Московский физико-технический институт
(национальный исследовательский университет)
Физтех-школа прикладной математики и информатики
Кафедра интеллектуальных систем

Крейнин Матвей Вадимович

Методы предобуславливания с затуханием весов

03.03.01 — Прикладные математика и физика

Выпускная квалификационная работа бакалавра

Научный руководитель:
к.ф.-м.н. А. Н. Безносиков

Москва — 2024

Аннотация

Исследуется задача минимизации целевой функции потерь. Рассматривается проблема оптимизации целевой функции потерь градиентными методами первого порядка. Исследуется сходимость методов градиентной оптимизации с предобуславливанием, использующих регуляризацию с затуханием весов. Специально рассматриваются популярные методы оптимизации из данного класса методов такие как AdamW и OASIS. Исследуются различные альтернативы этим методам с целью изучения их скорости сходимости и точности, показываемой моделью. Предлагается новый способ добавления регуляризации в метод оптимизации Adam. Доказывается теорема о скорости сходимости данных методов при различных допущениях на функцию потерь и показывается сходимость к исходной функции потерь. Проводятся вычислителные эксперименты с различными эталонными наборами данных, моделями и проводится анализ гиперпараметров, чтобы сравнить их на реальных задачах.

Содержание

1 Введение

Огромная часть машинного обучения основана на решении задачи оптимизации без ограничений

$$\min_{w \in \mathbb{R}^d} f(w). \tag{1}$$

Задачи вида (1) охватывают множество приложений, включая минимизацию эмпирического риска [1], глубокое обучение [2], и задачи обучения с учителем [3] такие, как наименьшие квадраты с регуляризацией [4] или логистическая регрессия [5].

Классический метод решения задачи оптимизации (1) это градиентный спуск.

$$w_{t+1} = w_t - \eta \nabla f(w_t), \tag{2}$$

Задача минимизации (1) может быть трудноразрешимой особенно, когда размер выборки крайне велик или размерность задачи велика.

В таких случаях подсчет полного градиента на каждой итерации в градиентном спуске становится очень дорогим в плане времени или вычислительных ресурсов, которые нужны для этого, особенно учитывая, что градиентному спуску часто требуется большое количество итераций для сходимости. В современном машинном обучении, особенно с появлением глубокого обучения, растет интерес к решению все более больших и сложных задач. Популярным решением для таких проблем стал стохастический градиентный спуск [6].

С течением времени методы оптимизации постоянно совершенствовались и развивались, становясь все более сложными и запутанными. Одним из важнейших аспектов в оптимизации является правильный подбор размера шага в ходе итерационного процесса. Адаптивные методы с градиентным масштабированием динамически регулируют этот размер шага на основе информации о градиенте. Такое адаптивное поведение, применяемое к каждой переменной, улучшает процесс оптимизации, эффективно перемещаясь по сложным ландшафтам и обеспечивая оптимальный прогресс для каждой переменной [7]. В частности, эти методы приобрели значительную популярность в области машинного обучения, где преобладают высокоразмерные задачи [8, 9].

Более подробно под методами с масштабированным градиентом понимаются техники, предполагающие предобуславливание градиента задачи по определенной матрице D_t, что позволяет градиенту учитывать геометрию задачи. В общем случае шаг алгоритмов с предуславливанием может быть выражен как следующая модернизация шага (2):

$$w_{t+1} = w_t - \eta \cdot D_t^{-1} g_t, \tag{3}$$

где g_t - несмещенный стохастический градиент.

Идея использования матрицы шкалирования отсылает нас к методу Ньютона, где $D_t = \nabla^2 f(w)$. Однако вычисление и обращение гессиана сопряжено со значительными трудностями, что приводит нас к необходимости использования определенных эвристик в качестве замены матрицы D_t. Примерами таких эвристических методов являются Adagrad [10], Adam [11], RMSProp, OASIS [12] и так далее, где стратегии вычислений для D_t не требуют оценки гессиана. Например, в Adagrad предусловие представлено в виде:

$$D_t = \text{diag}\left\{\sqrt{\sum_{t'=0}^{t} g_{t'} \odot g_{t'}}\right\},$$

где $\odot$ - Адамарово произведение. На самом деле этот подход использует только стохастические градиенты.

RMSProp и Adam используют похожие идеи:

$$D_t^2 = \beta D_{t-1}^2 + (1-\beta)\text{diag}\left\{g_t \odot g_t\right\}$$

где $\beta \in (0, 1)$ представляет собой степень учета предыдущих итераций [11].

В OASIS используется другой подход:

$$D_t = \text{diag}\left\{z_k \odot \nabla^2 f(w_t) z_k\right\},$$

где z_k - случайный вектор из распределения Рандамахера, т.е. каждый элемент вектора $z_k^i \in \{-1, 1\}$ с вероятностью $\frac{1}{2}$ [12]. На первый взгляд кажется, что используется матрица гессиана, но на самом деле она аппроксимируется через дифференцирование скалярной функции.

Несмотря на преимущества методов предобусловливания, они склонны к переобучению, таким образом, возникает необходимость в их совместном применении с регуляризацией. Этот подход широко применяется для решения различных задач машинного обучения, включая классификацию изображений [13], распознавание речи [14] и обработку естественного языка [15], и показал свою эффективность в улучшении обобщающей способности нейронных сетей [16].

С регуляризацией задача (1) переформулируется как

$$\min_{w \in \mathbb{R}^d} F(w) := f(w) + r(w), \tag{4}$$

где r - функция регуляризации.

В методах с предуславливанием есть несколько способов добавления регуляризации. Можно добавить регуляризатор r в подсчет g_t, и тогда он будет учитываться при вычислении D_t. Этот способ равносилен рассмотрению оптимизационной задачи (4). Или же мы можем добавить регуляризатор только на последнем шаге, уменьшая норму w [17]. Такой способ регуляризации называется затуханием весов и, как ни странно, оказывается более эффективным в практических задачах. Существует и другой способ рассмотрения регуляризатора, который будет рассмотрен далее в статье.

Несмотря на свою практическую эффективность, методы, использующие затухание весов, относительно мало изучены с точки зрения теории сходимости методов. В связи с этим возникает ряд исследовательских вопросов:

- *Сходятся ли с теоретической точки зрения методы с предобуславливанием и затуханием весов?*
- *Если сходятся, то какова скорость их сходимости?*
- *К какой задаче они сходятся?*

2 Обзор литературы

Стохастические методы имеют обширный анализ их сходимости [18, 19, 20], в то время как методы, включающие предобуславливаниям, являются относительно новыми и неизученными. В одной из первых работ по предобуславливанию [10] авторы провели тщательный анализ теории сходимости Adagrad. Однако в более поздних работах, например, обсуждающих RMSProp [21] или Adam [11], теоретическим аспектам уделяется мало внимания, либо существующая теория содержит неточности в доказательстве.

Со временем ошибки были исправлены, что привело к разработке надежных теорий сходимости для методов с предобуславливанием [22, 23]. В другом исследовании Лощилов и Хуттер [17] исследовали свойства алгоритмов Adam и AdamW в терминах гиперпараметров, а также изучили методы рестартов. Чжан и др.[24] исследовали механизм заглядывания в будущее в Adam. В [25] исследователи из Nvidia предложили новый способ добавления регуляризации в алгоритм Adam, который на их экспериментах дал прирост в качестве обучения. Совсем недавно были созданы теории сходимости для современных методов, таких как OASIS [12, 26]. Кроме того, появилась теория, рассматривающая изменяющиеся во времени матрицы предобуславливания [27]. Тем не менее, многие вопросы в этой области остаются без ответа. Некоторые из них сформулированы в конце предыдущего параграфа и рассматриваются в нашей статье.

3 Основная часть

3.1 Обозначения

- Мы использкем x^i, где x это вектор и $i \in \overline{1, d}$ обозначает i-ю компоненту d-мерного вектора x.

- Для любых $x, y \in \mathbb{R}^d$ скалярное произведение обозначается, как $\langle x, y \rangle := \sum_{i=1}^{d} x^i y^i$.

- L_f – константа липшица функции f, то есть $\forall x, y \in \mathbb{R}^d \to f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L_f}{2} \|x - y\|^2$

- $||x|| := \sqrt{\langle x, x \rangle}$, где $x \in \mathbb{R}^d$ это l_2 норма вектора x.

- $||x||_A^2 := x^T A x$, где $x \in \mathbb{R}^d, A \in \mathbb{R}^{d \times d}$

- Для матрицы $A \in \mathbb{R}^{d \times d}$, A^{-1} – обратная матрица.

- Мы используем $A \preccurlyeq B$ для двух матриц $A, B \in \mathbb{R}^{d \times d}$, чтобы обозначить что $x^T A x \leq x^T B x$ для любых $x \in \mathbb{R}^d$.

- $\mathrm{diag}\left\{\beta_1 \ldots, \beta_d\right\}$ – диагональная матрица, состоящая из элементов: $\beta_1, ..., \beta_d \in \mathbb{R}$.

3.2 Затухание весов

Как было сказано выше, в методах с предобуславливанием существует несколько техник добавления регуляризации в оптимзируемую функцию. Мы рассмотрим три различных подхода, которые проиллюстрированы в Алгоритм 1 с помощью различных цветов (каждый отдельный цвет это отдельный алгоритм).

Algorithm 1 Различные способы использования предобуславливания с регуляризацией

Require: η — шаг обучения, f — оптимзируемая функция

while w не сойдется **do**

$t = t + 1$

$g_t \leftarrow$ стохастический градиент f

$g_t \leftarrow g_t + \nabla r(w_t)$ обычная регуляризация

$D_t \leftarrow$ матрица предобуславливания с помощью g_t

$w_t \leftarrow w_{t-1} - \eta \cdot D_t^{-1} g_t$ обычная регуляризация,

$w_t \leftarrow w_{t-1} - \eta \cdot D_t^{-1} (g_t + \nabla r(w_t))$ масштабированное затухание весов,

$w_t \leftarrow w_{t-1} - \eta \cdot D_t^{-1} g_t - \eta \cdot \nabla r(w_t)$ затухание весов,

end while

Если говорить более конкретно, то первая техника регуляризации, показанная в синим, заключается в простом добавлении регуляризационного члена к оптизируемой функции. Этот регуляризатор включается в стохастический градиент и учитывается при вычислении D_t. По сути, этот подход предполагает применение базового метода оптимизации с предобусловляиванием к задаче (4). Вторая техника регуляризации, показанная на рисунке оранжевым, является новым подходом. Хотя член регуляризатора не влияет на вычисление матрицы предобуславливания D_t, он добавляется перед применением D_t. Это означает, что скорость обучения принимается одинаковой для градиента и регуляризатора. Последний рассматриваемый нами подход к регуляризации известен как затухание веса, в алгоритме он подсвечивается цветом красным в алгоритме 1. Как и во втором методе, матрица D_t вычисляется без использования регуляризатора, а в этом методе регуляризатор включен на шаге обновления весов, что позволяет изьежать влияния регуляризации на матрицу предобуславливания.

Важно учитывать влияние регуляризации при разработке алгоритмов оптимизации, и я надеюсь, что моё исследование окажется полезным для исследователей в этой области.

3.3 Скорость сходимости методов с предобуславливанием и затуханием весов

Давайте попробуем оценить скорость сходимости методов с предобуславливанием и затуханием весов.

Хотя шаг оптимизации весов модели может показаться простым, он может быть рассмотрен с другой стороны. Давайте вынесем матрицу D_t^{-1} за скобки, что даёт нам следующий шаг:

$$w_{t+1} = w_t - \eta \cdot D_t^{-1}(\nabla f(w_t) + D_t \nabla r(w_t)). \tag{5}$$

Это подталкивает нас к тому, чтобы вести новую функцию $\tilde{r}$, такую, что $\nabla \tilde{r}_t(w) = D_t \nabla r(w)$ и новую целевую функцию

$$\tilde{F}_t(w) := f(w) + \tilde{r}_t(w), \tag{6}$$

, где новая целевая функция $\widetilde{F}_t$ меняется на каждом оптимизационном шаге, так как D_t тоже обновляется на каждом оптимизационном шаге.

Новый адаптивный регуляризатор $\widetilde{r}_t$ в общем случае к сожалению не существует. Поэтому мы наложим ограничения на начальный регуляризатор и структуру предобусловливателя, которые будут оформлены в виде следующих предположений на функцию регуляризации.

Предположение 1. *(Структура регулязитора) Регуляризатор r сепарабелен, то есть он может быть представлен в следующем виде:*

$$r(w) = \sum_{i=1}^{d} r_i(w^i),$$

где $r_i(x) \geq 0$ для $i \in \overline{1,d}$ и $x \in \mathbb{R}$.

Предположение 2. *(Структура матрицы предобуславливания) Матрица предобуславливания D_t может быть представлена в следующем виде:*

$$D_t = diag\left\{d_t^1 \ldots, d_t^d\right\}.$$

Хотя эти предположения являются достаточно сильными, но они выполняются для упомянутых ранее методов с предобуславливанием и затуханием весов, также это верно для таких популярных функций регуляризации как регуляризация Тиханова и LASSO регуляризация. Скорость сходимости обычно исчисляется количеством итераций, которые необходимы для достижения определенного уровня погрешности. Чтобы получить оценки количества итераций, необходимых для сходимости к заданной ошибке, мы должны наложить определенные предположения на оптимизируемую функцию потерь. На протяжении всего последующего анализа я предполагаю, что $f : \mathbb{R}^d \to \mathbb{R}$ является $L-$ гладким и дважды дифференцируемым.

Предположение 3. *(L-гладкость)*

- *Градиент функции f является L_f-гладким, то есть существует такая константа $L_f > 0$ такая, что $\forall x, y \in \mathbb{R}^d$,*

$$f(x) \leq f(y) + \langle \nabla f(y), x - y\rangle + \frac{L_f}{2}\|x - y\|^2.$$

- *Градиент функции r является L_r-гладким, то есть существует такая константа $L_r > 0$ такая, что $\forall x, y \in \mathbb{R}^d$,*

$$r(x) \leq r(y) + \langle \nabla r(y), x - y\rangle + \frac{L_r}{2}||x - y||^2.$$

Для того чтобы работать в невыпуклом случае, необходимо ввести ограничение на значения функции регуляризации, это описано в 4.

Предположение 4. *(Ограниченность регуляризатора) Регуляризатор ограничен, то есть существует константа $\Omega > 0$ такая, что $\forall w \in \mathbb{R}^d$*

$$|r(w)| \leq \Omega.$$

Мы используем обычное ограничение на матрицу предобуславливания, которое сформулировано в предположении 5.

Предположение 5. *(Ограниченность предобуславливателя) Существуют константы* $\alpha, \Gamma \in \mathbb{R} : 0 < \alpha < \Gamma$ *такие, что*

$$\alpha I \preccurlyeq D_t \preccurlyeq \Gamma I \Leftrightarrow \frac{I}{\Gamma} \preccurlyeq D_t^{-1} \preccurlyeq \frac{I}{\alpha}.$$

Это было доказано в [27], что это предположение справедливо для всех современных и популярных алгоритмов с предобуславливанием, таких как Adam, Adagrad, OASIS.

В нашем анализе мы рассматриваем два способа обновления матрицы предобуславливания. В первом методе матрица обновляется через квадраты:

$$(D_{t+1})^2 = \beta(D_t)^2 + (1-\beta)(H_t)^2, \tag{7}$$

, где H_t - матрица, содержащая новую информацию, а β $[0,1]$ - параметр импульса. Этот подход используется в Adam, а также в более старых методах, таких как RMSProp и AdaHessian. Второй способ является более современным и предполагает использование первых степеней матриц, сохраняя форму преобразования

$$D_{t+1} = \beta D_t + (1-\beta)H_t, \tag{8}$$

Этот подход используется в OASIS, недавно придуманным методом. В обеих случаях параметр импульса β обычно подбирается близким к 1, что означает, что D_t незначительно меняется в ходе обучения, что может быть формально сформулировано в лемме 1.

Выполнение предположения 5 имеет решающее значение для сходимости и теоретического анализа, и поэтому в алгоритмах часто используется метод императивного выбора для нижней границы матрицы D_t

$$\hat{D}_{t+1}^{ii} = \max\{\alpha, D_t^{ii}\}. \tag{9}$$

Лемма 1. *(Эволюция D_t, Безносиков) Предположим, что для начальной матрицы D_0 выполнены предположения 2 и 5, H_t диагональна с максимальным значением меньше или равным Γ на каждом временном шаге t, и D_t*

эволюционирует в соответствии с (7), (9) *или* (8), (9), *тогда справедливы следующие утверждения:*

1. *2 и 5 выполняется для* $\hat{D}_t$ *для всех* t;
2. $||\hat{D}_{t+1} - \hat{D}_t||_\infty \le \frac{(1-\beta)\Gamma^2}{2\alpha}$ *for* (7);
3. $||\hat{D}_{t+1} - \hat{D}_t||_\infty \le 2(1-\beta)\Gamma$ *for* (8).

Эта лемма доказана в **??**, где мы опираемся на [27].

Чтобы проводить стохастический анализ, мы должны включить ограничения на стохастический градиент функции. Это формализуется в следующем предположении

Предположение 6. *(Ожидания)* g_t *являются несмещенными и имеют ограниченную вариацию на любом шаге, то есть*

$$\mathbb{E}\left[g_t\right] = \nabla f(w_t), \mathbb{E}\left[||g_t - \nabla f||^2\right] \le \sigma^2. \tag{10}$$

Чтобы получить дополнительные оценки на сходимость методов с предобуславливанием и затуханием весов мы накладываем сильную выпуклость 7 на целевую функцию потерь.

Предположение 7. *(Сильная выпуклость) Существует* μ_f *такая, что* $\forall x, y \in \mathbb{R}^d$ *выполняется:*

$$f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu_f}{2}||x - y||_2^2$$

С помощью предположений 1 и 2 мы можем доказать существование $\widetilde{r}$ и, следовательно, $\widetilde{F}$. Мы оформим это в лемме 2. Мы показываем только существование, но не единственность функции, но в наших оценках $\widetilde{F}$ может быть найдена до константы.

Лемма 2. *(Существование* $\widetilde{r}$*) Предполагая, что 1, 2 выполняются, функция* $\widetilde{r}$ *существует и имеет следующую форму:*

$$\widetilde{r}_t(w) = \sum_{i=1}^{d} d_t^i r_i(w_i)$$

Используя введенное предположение 3, мы можем гарантировать гладкость для $\widetilde{r}$ и оценить его константу Липшица, что формально сформулировано и доказано в лемме 3.

Лемма 3. *(L-гладкость $\widetilde{r}$) Предполагая, что 1, 2, 3, 5 выполняются, градиент $\widetilde{r}$ является $L_{\tilde{r}}$-непрерывным, то есть существует константа $L_{\tilde{r}} > 0$ такая, что $\forall x, y \in \mathbb{R}^d$,*

$$\widetilde{r}_t(x) \leq \widetilde{r}_t(y) + \langle \nabla \widetilde{r}_t(y), x - y \rangle + \frac{L_{\tilde{r}}}{2} ||x - y||^2,$$

и $L_{\tilde{r}} = \Gamma L_r$.

Используя введенные предположения, мы доказали сходимость методов с предобусловливанием и затуханием в общем виде. Наши результаты оформлены в Теорему 1 и Теорему 2. Доказательства теорем можно найти в Приложении A.2.

Теорема 1. *Предполагая, что 1, 2, 3, 4, 5 выполняются, положим ошибку $\varepsilon > 0$ и шаг обучения удовлетворяют условию:*

$$\eta < \frac{2\alpha}{L_f + \Gamma L_r},$$

где L_f, L_r - константа Липшица функций f и r. Пусть существует начальная матрица предобуславливания, которая обновляется в соответствии с условиями леммы 1. Тогда количество итераций, выполняемых алгоритмами с предусловием и убывающим весом, начиная с начальной точки $w_0 \in \mathbb{R}^d$ с $\Delta_0 = \tilde{F}_0(w_0) - f^*$, где $\widetilde{F}_t$ определено в (6) и f^* решением задачи (1), необходимое для ε-приближения нормы градиента к 0, может быть ограничено количеством шагов

$$T = \mathcal{O}\left(\frac{\Delta_0 \Gamma}{(\eta - \frac{\tilde{L}\eta^2}{2\alpha})\left(\varepsilon - \frac{\delta\Gamma}{\eta - \frac{\tilde{L}\eta^2}{2\alpha}}\right)}\right),$$

где $\widetilde{L} = L_f + \Gamma L_r$ и δ может выбрано сколь угодно малым с помощью выбора гиперпараметров α, β, Γ

1. $\delta = \frac{(1-\beta)\Gamma^2}{2\alpha}$ *for* (7);

2. $\delta = 2(1-\beta)\Gamma$ *for* (8).

Теорема 2. *Преполагая, что 1, 2, 3, 4, 5, 7 выполняются, положим ошибку* $\varepsilon > 0$ *и шаг обучения удовлетворяют условию:*

$$\eta < \eta_{min} = \min\left\{\frac{2L_f\Omega_0^2}{\alpha\beta^2}; \frac{\alpha}{4L_f}; \frac{8\mu_f L_f^2\Omega_0^4}{\alpha^2\beta^4} + \frac{L_f\Omega_0^2}{\alpha\beta^2}\right\},$$

гиперпараметры удовлетворяют условиям: $\lambda < \frac{\alpha\beta^2}{8L_f\Omega_0^2}$, $\beta \geq 1 - \frac{\eta(\mu_f+\lambda)\alpha}{2\Gamma^2}$. *Получаем оценку на необходимое количество шагов для сходимости алгоритма к заданной точности:*

$$T = \mathcal{O}\left(\log\left(\frac{R_0^2 + \frac{8\lambda\Omega_0^2\Gamma^2}{\alpha^2(\mu_f+\lambda)}\sigma_0^2}{\varepsilon}\right)\frac{4}{\eta_{min}(\mu_f+\lambda)\cdot\min\left\{1; \frac{2\alpha}{\Gamma^2}\right\}}\right)$$

Эти теоремы устанавливают сходимость методов с предобуславливанием и затуханием весов различных предположениях, а также определяют необходимое количество итераций для заданной точности. Для наших задач простой факт сходимости этих методов имеет огромное значение.

Однако характеристики решения $\widetilde{w}^*$ задачи

$$\min_{w\in\mathbb{R}^d} \tilde{F}(w) = f(w) + \tilde{r}(w), \tag{11}$$

к которым сходится этот метод, требуют более глубокого исследования, которое будет рассмотрено в следующем разделе.

3.4 Решение методов с предобуславливанием и затуханием весов

В предыдущем подразделе мы доказали сходимость методов с предобуславливанием, однако выше мы указали, что методы с затуханием весов сходятся к исходному решению задачи оптимизации (4) w^*, а к исходному решению $\widetilde{w}^*$ задачи (1), это достигается за счет следующего. Мы получили новую целевую функцию потерь , в которой величина регуляризации динамически изменяется от шага к шагу, на основании матрицы предобуславливания, учитывая, что матрица составляется на основе стохастических градиентов, полученных в ходе обновления весов модели, получается, что регуляризация получается тем больше, чем больше градиент по данному весу модели, и наоборот тем меньше, чем меньше градиент по весу модели. То есть регуляризация не штрафует веса модели, где градиент вышел на значения близкие к нулю. То есть мы стараемся выйти делать больший шаг там, где стохастический градиент не приблизился к каким-то околнулевым значениям, то есть пока мы не оказались в окрестности какого-то экстремума. За счёт этого получается более разнообразная траектория обновления весов модели, которая позволяет нам получать лучшую сходимость на практике. Эти рассуждения подтверждаются экспериментами, подробно описанными в разделе 3.5.

Оценим разницу между решениями задач (4) и (6). Это ограничение основано на предположениях (3) и свойствах матрицы D_t.

Лемма 4. *(Lower bound) Предполагая, что 1, 2 и 3 выполняются, также предполагая, что задачи* (6) *и* (4) *имеют соответствующие решения* $\widetilde{w}^*$ *и* w^**, тогда разница между решениями может быть ограничена снизу:*

$$\|\widetilde{w}^* - w^*\| L_F \geq \|\nabla r(\widetilde{w}^*)(I - D_t)\|.$$

Следовательно, можно заметить, что использование регуляризации весов не в прямом подсчете градиента, которое влечет и учет их в матрице предобуславливания приводит к сходимости к решению исходной задачи, в то время, как прямое использование функции регуляризации для подсчета стохастического градиента приводит нас к альтернативному решению. Расхождение между этими решениями зависит от нормы разности между D_t и матрицей тождества ($||D_t - I||$). В результате анализ распределения элементов $D := \lim\limits_{t\to\infty} D_t$ может дать представление о сходимости метода с затуханием весов.

3.5 Эксперименты

Мы рассмотрим два алгоритма OASIS [12] и Adam [11], а также их вариации. Их основное отличие заключается в вычислении матрицы предобуславливания. В Adam это диагональная матрица, состоящая из квадратов производных, в OASIS - стохастический гессиан, который вычисляется через случайную величину из распределения Рандемахера. Я показываю три варианта регуляризации для Adam и OASIS в Алгоритме 2 и Алгоритме 3 соответственно.

Algorithm 2 Различные способы добавления регуляризации для Adam

---

Require: $\eta, \beta_1, \beta_2, \epsilon, f, r$

while θ не сойдется **do**

$t = t + 1$

$g_t = \nabla f(w_{t-1}) + \nabla r(w_{t-1})$ AdamL2

$m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$

$v_t = \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$

$\hat{m}_t = \frac{m_t}{1-\beta_1^t} + \nabla r(w_{t-1})$ AdamWH

$\hat{v}_t = \frac{v_t}{1-\beta_2^t}$

$w_t = w_{t-1} - \eta \cdot \frac{\hat{m}_t}{\sqrt{v_t}+\epsilon} - \eta \nabla r(w_{t-1})$ AdamW

end while

Algorithm 3 Различные способы добавления регуляризации для OASIS

Require: $w_0, \eta_0, D_0, \theta_0 = +\infty$

$w_1 = w_0 - \eta \hat{D_0}^{-1} \nabla f(w_0)$

for $k = 1, 2, ...$ **do**

$g_k = \nabla f(w_k) + \nabla r(w_{t-1})$ OASISL2

$D_k = \beta D_{k-1} + (1 - \beta_2) \cdot diag\left(z_k \odot \nabla^2 \left(f(w_k) + r(w_k)\right) z_k\right)$ OASISWH

$(\hat{D}_k)_{ii} = max\{|D_k|_{i,i}; \alpha\}, \forall i = \overline{1, d}$

$\eta_k = min\{\sqrt{1 + \theta_{k-1}} \cdot \eta_{k-1}; \frac{||w_k - w_{k-1}||_{\hat{D_k}}}{2||\nabla f(w_k) - \nabla f(w_{k-1})||^*_{\hat{D_k}}}\}$

$w_{k+1} = w_k - \eta_k g_k D_k^{-1} - \eta \nabla r(w_{t-1})$ OASISW

$\theta_k = \frac{\eta_k}{\eta_{k-1}}$

end for

В этом разделе приводится численные эксперименты для вышеупомянутых методов оптимизации. Эксперименты проводились на процессоре x86 и графическим ускорителем NVIDIA GeForce RTX 3090, эскперименты были воспроизведены на 8-ми ядерном процессоре на архитектуре ARM-64.

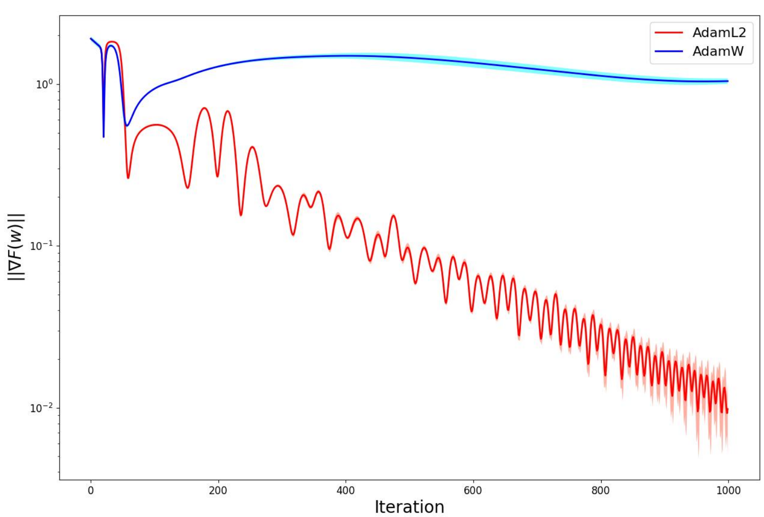

Рис. 1: Adam и AdamW по классическому критерию

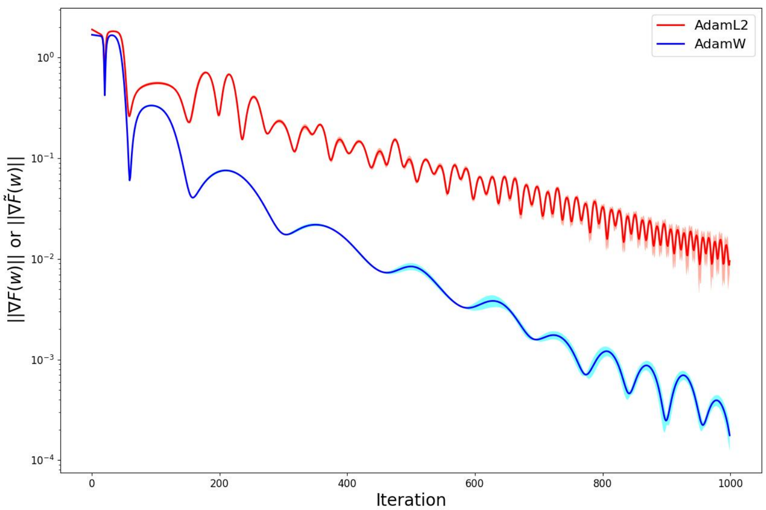

Рис. 2: Adam и AdamW с модифицированным критерием

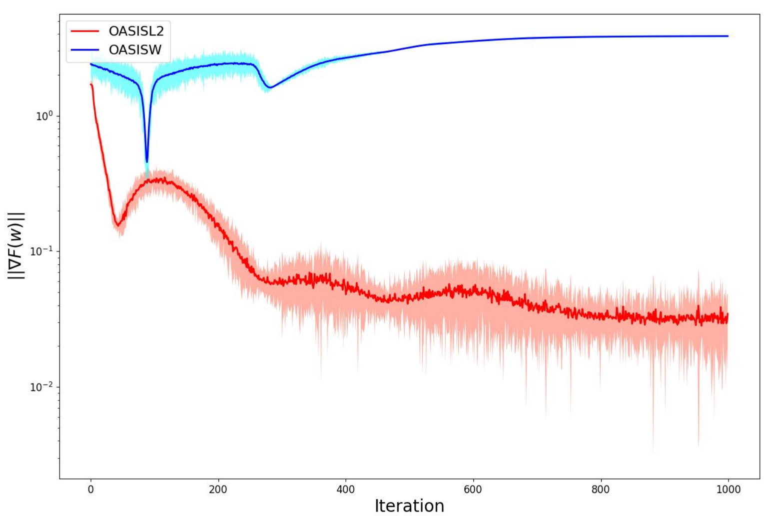

Рис. 3: OASIS и OASISW по классическому критерию

Нужно пояснить данные графики, в первой теореме мы приводим оценку сходимости необходимого количества шагов для нормы градиента измененной функции $\tilde{F}$. На графиках 1 и 3 приведена норма градиента изначальной функции потерь от итерации, как видно из графиков норма изначального градиента для метода с затуханием весов не убывает с итерациями, на графиках 2 и 4 построены графики для нормы градиента модифицированной функции потерь

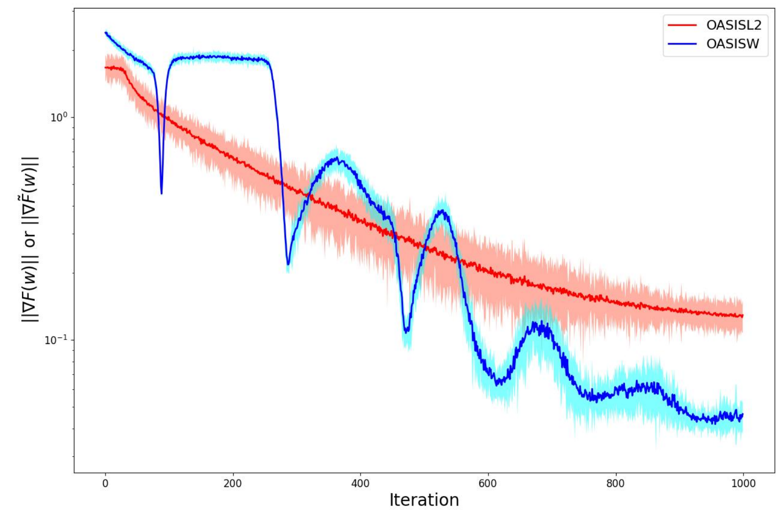

Рис. 4: OASIS и OASISW с модифицированным критерием

$\tilde{F}$. Различие сходимости в методах Adam и OASIS иллюстрирует выбранный критерий сходимости в первой теореме, это подтверждает, что методы с предобуславливанием и затуханием весов оптимизируют не функцию потерь F, а $\tilde{F}$. Именно градиент $\tilde{F}$ убывает с итерациями в то время, как градиент F остается неизменным.

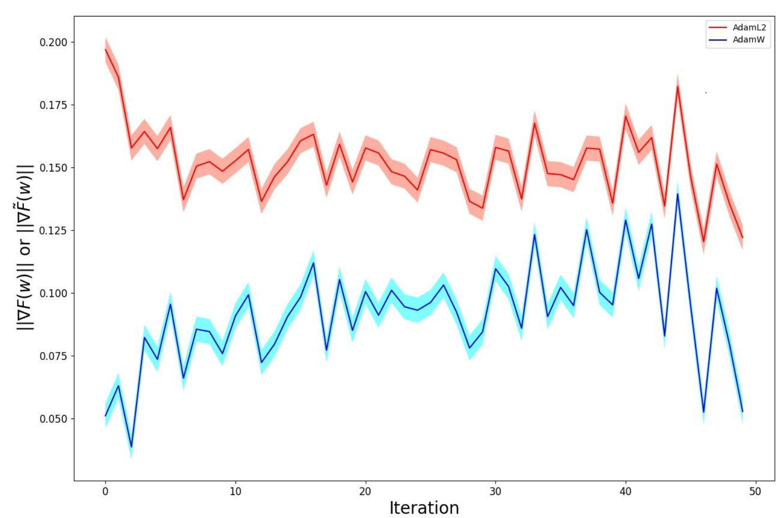

Рис. 5: Adam и AdamW по классическому критерию в слое нейронной сети

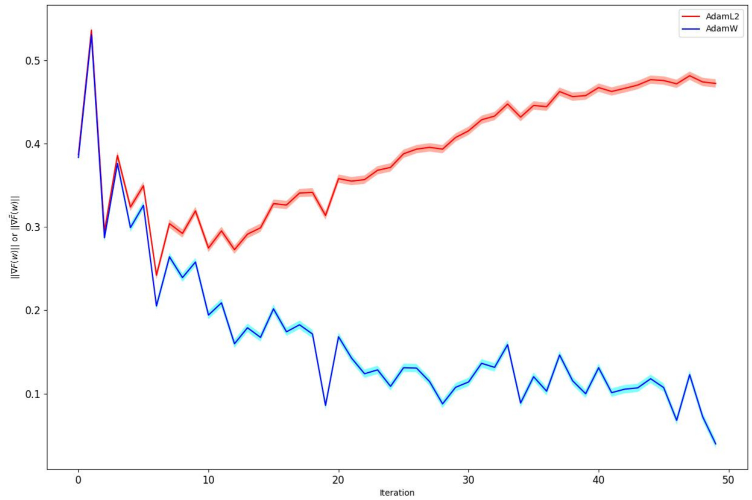

Рис. 6: Adam и AdamW с модифицированным критерием в слое нейронной сети

Это явление можно наблюдать при обучении нейроных сетей на задачу классификации, в силу сложности современных нейронных сетей этот эффект наблюдается в отдельных слоях нейронной сети, но все таки он есть и он влияет на сходимость.

4 Заключение

Доказаны две теоремы, анализирующие скорость сходимости методов оптимизации с предобуславливанием, при различных предположениях на функцию потерь и регуляризации. Первая теорема доказана Проведен вычислительный эксперимент, позволящий анализировать свойства методов предобуславливания методов и их эффективность. Проведен анализ альтернативного метода добавления затухания весов в алгоритмы с предобуславливанием.

Список литературы

[1] Olivier Chapelle, Jason Weston, Léon Bottou, and Vladimir Vapnik. Vicinal risk minimization. *Advances in neural information processing systems*, 13, 2000.

[2] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[3] Pádraig Cunningham, Matthieu Cord, and Sarah Jane Delany. Supervised learning. *Machine learning techniques for multimedia: case studies on organization and retrieval*, pages 21–49, 2008.

[4] Ryan M Rifkin and Ross A Lippert. Notes on regularized least squares. 2007.

[5] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[6] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[7] Elad Hazan, Alexander Rakhlin, and Peter Bartlett. Adaptive online gradient descent. *Advances in Neural Information Processing Systems*, 20, 2007.

[8] Guodong Zhang, Chaoqi Wang, Bowen Xu, and Roger Grosse. Three mechanisms of weight decay regularization. *arXiv preprint arXiv:1810.12281*, 2018.

[9] Zhewei Yao, Amir Gholami, Sheng Shen, Mustafa Mustafa, Kurt Keutzer, and Michael W. Mahoney. Adahessian: An adaptive second order optimizer for machine learning, 2021.

[10] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[12] Andrew Goldberg, Xiaojin Zhu, Alex Furger, and Jun-Ming Xu. Oasis: Online active semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pages 362–367, 2011.

[13] Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu, and Xiaogang Wang. Learning spatial regularization with image-level supervisions for multi-label image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5513–5522, 2017.

[14] Yingbo Zhou, Caiming Xiong, and Richard Socher. Improved regularization techniques for end-to-end speech recognition. *arXiv preprint arXiv:1712.07108*, 2017.

[15] Tingting Wu, Xiao Ding, Minji Tang, Hao Zhang, Bing Qin, and Ting Liu. Stgn: an implicit regularization method for learning with noisy labels in natural language processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7587–7598, 2022.

[16] Federico Girosi, Michael Jones, and Tomaso Poggio. Regularization theory and neural networks architectures. *Neural computation*, 7(2):219–269, 1995.

[17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[18] Johannes Schneider and Scott Kirkpatrick. *Stochastic optimization*. Springer Science & Business Media, 2007.

[19] Daniel P Heyman and Matthew J Sobel. *Stochastic models in operations research: stochastic optimization*, volume 2. Courier Corporation, 2004.

[20] James C Spall. Implementation of the simultaneous perturbation algorithm for stochastic optimization. *IEEE Transactions on aerospace and electronic systems*, 34(3):817–823, 1998.

[21] T. Tieleman and G. Hinton. Lecture 6.5 - rmsprop: Divide the gradient by a running average of its recent magnitude. *Lecture 6.5 - rmsprop: Divide the gradient by a running average of its recent magnitude*, 2012.

[22] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond, 2019.

[23] Alexandre Défossez, Léon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of adam and adagrad. *arXiv preprint arXiv:2003.02395*, 2020.

[24] Michael R. Zhang, James Lucas, Geoffrey Hinton, and Jimmy Ba. Lookahead optimizer: k steps forward, 1 step back, 2019.

[25] Boris Ginsburg, Patrice Castonguay, Oleksii Hrinchuk, Oleksii Kuchaiev, Vitaly Lavrukhin, Ryan Leary, Jason Li, Huyen Nguyen, Yang Zhang, and Jonathan M. Cohen. Stochastic gradient methods with layer-wise adaptive moments for training of deep networks, 2020.

[26] Abdurakhmon Sadiev, Aleksandr Beznosikov, Abdulla Jasem Almansoori, Dmitry Kamzolov, Rachael Tappenden, and Martin Takáč. Stochastic gradient methods with preconditioned updates. *arXiv preprint arXiv:2206.00285*, 2022.

[27] Aleksandr Beznosikov, Aibek Alanov, Dmitry Kovalev, Martin Takáč, and Alexander Gasnikov. On scaled methods for saddle point problems. *arXiv preprint arXiv:2206.08303*, 2022.

A Приложение

A.1 Доказательство лемм

Доказательство. (Лемма 1)

1. Благодаря выражению (9) $\hat{D}_t$ ограничена снизу. Из уравнений обновления матрицы предобуславливания (8) и (9) можно сделать вывод, что условие 2 выполняется из итерационного обновления матрицы, а также абсолютное значение диагональных элементов матрицы D_t ограниченно сверху после каждого обновления, по причине того, что матрица H_t ограниченна сверху.

2. Можем ограничить норму разности сверху, используя (7) и тот факт, что все матрицы диагональные

$$\begin{aligned} ||\hat{D}_{t+1} - \hat{D}_t||_\infty \leq ||D_{t+1} - D_t||_\infty &= ||((D_{t+1})^2 - (D_t)^2)(D_{t+1} + D_t)^{-1}||_\infty \\ &\leq (1-\beta)||((H_t)^2 - (D_t)^2)(D_{t+1} + D_t)^{-1}||_\infty \\ &\leq (1-\beta)\frac{\Gamma^2}{2\alpha}. \end{aligned}$$

   Ограничили каждый фактор по отдельности: $||(D_t)^2 - (H_t)^2||_\infty$ ограничен Γ^2, так как каждый из диагональных элементов $(D_t)^2$ и $(H_t)^2$ больше 0 и меньше Γ^2. Второй коэффициент ограничен, поскольку и D_{t+1}, и D_t больше или равны α, как доказано в первом утверждении этой леммы, следовательно,

$$(D_{t+1} + D_t)^{-1} \preccurlyeq \frac{1}{2\alpha}.$$

3. Можем ограничить норму разности сверху, используя (8), аналогично доказательству второго утверждения этой леммы

$$||\hat{D}_{t+1} - \hat{D}_t||_\infty \leq ||D_{t+1} - D_t||_\infty \leq (1-\beta)||H_t - D_t||_\infty \leq 2\Gamma(1-\beta).$$

   Используем первое утверждение этой леммы об ограниченности диагональных элементов.

$\square$

Доказательство. (Лемма 2)

Используя предположения 1, 2, мы можем записать градиент $\widetilde{r}$

$$\nabla \widetilde{r} = \nabla \left(\sum_{i=1}^{d} D_t^i r_i(w_i) \right) = D_t \begin{pmatrix} r_1'(w_1) \\ \vdots \\ r_d'(w_d) \end{pmatrix} = D_t \nabla r.$$

Доказали необходимое на утверждение. $\square$

Доказательство. (Лемма 3) Мы можем написать определение гладкости, используя лемму 2, а затем применить 3 и 5.

$$\begin{aligned} ||\nabla \widetilde{r}(x) - \nabla \widetilde{r}(y)|| &= \left\| \nabla \left(\sum_{i=1}^{d} D_t^i r_i(x_i) \right) - \nabla \left(\sum_{i=1}^{d} D_t^i r_i(y_i) \right) \right\| \\ &= ||D_t \left(\nabla r(x) - \nabla r(y) \right)|| \leq ||D_t|| L_r \leq \Gamma L_r \end{aligned}$$

Доказали необходимое нам утверждение. $\square$

Доказательство. (Proof of lemma 4) Напишем определения решений w^*, $\widetilde{w}^*$:

$$\begin{cases} \nabla f(\widetilde{w}^*) + D_t \nabla r(\widetilde{w}^*) = 0 \\ \nabla f(w^*) + \nabla r(w^*) = 0 \end{cases},$$

Тогда мы можем получить нижнюю границу из определения L_F-контрастности функции F.

$$\begin{aligned} \|\widetilde{w}^* - w^*\| L_F &\geq \|\nabla f(\widetilde{w}^*) + \nabla r(\widetilde{w}^*) - \nabla f(w^*) - \nabla r(w^*)\| \\ &= \| - D_t \nabla r(\widetilde{w}^*) + \nabla r(\widetilde{w}^*)\| = \|\nabla r(\widetilde{w}^*)(I - D_t)\|. \end{aligned}$$

Доказали утверждение. $\square$

A.2 Доказательства теорем

Доказательство. (Теорема 1)

Используем предположение (3) для шагов t и $t+1$:

$$f(w_{t+1}) \leq f(w_t) + \langle \nabla f(w_t), w_{t+1} - w_t \rangle + \frac{L_f}{2} ||w_{t+1} - w_t||^2, \tag{12}$$

По определению нашего алгоритма мы имеем:

$$w_{t+1} - w_t = -\eta D_t^{-1} \nabla f(w_t) - \eta \nabla r(w_t).$$

Из предыдущего выражения выбираем градиент функции

$$\nabla f(w_t) = \frac{1}{\eta} D_t (w_t - w_{t+1}) - D_t \nabla r(w_t),$$

заменим $\nabla f(w_t)$ в 12 и по предположению 5,, $I \preccurlyeq \frac{D_t}{\alpha}$

$$\begin{aligned} f(w_{t+1}) &\leq f(w_t) + \langle \frac{1}{\eta} D_t (w_t - w_{t+1}) - D_t \nabla r(w_t), w_{t+1} - w_t \rangle + \frac{L_f}{2\alpha} ||w_{t+1} - w_t||^2_{D_t} \\ &= f(w_t) + \left(\frac{L_f}{2\alpha} - \frac{1}{\eta} \right) ||w_{t+1} - w_t||^2_{D_t} - \langle D_t \nabla r(w_t), w_{t+1} - w_t \rangle, \end{aligned}$$

используя обозначение $\widetilde{r}_t : \nabla \widetilde{r}_t = D_t \nabla r(w_t)$, мы можем переписать шаг, используя переменную и предположение (3)

$$\widetilde{r}(w_{t+1}) \leq \widetilde{r}(w_t) + \langle \nabla \widetilde{r}(w_t), w_{t+1} - w_t \rangle + \frac{L_{\tilde{r}}}{2} ||w_{t+1} - w_t||^2_2.$$

Заменим старую функцию регуляризации на новую

$$f(w_{t+1}) \leq f(w_t) + \left(\frac{L_f}{2\alpha} - \frac{1}{\eta} \right) ||w_{t+1} - w_t||^2_{D_t} + \tilde{r}(w_t) - \tilde{r}(w_{t+1}) + \frac{\Gamma L_{\tilde{r}}}{2} ||w_{t+1} - w_t||^2_{D_t}.$$

Теперь давайте определим новую функцию потерь $\widetilde{F}_t(w) := f(w) + \tilde{r}_t(w)$, $(\tilde{L} = L_f + \Gamma L_r)$, мы получаем:

$$\widetilde{F}_t(w_{t+1}) \leq \widetilde{F}_t(w_t) + \left(\frac{\widetilde{L}}{2\alpha} - \frac{1}{\eta} \right) ||w_{t+1} - w_t||^2_{D_t},$$

мы выбираем шаг таким образом, чтобы $\frac{\tilde{L}}{2\alpha} - \frac{1}{\eta} < 0 \Leftrightarrow \eta < \frac{2\alpha}{\tilde{L}}$

$$\left(\frac{1}{\eta} - \frac{\tilde{L}}{2\alpha} \right) ||w_{t+1} - w_t||^2_{D_t} \leq \tilde{F}_t(w_t) - \tilde{F}_t(w_{t+1}). \tag{13}$$

Тогда заметим, что согласно алгоритму

$$\begin{aligned}
||w_{t+1} - w_t||^2_{D_t} &= || - \eta D_t^{-1} \nabla f(w_t) - \eta \nabla r(w_t)||^2_{D_t} \\
&= \eta^2 ||D_t^{-1}(\nabla f(w_t) + \nabla \widetilde{r}_t(w_t))||^2_{D_t} \\
&= \eta^2 (\nabla f(w_t) + \nabla \widetilde{r}_t(w_t))^T D_t^{-1} D_t D_t^{-1} (\nabla f(w_t) + \nabla \widetilde{r}_t(w_t)) \\
&\geq \frac{\eta^2}{\Gamma} ||\nabla f(w_t) + \nabla \widetilde{r}_t(w_t)||^2 = \frac{\eta^2}{\Gamma} ||\nabla \widetilde{F}_t(w_t)||^2,
\end{aligned}$$

и заменим $||w_{t+1} - w_t||^2_{D_t}$ в (13) предыдущим уравнением и получим

$$\left(\frac{1}{\eta} - \frac{\tilde{L}}{2\alpha}\right) \frac{\eta^2}{\Gamma} ||\nabla \widetilde{F}_t(w_t)||^2 \leq \tilde{F}_t(w_t) - \tilde{F}_t(w_{t+1}). \tag{14}$$

Чтобы получить уравнение, мы связываем норму разности $\tilde{F}_{t+1}$ и $\tilde{F}_t$ при равных переменных w:

$$\begin{aligned}
|\widetilde{F}_{t+1}(w) - \widetilde{F}_t(w)| = |\widetilde{r}_{t+1}(w) - \widetilde{r}_t(w)| &= \left| \sum_{i=0}^{d} (d^i_{t+1} - d^i_t) r_i(w^i) \right| \\
&\leq \sum_{i=0}^{d} |d^i_{t+1} - d^i_t| r_i(w^i) \leq ||D_{t+1} - D_t||_\infty |r(w)| \\
&\leq \Omega ||D_{t+1} - D_t||_\infty,
\end{aligned}$$

где мы используем предположение 1 и предположение 4.

Затем нам нужно оценить $|\widetilde{F}_{t+1}(w) - \widetilde{F}_t(w)|$, используя лемму 1. Мы связываем последнее уравнение с δ и уточняем δ для случаев (7) и (8)

$$|\widetilde{F}_{t+1}(w) - \widetilde{F}_t(w)| \leq \Omega ||D_{t+1} - D_t||_\infty \leq \delta, \tag{15}$$

где $\delta = \begin{cases} \frac{(1-\beta)\Gamma^2}{2\alpha}\Omega & \text{для (7)} \\ 2(1-\beta)\Gamma\Omega & \text{для (8)} \end{cases}$.

Теперь мы можем оценить следующую разность, используя (14) и (15).

$$\begin{aligned}
\widetilde{F}_t(w_t) - \widetilde{F}_{t+1}(w_{t+1}) &= \widetilde{F}_t(w_t) - \widetilde{F}_t(w_{t+1}) + \widetilde{F}_t(w_{t+1}) - \widetilde{F}_{t+1}(w_{t+1}) \\
&\geq \left(\frac{1}{\eta} - \frac{\tilde{L}}{2\alpha}\right) \frac{\eta^2}{\Gamma} ||\nabla \widetilde{F}_t(w_t)||^2 - \delta,
\end{aligned}$$

и перепишем

$$\left(\frac{1}{\eta} - \frac{\tilde{L}}{2\alpha}\right) \frac{\eta^2}{\Gamma} ||\nabla \widetilde{F}_t(w_t)||^2 \leq \widetilde{F}_t(w_t) - \widetilde{F}_{t+1}(w_{t+1}) + \delta.$$

Теперь просуммируем все итерации предыдущего выражения

$$\begin{aligned}\frac{\eta^2(T+1)}{\Gamma}\left(\frac{1}{\eta} - \frac{\tilde{L}}{2\alpha}\right) \cdot \min_{t\in\overline{0,T}} ||\nabla \widetilde{F}_t(w_t)||^2 &\leq \frac{\eta^2}{\Gamma}\left(\frac{1}{\eta} - \frac{\tilde{L}}{2\alpha}\right) \cdot \sum_{t=0}^{T} ||\nabla \widetilde{F}_t(w_t)||^2 \\ &\leq \tilde{F}(w_0) - \tilde{F}^* + \delta \cdot (T+1)\end{aligned}$$

Переместив все в правую часть, мы получим следующую оценку

$$\min_{t\in\overline{0,T}} ||\nabla f(w_t) + \nabla \tilde{r}(w_t)||^2 \leq \frac{(\tilde{F}(w_0) - \tilde{F}(w_*))\Gamma}{(\frac{1}{\eta} - \frac{\tilde{L}}{2\alpha})\eta^2(T+1)} + \frac{\delta\Gamma}{(\frac{1}{\eta} - \frac{\tilde{L}}{2\alpha})\eta^2} = \varepsilon,$$

$$T+1 \geq \frac{\Delta_0\Gamma}{(\eta - \frac{\tilde{L}\eta^2}{2\alpha})\left(\varepsilon - \frac{\delta\Gamma}{\eta - \frac{\tilde{L}\eta^2}{2\alpha}}\right)}.$$

Мы получаем оценку количества шагов, необходимых для достижения заданной точности

$$T = \mathcal{O}\left(\frac{\Delta_0\Gamma}{(\eta - \frac{\tilde{L}\eta^2}{2\alpha})\left(\varepsilon - \frac{\delta\Gamma}{\eta - \frac{\tilde{L}\eta^2}{2\alpha}}\right)}\right).$$

□

Доказательство. (Теорема 2) В дальнейшем доказательстве нам понадобятся вспомогательный термин - σ_{t+1}^2, он понадобится нам для записи рекурсии:

$$\sigma_{t+1}^2 = ||D_{t+1}||_2^2 = ||\beta D_t + (1-\beta)H_t||_2^2 = \beta^2||D_t + \frac{1-\beta}{\beta}H_t||_2^2.$$

Тогда давайте перепишем:

$$\begin{aligned}\sigma_{t+1}^2 &\leq \beta^2(1+\frac{1}{a})\sigma_t^2 + \left(\frac{1-\beta}{\beta}\right)^2(1+a)||\nabla f(w_t)||_2^2 \\ &= \beta^2(1+\frac{1}{a})\sigma_t^2 + \left(\frac{1-\beta}{\beta}\right)^2(1+a)||\nabla f(w_t) - \nabla f(w_*)||_2^2 \\ &\leq \beta^2(1+\frac{1}{a})\sigma_t^2 + 2\left(\frac{1-\beta}{\beta}\right)^2(1+a)L_f(f(w_t) - f(w_*)).\end{aligned}$$

Выберем $a = \frac{\beta}{1-\beta}$, чтобы получить $\beta^2(1+\frac{1}{a}) = \beta$, учтем это в уравнении σ^2:

$$\sigma_{t+1}^2 \leq \beta\sigma_t^2 + 2\frac{1-\beta}{\beta^2}L_f(f(w_t) - f(w_*)).$$

Запишем норму между текущими весами и решением исходной задачи:

$$||w_{t+1}-w_*||_{D_t}^2 = ||w_t-w_*||_{D_t}^2 - 2\eta\langle\nabla f(w_t)+D_t\nabla r(w_t), w_t-w_*\rangle + \eta^2||\nabla f(w_t)+D_t\nabla r(w_t)||^2$$

С предположением 7 на f:

$$\begin{aligned}||w_{t+1} - w_*||_{D_t}^2 \leq ||w_t - w_*||_{D_t}^2 &+ 2\eta\left(f(w_*) - f(w_t) - \frac{\mu_f}{2}||w_t - w_*||_2^2\right) \\ &- 2\eta\langle\nabla r(w_t), w_t - w_*\rangle_{D_t} + \eta^2||\nabla f(w_t) + D_t\nabla r(w_t)||_{D_t^{-1}}^2.\end{aligned}$$

В случае регуляризации ℓ_2 можем записать третий член:

$$\begin{aligned}-2\eta\langle D_t\nabla r(w_t), w_t - w_*\rangle &= -2\lambda\eta\langle D_t w_t, w_t - w_*\rangle \\ &= -2\lambda\eta\langle w_t - w_*, w_t - w_*\rangle_{D_t} - 2\lambda\eta\langle w_*, w_t - w_*\rangle_{D_t} \\ &= -2\eta\lambda||w_t - w_*||_{D_t}^2 - 2\lambda\eta\langle w_*\sqrt{D_t}, \sqrt{D_t}(w_t - w_*)\rangle \\ &\leq -2\eta\lambda||w_t - w_*||_{D_t}^2 + \lambda\eta||w_*\sqrt{D_t}||_2^2 + \lambda\eta||w_t - w_*||_{D_t}^2 \\ &\leq -\eta\lambda||w_t - w_*||_{D_t}^2 + \frac{\lambda\eta\Omega_0^2}{\alpha}||D_t||_2^2.\end{aligned}$$

Здесь мы использовали, что 5, $\alpha I \preccurlyeq D_t \preccurlyeq \Gamma I$ и $||w_*|||_2^2 \leq \Omega_0^2$. С помощью леммы 2 и L-гладкости 3 функции f, $||\nabla f(w_t) - \nabla f(w_*)||_2^2 \leq 2L_f(f(w_t) - f(w_*))$:

$$\begin{aligned}\eta^2||\nabla f(w_t) + D_t\nabla r(w_t)||_{D_t^{-1}}^2 &\leq 2\eta^2||\nabla f(w_t) - \nabla f(w^*)||_{D_t^{-1}}^2 + 4\eta^2||\nabla r(w_t) \\ &\quad - \nabla r(w_*)||_{D_t}^2 + 4\eta^2||\nabla r(w_*)||_{D_t}^2 \\ &\leq 4\eta^2\lambda^2||w_t - w_*||_{D_t}^2 + \frac{4\eta^2 L_f}{\alpha}\left(f(w_t) - f(w_*)\right) \\ &\quad + 4\eta^2\lambda^2\Omega_0^2||D_t||_2 \\ &\leq 4\eta^2\lambda^2||w_t - w_*||_{D_t}^2 + \frac{4\eta^2 L_f}{\alpha}\left(f(w_t) - f(w_*)\right) \\ &\quad + 4\eta^2\frac{\lambda^2\Omega_0^2}{\alpha}||D_t||_2^2.\end{aligned}$$

Наконец, используя дополнительные обозначения для $R_{t+1}^2 = ||w_{t+1} - w_*||_{D_t}^2$ и леммы 1 об изменении D_t, мы получаем:

$$\begin{aligned} R_{t+1}^2 \le \left(1 - \eta\mu_f - \eta\lambda + 4\eta^2\lambda^2\right)\left(1 + \frac{(1-\beta)\Gamma^2}{2\alpha}\right) R_t^2 + \left(\frac{4\eta^2\lambda^2\Omega_0^2}{\alpha} + \frac{\lambda\eta\Omega_0^2}{\alpha}\right)\sigma_t^2 \\ + \left(\frac{4\eta^2 L_f}{\alpha} - 2\eta\right)\left(f(w_t) - f(w_*)\right). \end{aligned}$$

$$\sigma_{t+1}^2 \le \beta\sigma_t^2 + 2\frac{1-\beta}{\beta^2} L_f (f(w_t) - f(w_*)).$$

$$\begin{aligned} R_{t+1}^2 + M\sigma_{t+1}^2 \le \left(1 - \eta\mu_f - \eta\lambda + 4\eta^2\lambda^2\right)\left(1 + \frac{(1-\beta)\Gamma^2}{2\alpha}\right) R_t^2 \\ + \left(\frac{4\eta^2\lambda^2\Omega_0^2}{\alpha} + \frac{\lambda\eta\Omega_0^2}{\alpha} + M\beta\right)\sigma_t^2 \\ + \left(\frac{4\eta^2 L_f}{\alpha} + 2M\frac{1-\beta}{\beta^2} L_f - 2\eta\right)\left(f(w_t) - f(w_*)\right). \end{aligned}$$

Напишем ограничения на шаг алгоритма, то есть на η:

$$\beta \ge 1 - \frac{\eta(\mu_f + \lambda)\alpha}{2\Gamma^2},$$

$$\eta \le \frac{\mu_f + \lambda}{8\lambda^2}$$

$$\begin{aligned} \left(1 - \eta\mu_f - \eta\lambda + 4\eta^2\lambda^2\right)\left(1 + \frac{(1-\beta)\Gamma^2}{2\alpha}\right) &\le \left(1 - \eta\frac{\mu_f + \lambda}{2}\right)\left(1 + (1-\beta)\frac{\Gamma^2}{2\alpha}\right) \\ &\le \left(1 - \eta\frac{\mu_f + \lambda}{2}\right)\left(1 + \eta\frac{\mu_f + \lambda}{4}\right) \\ &= 1 + \eta\frac{\mu_f + \lambda}{4} - \eta\frac{\mu_f + \lambda}{2} - \eta^2\frac{(\mu_f + \lambda)^2}{8} \\ &= 1 - \eta\frac{\mu_f + \lambda}{4} - \eta^2\frac{(\mu_f + \lambda)^2}{8} \\ &< 1 - \eta\frac{\mu_f + \lambda}{4}. \end{aligned}$$

Запишем ограничения на второй множитель выражения, причем ещё ограничение на шаг обучения $\eta < \frac{1}{4\lambda}$:

$$\left(\frac{4\eta^2\lambda^2\Omega_0^2}{\alpha} + \frac{\lambda\eta\Omega_0^2}{\alpha} + M\beta\right) \le \frac{2\lambda\eta\Omega_0^2}{\alpha} + M\beta = \left(\frac{1+\beta}{2}\right) M.$$

$$M = \frac{4\eta\lambda\Omega_0^2}{\alpha(1-\beta)}.$$

Запишем ограничения на третий множитель:

$$2\eta^2\frac{L_f}{\alpha} - \eta + \frac{1-\beta}{\beta^2}L_f M = 2\eta^2\frac{L_f}{\alpha} - \eta + \frac{1-\beta}{\beta^2}L_f\frac{4\eta\lambda\Omega_0^2}{\alpha(1-\beta)} \leq 0.$$

Поделим обе части на η:

$$2\eta\frac{L_f}{\alpha} - 1 + \frac{1-\beta}{\beta^2}L_f\frac{4\lambda\Omega_0^2}{\alpha(1-\beta)} \leq 0.$$

С ограничениями на λ :

$$\lambda \leq \frac{\alpha\beta^2}{8L_f\Omega_0^2},$$

получили, что

$$\eta < \frac{\alpha}{4L_f} \leq \frac{\alpha}{2L_f}\left(1 - \frac{4L_f\lambda\Omega_0^2}{\alpha\beta^2}\right).$$

Наконец, мы можем запустить рекурсию:

$$\begin{aligned}
R_{T+1}^2 + M\sigma_{T+1}^2 &\leq \left(1 - \eta\frac{\mu_f+\lambda}{4}\right)R_T^2 + \left(\frac{1+\beta}{2}\right)\cdot M\sigma_T^2 \\
&\leq \exp\left(-\min\left\{\eta\frac{\mu_f+\lambda}{4}; -\log\left(\frac{1+\beta}{2}\right)\right\}\right)\left(R_T^2 + M\sigma_T^2\right) \\
&\leq \exp\left(-\min\left\{\eta\frac{\mu_f+\lambda}{4}; \eta\frac{(\mu_f+\lambda)\alpha}{2\Gamma^2}\right\}\right)\left(R_T^2 + M\sigma_T^2\right) \\
&\leq \exp\left(-T\eta\frac{\mu_f+\lambda}{4}\cdot\min\left\{1; \frac{2\alpha}{\Gamma^2}\right\}\right)\left(R_0^2 + M\sigma_0^2\right).
\end{aligned}$$

У нас есть список ограничений на гиперпараметры алгоритма:

1. $\lambda < \frac{\alpha\beta^2}{8L_f\Omega_0^2}$.

2. $\eta \leq \frac{8\mu_f L_f^2\Omega_0^4}{\alpha^2\beta^4} + \frac{L_f\Omega_0^2}{\alpha\beta^2} < \frac{\mu_f+\lambda}{8\lambda^2}$.

3. $\eta < \frac{2L_f\Omega_0^2}{\alpha\beta^2} \leq \frac{1}{4\lambda}$.

4. $\beta \geq 1 - \frac{\eta(\mu_f+\lambda)\alpha}{2\Gamma^2}$.

5. $\eta < \frac{\alpha}{4L_f}$.

$$\eta_{min} = \min\left\{\frac{2L_f\Omega_0^2}{\alpha\beta^2}; \frac{\alpha}{4L_f}; \frac{8\mu_f L_f^2\Omega_0^4}{\alpha^2\beta^4} + \frac{L_f\Omega_0^2}{\alpha\beta^2}\right\}.$$

Получили оценку количества шагов, необходимых для достижения заданной точности

$$T = \mathcal{O}\left(\log\left(\frac{R_0^2 + \frac{8\lambda\Omega_0^2\Gamma^2}{\alpha^2(\mu_f+\lambda)}\sigma_0^2}{\varepsilon}\right)\frac{4}{\eta_{min}(\mu_f+\lambda)\cdot\min\left\{1; \frac{2\alpha}{\Gamma^2}\right\}}\right)$$

□