -\nabla_y F(x_f^k, y_f^k) \end{pmatrix} \right\|^2 + \frac{1}{2\theta} \left\| \begin{pmatrix} x_f^k - \overline{x}_f^k \\ y_f^k - \overline{y}_f^k \end{pmatrix} \right\|^2$$

Рассмотрим следующее выражение

$$\left\langle B_\theta^k(x_1, y_1) - B_\theta^k(x_2, y_2);\ \begin{pmatrix} x_1 - x_2 \\ y_1 - y_2 \end{pmatrix} \right\rangle =$$

$$= \left\langle \begin{pmatrix} \nabla_x F(x_1, y_1) - \nabla_x F(x_2, y_2) + \frac{x_1 - x_2}{\theta} \\ -\nabla_y F(x_1, y_1) + \nabla_y F(x_2, y_2) + \frac{y_1 - y_2}{\theta} \end{pmatrix};\ \begin{pmatrix} x_1 - x_2 \\ y_1 - y_2 \end{pmatrix} \right\rangle \ge \frac{1}{\theta} \left\| \begin{pmatrix} x_1 - x_2 \\ y_1 - y_2 \end{pmatrix} \right\|^2$$

где последнее неравенство выполняется вследствие монотонности градиента для выпуклой-вогнутой седловой задачи. Также можно записать

$$\left\langle B_\theta^k(x_1, y_1) - B_\theta^k(x_2, y_2);\ \begin{pmatrix} x_1 - x_2 \\ y_1 - y_2 \end{pmatrix} \right\rangle \le \left\| B_\theta^k(x_1, y_1) - B_\theta^k(x_2, y_2) \right\| \cdot \left\| \begin{pmatrix} x_1 - x_2 \\ y_1 - y_2 \end{pmatrix} \right\|$$

$$\frac{1}{\theta} \left\| \begin{pmatrix} x_1 - x_2 \\ y_1 - y_2 \end{pmatrix} \right\| \le \left\| B_\theta^k(x_1, y_1) - B_\theta^k(x_2, y_2) \right\|$$

Решение подзадачи (1.7) является также решением уравнения $B_\theta^k(x, y) = 0$, тогда

$$\frac{1}{\theta^2} \left\| \begin{pmatrix} x_f^k - \overline{x}_f^k \\ y_f^k - \overline{y}_f^k \end{pmatrix} \right\|^2 \le \left\| B_\theta^k(x_f^k, y_f^k) - B_\theta^k(\overline{x}_f^k, \overline{y}_f^k) \right\|^2 = \left\| B_\theta^k(x_f^k, y_f^k) \right\|^2$$

Теперь используя полученное неравенство можем записать

$$2\left\langle \begin{pmatrix} x^* - x^k \\ y^* - y^k \end{pmatrix}, \begin{pmatrix} \nabla f(x_f^k) + \nabla_x F(x_f^k, y_f^k) \\ -\nabla_y F(x_f^k, y_f^k) \end{pmatrix} \right\rangle \leq$$

$$\leq -2\mu \left\| \begin{pmatrix} x_f^k - x^* \\ y_f^k - y^* \end{pmatrix} \right\|^2 - \frac{1}{4\theta} \left\| \begin{pmatrix} x^k - \overline{x}_f^k \\ y^k - \overline{y}_f^k \end{pmatrix} \right\|^2 + 2\theta \left\| B_\theta^k(x_f^k, y_f^k) \right\|^2 -$$

$$-\theta \left\| \begin{pmatrix} \nabla f(x_f^k) + \nabla_x F(x_f^k, y_f^k) \\ -\nabla_y F(x_f^k, y_f^k) \end{pmatrix} \right\|^2 + \frac{1}{2\theta} \left\| \begin{pmatrix} x_f^k - \overline{x}_f^k \\ y_f^k - \overline{y}_f^k \end{pmatrix} \right\|^2 \leq$$

$$\leq -2\mu \left\| \begin{pmatrix} x_f^k - x^* \\ y_f^k - y^* \end{pmatrix} \right\|^2 - \frac{1}{4\theta} \left\| \begin{pmatrix} x^k - \overline{x}_f^k \\ y^k - \overline{y}_f^k \end{pmatrix} \right\|^2 + 2\theta \left\| B_\theta^k(x_f^k, y_f^k) \right\|^2 -$$

$$-\theta \left\| \begin{pmatrix} \nabla f(x_f^k) + \nabla_x F(x_f^k, y_f^k) \\ -\nabla_y F(x_f^k, y_f^k) \end{pmatrix} \right\|^2 + \frac{\theta}{2} \left\| B_\theta^k(x_f^k, y_f^k) \right\|^2 \leq$$

$$\leq -2\mu \left\| \begin{pmatrix} x_f^k - x^* \\ y_f^k - y^* \end{pmatrix} \right\|^2 - \frac{1}{4\theta} \left\| \begin{pmatrix} x^k - \overline{x}_f^k \\ y^k - \overline{y}_f^k \end{pmatrix} \right\|^2 + 3\theta \left\| B_\theta^k(x_f^k, y_f^k) \right\|^2 -$$

$$-\theta \left\| \begin{pmatrix} \nabla f(x_f^k) + \nabla_x F(x_f^k, y_f^k) \\ -\nabla_y F(x_f^k, y_f^k) \end{pmatrix} \right\|^2 =$$

$$= -2\mu \left\| \begin{pmatrix} x_f^k - x^* \\ y_f^k - y^* \end{pmatrix} \right\|^2 - \theta \left\| \begin{pmatrix} \nabla f(x_f^k) + \nabla_x F(x_f^k, y_f^k) \\ -\nabla_y F(x_f^k, y_f^k) \end{pmatrix} \right\|^2 +$$

$$+3\theta \left(\left\| B_\theta^k(x_f^k, y_f^k) \right\|^2 - \frac{1}{12\theta^2} \left\| \begin{pmatrix} x^k - \overline{x}_f^k \\ y^k - \overline{y}_f^k \end{pmatrix} \right\|^2 \right) =$$

$$= -2\mu \left\| \begin{pmatrix} x_f^k - x^* \\ y_f^k - y^* \end{pmatrix} \right\|^2 - \theta \left\| \begin{pmatrix} \nabla f(x_f^k) + \nabla_x F(x_f^k, y_f^k) \\ -\nabla_y F(x_f^k, y_f^k) \end{pmatrix} \right\|^2 +$$

$$+3\theta \left(\left\| B_\theta^k(x_f^k, y_f^k) \right\|^2 - \frac{L_f^2}{3} \left\| \begin{pmatrix} x^k - \overline{x}_f^k \\ y^k - \overline{y}_f^k \end{pmatrix} \right\|^2 \right)$$

Лемма доказана. □

Лемма 2. Пусть для Алгоритма 2 в условиях задачи (1.1) при выборе параметров $\theta = \frac{1}{2L_f}$, $\eta = \min\left[\frac{1}{4\mu}, \frac{1}{4L_f}\right]$, $\alpha = 2\mu$ и при решении задачи (1.7) так, что выполняется

$$\left\|B_\theta^k(x_f^k, y_f^k)\right\|^2 \le \frac{L_f^2}{3}\left\|\begin{pmatrix} x^k - \overline{x}_f^k \\ y^k - \overline{y}_f^k \end{pmatrix}\right\|^2 \tag{A.1}$$

выполняется следующее неравенство

$$\left\|\begin{pmatrix} x^{k+1} - x^* \\ y^{k+1} - y^* \end{pmatrix}\right\|^2 \le (1 - 2\mu\eta)^k \left\|\begin{pmatrix} x^0 - x^* \\ y^0 - y^* \end{pmatrix}\right\|^2 \tag{A.2}$$

где $(\overline{x}_f^k, \overline{y}_f^k)$ - точное решение подзадачи на k-ой итерации Алгоритма 2.

Доказательство (приведено из Леммы 5 [6]): Используя выражения для обновления переменных оптимизации мы получаем:

$$\left\|\begin{pmatrix} x^{k+1} - x^* \\ y^{k+1} - y^* \end{pmatrix}\right\|^2 = \left\|\begin{pmatrix} x^k - x^* \\ y^k - y^* \end{pmatrix}\right\|^2 + 2\left\langle \begin{pmatrix} x^{k+1} - x^k \\ y^{k+1} - y^k \end{pmatrix}; \begin{pmatrix} x^k - x^* \\ y^k - y^* \end{pmatrix} \right\rangle +$$

$$+\left\|\begin{pmatrix} x^{k+1} - x^k \\ y^{k+1} - y^k \end{pmatrix}\right\|^2 = \left\|\begin{pmatrix} x^k - x^* \\ y^k - y^* \end{pmatrix}\right\|^2 + 2\eta\alpha\left\langle \begin{pmatrix} x_f^k - x^k \\ y_f^k - y^k \end{pmatrix}; \begin{pmatrix} x^k - x^* \\ y^k - y^* \end{pmatrix} \right\rangle -$$

$$-2\eta\left\langle \begin{pmatrix} \nabla f(x_f^k) + \nabla_x F(x_f^k, y_f^k) \\ -\nabla_y F(x_f^k, y_f^k) \end{pmatrix}; \begin{pmatrix} x^k - x^* \\ y^k - y^* \end{pmatrix} \right\rangle + \left\|\begin{pmatrix} x^{k+1} - x^k \\ y^{k+1} - y^k \end{pmatrix}\right\|^2 =$$

$$= \left\|\begin{pmatrix} x^k - x^* \\ y^k - y^* \end{pmatrix}\right\|^2 + \eta\alpha\left\|\begin{pmatrix} x_f^k - x^* \\ y_f^k - y^* \end{pmatrix}\right\|^2 - \eta\alpha\left\|\begin{pmatrix} x^k - x^* \\ y^k - y^* \end{pmatrix}\right\|^2 - \eta\alpha\left\|\begin{pmatrix} x_f^k - x^k \\ y_f^k - y^k \end{pmatrix}\right\|^2$$

$$+\left\|\begin{pmatrix} x^{k+1} - x^k \\ y^{k+1} - y^k \end{pmatrix}\right\|^2 - 2\eta\left\langle \begin{pmatrix} \nabla f(x_f^k) + \nabla_x F(x_f^k, y_f^k) \\ -\nabla_y F(x_f^k, y_f^k) \end{pmatrix}; \begin{pmatrix} x^k - x^* \\ y^k - y^* \end{pmatrix} \right\rangle$$

Используем ранее доказанную Лемму 1:

$$\left\|\begin{pmatrix} x^{k+1} - x^* \\ y^{k+1} - y^* \end{pmatrix}\right\|^2 \le (1 - \eta\alpha)\left\|\begin{pmatrix} x^k - x^* \\ y^k - y^* \end{pmatrix}\right\|^2 + \eta\alpha\left\|\begin{pmatrix} x_f^k - x^* \\ y_f^k - y^* \end{pmatrix}\right\|^2 +$$

$$+\left\|\begin{pmatrix} x^{k+1} - x^k \\ y^{k+1} - y^k \end{pmatrix}\right\|^2 - \eta\alpha\left\|\begin{pmatrix} x_f^k - x^k \\ y_f^k - y^k \end{pmatrix}\right\|^2 - 2\eta\mu\left\|\begin{pmatrix} x_f^k - x^* \\ y_f^k - y^* \end{pmatrix}\right\|^2 -$$

$$-\left\|\begin{pmatrix}\nabla f(x_f^k)+\nabla_x F(x_f^k,y_f^k)\\ -\nabla_y F(x_f^k,y_f^k)\end{pmatrix}\right\|^2 + 3\eta\theta\left(\|B_\theta^k(x_f^k,y_f^k)\|^2 - \frac{L_f^2}{3}\left\|\begin{pmatrix}x^k-\overline{x}_f^k\\ y^k-\overline{y}_f^k\end{pmatrix}\right\|^2\right) =$$

$$= (1-\eta\alpha)\left\|\begin{pmatrix}x^k-x^*\\ y^k-y^*\end{pmatrix}\right\|^2 + \left\|\eta\alpha\begin{pmatrix}x_f^k-x^*\\ y_f^k-y^*\end{pmatrix} - \eta\begin{pmatrix}\nabla f(x_f^k)+\nabla_x F(x_f^k,y_f^k)\\ -\nabla_y F(x_f^k,y_f^k)\end{pmatrix}\right\|^2 -$$

$$-\eta\alpha\left\|\begin{pmatrix}x_f^k-x^k\\ y_f^k-y^k\end{pmatrix}\right\|^2 - \eta(2\mu-\alpha)\left\|\begin{pmatrix}x_f^k-x^*\\ y_f^k-y^*\end{pmatrix}\right\|^2 - \eta\theta\left\|\begin{pmatrix}\nabla f(x_f^k)+\nabla_x F(x_f^k,y_f^k)\\ -\nabla_y F(x_f^k,y_f^k)\end{pmatrix}\right\|^2 +$$

$$+3\eta\theta\left(\|B_\theta^k(x_f^k,y_f^k)\|^2 - \frac{L_f^2}{3}\left\|\begin{pmatrix}x^k-\overline{x}_f^k\\ y^k-\overline{y}_f^k\end{pmatrix}\right\|^2\right) \leq$$

$$\leq (1-\eta\alpha)\left\|\begin{pmatrix}x^k-x^*\\ y^k-y^*\end{pmatrix}\right\|^2 - \eta\alpha(1-2\eta\alpha)\left\|\begin{pmatrix}x_f^k-x^k\\ y_f^k-y^k\end{pmatrix}\right\|^2 - \eta(2\mu-\alpha)\left\|\begin{pmatrix}x_f^k-x^*\\ y_f^k-y^*\end{pmatrix}\right\|^2 -$$

$$-\eta(\theta-2\eta)\left\|\begin{pmatrix}\nabla f(x_f^k)+\nabla_x F(x_f^k,y_f^k)\\ -\nabla_y F(x_f^k,y_f^k)\end{pmatrix}\right\|^2 +$$

$$+3\eta\theta\left(\|B_\theta^k(x_f^k,y_f^k)\|^2 - \frac{L_f^2}{3}\left\|\begin{pmatrix}x^k-\overline{x}_f^k\\ y^k-\overline{y}_f^k\end{pmatrix}\right\|^2\right) \Rightarrow$$

$$\Rightarrow \left\|\begin{pmatrix}x^{k+1}-x^*\\ y^{k+1}-y^*\end{pmatrix}\right\|^2 \leq (1-2\eta\mu)\left\|\begin{pmatrix}x^k-x^*\\ y^k-y^*\end{pmatrix}\right\|^2 +$$

$$+3\eta\theta\left(\|B_\theta^k(x_f^k,y_f^k)\|^2 - \frac{L_f^2}{3}\left\|\begin{pmatrix}x^k-\overline{x}_f^k\\ y^k-\overline{y}_f^k\end{pmatrix}\right\|^2\right)$$

Здесь последнее следствие получено подстановкой заданных в условии Леммы 2 параметров алгоритма. Так как в условии леммы также задана точность решения подзадачи, то получаем:

$$\left\|\begin{pmatrix}x^{k+1}-x^*\\ y^{k+1}-y^*\end{pmatrix}\right\|^2 \leq (1-2\eta\mu)\left\|\begin{pmatrix}x^k-x^*\\ y^k-y^*\end{pmatrix}\right\|^2$$

Лемма доказана. $\square$

Лемма 3. Условие остановки внутреннего метода (A.1) можно записать в терминах ε-точного решения (1.4), как

$$\left\| \begin{pmatrix} x_f^k - \overline{x}_f^k \\ y_f^k - \overline{y}_f^k \end{pmatrix} \right\|^2 \leq \varepsilon_k \tag{A.3}$$

где

$$\varepsilon_k \leq \frac{L_f^2}{3(2L_f + L_F)^2} \left\| \begin{pmatrix} x^k - \overline{x}_f^k \\ y^k - \overline{y}_f^k \end{pmatrix} \right\|^2 \tag{A.4}$$

Доказательство: Запишем векторную форму критерия остановки внутреннего метода:

$$\left\| \begin{pmatrix} \nabla f(x^k) + \nabla_x F(x_f^k, y_f^k) + \frac{1}{\theta}(x_f^k - x^k) \\ -\nabla_y F(x_f^k, y_f^k) + \frac{1}{\theta}(y_f^k - y^k) \end{pmatrix} \right\|^2 \leq \frac{L_f^2}{3} \left\| \begin{pmatrix} x^k - \overline{x}_f^k \\ y^k - \overline{y}_f^k \end{pmatrix} \right\|^2$$

Рассматриваемые алгоритмы гарантируют ε-точное решение, тогда предположим, что на k-ой итерации мы решаем вспомогательную задачу с точностью ε_k:

$$\left\| \begin{pmatrix} x_f^k - \overline{x}_f^k \\ y_f^k - \overline{y}_f^k \end{pmatrix} \right\|^2 \leq \varepsilon_k$$

Так как теорема сходимости требует $\theta = \frac{1}{2L_f}$, то

$$\left\| \begin{pmatrix} \nabla f(x^k) + \nabla_x F(x_f^k, y_f^k) + 2L_f(x_f^k - x^k) \\ -\nabla_y F(x_f^k, y_f^k) + 2L_f(y_f^k - y^k) \end{pmatrix} \right\|^2 \leq \frac{L_f^2}{3} \left\| \begin{pmatrix} x^k - \overline{x}_f^k \\ y^k - \overline{y}_f^k \end{pmatrix} \right\|^2$$

$$\left\| \begin{pmatrix} \nabla f(x^k) + \nabla_x F(x_f^k, y_f^k) + 2L_f(x_f^k - x^k) \\ -\nabla_y F(x_f^k, y_f^k) + 2L_f(y_f^k - y^k) \end{pmatrix} \right\|^2 =$$

$$= \left\| \begin{pmatrix} \nabla f(x^k) + \nabla_x F(x_f^k, y_f^k) - \nabla_x F(\overline{x}_f^k, \overline{y}_f^k) + \nabla_x F(\overline{x}_f^k, \overline{y}_f^k) \\ -\nabla_y F(x_f^k, y_f^k) + \nabla_y F(\overline{x}_f^k, \overline{y}_f^k)) - \nabla_y F(\overline{x}_f^k, \overline{y}_f^k - x^k)) \end{pmatrix} + \right.$$

$$\left. + \begin{pmatrix} 2L_f(x_f^k - \overline{x}_f^k + \overline{x}_f^k - x^k) \\ 2L_f(y_f^k - \overline{y}_f^k + \overline{y}_f^k - y^k) \end{pmatrix} \right\|^2 \leq$$

$$\leq \left\| \begin{pmatrix} \nabla f(x^k) + 2L_f(\overline{x}_f^k - x^k) + \nabla_x F(\overline{x}_f^k, \overline{y}_f^k) \\ 2L_f(\overline{y}_f^k - y^k) - \nabla_y F(\overline{x}_f^k, \overline{y}_f^k) \end{pmatrix} \right\|^2 + \left\| 2L_f \begin{pmatrix} x_f^k - \overline{x}_f^k \\ y_f^k - \overline{x}_f^k \end{pmatrix} \right\|^2 +$$

$$+\left\|\begin{pmatrix}\nabla_x F(x_f^k, y_f^k) - \nabla_x F(\overline{x}_f^k, \overline{y}_f^k)\\ -\nabla_y F(x_f^k, y_f^k) + \nabla_y F(\overline{x}_f^k, \overline{y}_f^k)\end{pmatrix}\right\|^2 +$$

$$+2\left\|\begin{pmatrix}\nabla f(x^k) + 2L_f(\overline{x}_f^k - x^k) + \nabla_x F(\overline{x}_f^k, \overline{y}_f^k)\\ 2L_f(\overline{y}_f^k - y^k) - \nabla_y F(\overline{x}_f^k, \overline{y}_f^k)\end{pmatrix}\right\| \cdot \left\|2L_f\begin{pmatrix}x_f^k - \overline{x}_f^k\\ y_f^k - \overline{y}_f^k\end{pmatrix}\right\| +$$

$$+2\left\|\begin{pmatrix}\nabla f(x^k) + 2L_f(\overline{x}_f^k - x^k) + \nabla_x F(\overline{x}_f^k, \overline{y}_f^k)\\ 2L_f(\overline{y}_f^k - y^k) - \nabla_y F(\overline{x}_f^k, \overline{y}_f^k)\end{pmatrix}\right\| \cdot \left\|\begin{pmatrix}\nabla_x F(x_f^k, y_f^k) - \nabla_x F(\overline{x}_f^k, \overline{y}_f^k)\\ -\nabla_y F(x_f^k, y_f^k) + \nabla_y F(\overline{x}_f^k, \overline{y}_f^k)\end{pmatrix}\right\|$$

$$+2\left\|2L_f\begin{pmatrix}x_f^k - \overline{x}_f^k\\ y_f^k - \overline{y}_f^k\end{pmatrix}\right\| \cdot \left\|\begin{pmatrix}\nabla_x F(x_f^k, y_f^k) - \nabla_x F(\overline{x}_f^k, \overline{y}_f^k)\\ -\nabla_y F(x_f^k, y_f^k) + \nabla_y F(\overline{x}_f^k, \overline{y}_f^k)\end{pmatrix}\right\| \leq$$

$$\leq \left\|\begin{pmatrix}\nabla f(x^k) + 2L_f(\overline{x}_f^k - x^k) + \nabla_x F(\overline{x}_f^k, \overline{y}_f^k)\\ 2L_f(\overline{y}_f^k - y^k) - \nabla_y F(\overline{x}_f^k, \overline{y}_f^k)\end{pmatrix}\right\|^2 + 4L_f^2\varepsilon_k + L_F^2\varepsilon_k +$$

$$+2\left\|\begin{pmatrix}\nabla f(x^k) + 2L_f(\overline{x}_f^k - x^k) + \nabla_x F(\overline{x}_f^k, \overline{y}_f^k)\\ 2L_f(\overline{y}_f^k - y^k) - \nabla_y F(\overline{x}_f^k, \overline{y}_f^k)\end{pmatrix}\right\| \cdot (2L_f\sqrt{\varepsilon_k} + L_F\sqrt{\varepsilon_k}) + 4L_f L_F \varepsilon_k =$$

$$= \left(\left\|\begin{pmatrix}\nabla f(x^k) + 2L_f(\overline{x}_f^k - x^k) + \nabla_x F(\overline{x}_f^k, \overline{y}_f^k)\\ 2L_f(\overline{y}_f^k - y^k) - \nabla_y F(\overline{x}_f^k, \overline{y}_f^k)\end{pmatrix}\right\| + (2L_f + L_F)\sqrt{\varepsilon_k}\right)^2 \leq$$

$$\leq \frac{L_f^2}{3}\left\|\begin{pmatrix}x^k - \overline{x}_f^k\\ y^k - \overline{y}_f^k\end{pmatrix}\right\|^2$$

$$\left\|\begin{pmatrix}\nabla f(x^k) + 2L_f(\overline{x}_f^k - x^k) + \nabla_x F(\overline{x}_f^k, \overline{y}_f^k)\\ 2L_f(\overline{y}_f^k - y^k) - \nabla_y F(\overline{x}_f^k, \overline{y}_f^k)\end{pmatrix}\right\| + (2L_f + L_F)\sqrt{\varepsilon_k} \leq$$

$$\leq \frac{L_f}{\sqrt{3}}\left\|\begin{pmatrix}x^k - \overline{x}_f^k\\ y^k - \overline{y}_f^k\end{pmatrix}\right\|$$

Отсюда можно получить оценку на ε_k:

$$(2L_f + L_F)\sqrt{\varepsilon_k} \leq \frac{L_f}{\sqrt{3}}\left\|\begin{pmatrix}x^k - \overline{x}_f^k\\ y^k - \overline{y}_f^k\end{pmatrix}\right\| - \left\|\begin{pmatrix}\nabla f(x^k) + 2L_f(\overline{x}_f^k - x^k) + \nabla_x F(\overline{x}_f^k, \overline{y}_f^k)\\ 2L_f(\overline{y}_f^k - y^k) - \nabla_y F(\overline{x}_f^k, \overline{y}_f^k)\end{pmatrix}\right\|$$

Используя условие оптимальности для вспомогательной задачи, которое записывается как

$$B_\theta^k(\overline{x}_f^k, \overline{y}_f^k) = \begin{pmatrix}\nabla f(x^k)\\ 0\end{pmatrix} + \begin{pmatrix}\nabla_x F(\overline{x}_f^k, \overline{y}_f^k)\\ -\nabla_y F(\overline{x}_f^k, \overline{y}_f^k)\end{pmatrix} + \frac{1}{\theta}\begin{pmatrix}\overline{x}_f^k - x^k\\ \overline{y}_f^k - y^k\end{pmatrix} = \begin{pmatrix}0\\ 0\end{pmatrix}$$

Получаем следующее соотношение:

$$\begin{pmatrix} \nabla_x F(\overline{x}_f^k, \overline{y}_f^k) \\ \nabla_y F(\overline{x}_f^k, \overline{y}_f^k) \end{pmatrix} = \begin{pmatrix} -\nabla f(x^k) - 2L_f(\overline{x}_f^k - x^k) \\ 2L_f(\overline{y}_f^k - y^k) \end{pmatrix}$$

Тогда уточним оценку на ε_k:

$$(2L_f + L_F)\sqrt{\varepsilon_k} \leq \frac{L_f}{\sqrt{3}} \left\| \begin{pmatrix} x^k - \overline{x}_f^k \\ y^k - \overline{y}_f^k \end{pmatrix} \right\|$$

Итоговая оценка ε_k:

$$\varepsilon_k \leq \frac{L_f^2}{3(2L_f + L_F)^2} \left\| \begin{pmatrix} x^k - \overline{x}_f^k \\ y^k - \overline{y}_f^k \end{pmatrix} \right\|^2 \tag{A.5}$$

Лемма доказана. $\square$