

УДК 519.85

Унифицированный анализ методов решения вариационных  
неравенств: редукция дисперсии, сэмплирование, квантизация и  
покомпонентный спуск<sup>1</sup>

© 2021 г. А. Н. Безносиков<sup>1</sup>, А. В. Гасников<sup>1,2,3</sup>, К. Э.

Зайнуллина<sup>1</sup>, А. Ю. Масловский<sup>1</sup>, Д. А. Пасечнюк<sup>1,2</sup>,

<sup>1</sup>141701 Московская обл., Долгопрудный, Институтский пер. 9,

Московский физико-технический институт (национальный  
исследовательский университет), Россия

<sup>2</sup>127051 Москва, Большой Каретный переулок, д.19 стр. 1, Институт

проблем передачи информации им. А.А. Харкевича, Россия

<sup>3</sup>385000 Республика Адыгея, Майкоп, ул. Первомайская, д. 208

Кавказский математический центр Адыгейского государственного  
университета, Россия

Поступила в редакцию 27.01.2022 г.

Переработанный вариант 00.00.2022 г.

Принята к публикации 00.00.2022 г.

В данной статье предлагается унифицированный анализ методов для такого широкого класса задач, как вариационные неравенства, который в качестве частных случаев включает в себя задачи минимизации и задачи нахождения седловой точки. Предлагаемый анализ развивается на основе экстра-градиентного метода, являющегося стандартным для решения вариационных

---

<sup>1</sup>Исследование выполнено при поддержке Министерства науки и высшего образования Российской Федерации (госзадание) №075-00337-20-03, номер проекта 0714-2020-0005.

неравенств. Рассматриваются монотонный и сильно монотонный случаи, которые соответствуют выпукло-вогнутым и сильно-выпукло-сильно-вогнутым задачам нахождения седловой точки. Теоретический анализ основан на параметризованных предположениях для итераций экстраградиентного метода. Следовательно, он может служить прочной основой для объединения уже существующих методов различных типов, а также для создания новых алгоритмов. В частности, чтобы показать это, мы разрабатываем некоторые новые надежные методы, в том числе метод с квантизацией, покомпонентный метод, распределенные рандомизированные локальные методы и другие. Большинство из упомянутых подходов прежде никогда не рассматривались в общности вариационных неравенств и применялись лишь для задач минимизации. Стабильность новых методов подтверждается предоставляемыми численными экспериментами по обучению моделей GAN. Библ.35. Фиг.3. Табл.1.

**Ключевые слова:** Экстраградиентный метод, стохастические вариационные неравенства, квантизация, редукция дисперсии.

## 1. Введение

Основная постановка задачи, рассматриваемая в этой статье, является задачей решения вариационного неравенства (ВН), и имеет следующий вид:

$$\text{найти } z^* \in \mathcal{Z} \text{ такое, что } \langle F(z^*), z - z^* \rangle + h(z) - h(z^*) \geq 0, \quad \forall z \in \mathcal{Z}, \quad (1)$$

где  $h : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$  — подходящая полунепрерывная снизу выпуклая функция,  $F : \mathbb{R}^d \rightarrow \mathbb{R}^d$  — оператор,  $\mathcal{Z}$  — непустое замкнутое выпуклое подмножество  $\mathbb{R}^d$ .

Приведем сразу некоторые классические примеры задач, которые могут быть представлены в виде (1):

**Задача минимизации.** Положим, что  $F(x)$  является градиентом некоторой функции  $f(x)$ , а  $h(x) = \delta_{\mathcal{X}}(x)$  — индикаторная функция множества

$\mathcal{X}$ . В частности, при  $\mathcal{X} = \mathbb{R}^d$ ,  $x^* \in \mathcal{X}$  является решением (1) тогда и только тогда, когда  $\nabla f(x^*) = 0$ . В случае выпуклой функции, ему соответствует глобальный экстремум.

**Задача нахождения седловой точки (СТ).** Рассмотрим следующую минимаксную задачу:

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y). \quad (2)$$

Эта задача может быть также представлена в виде (1). Достаточно выбрать  $F$  и  $g$  следующим образом:

$$F(z) = \begin{pmatrix} \nabla_x f(x, y) \\ -\nabla_y f(x, y) \end{pmatrix}, \quad h(z) = \delta_{\mathcal{X}}(x) + \delta_{\mathcal{Y}}(y).$$

Как и в случае задачи минимизации, неравенство (1) является необходимым условием оптимальности. В частности, если  $f(x, y)$  выпукло-вогнута, и  $\mathcal{X} = \mathbb{R}^{d_x}$ ,  $\mathcal{Y} = \mathbb{R}^{d_y}$ , это условие является также и достаточным, причём решение  $(x^*, y^*)$  вариационного неравенства тогда является глобальной седловой точкой:

$$f(x^*, y) \leq f(x, y) \leq f(x, y^*), \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}.$$

Эти примеры показывают, что класс задач ВН достаточно широк. В частности, задачи минимизации могут рассматриваться в общности ВН. Но обычно предпочитают проводить анализ задачи минимизации отдельно и независимо. На данный момент, анализ задач минимизации разработан гораздо шире и глубже, чем для задач ВН. В первую очередь это связано со сложностью задачи ВН: многие техники из минимизации ещё не перенесены или не могут быть перенесены на ВН. Поэтому прежде всего решение ВН интересно с точки зрения нахождения СТ.

Задачи СТ имеют множество практических приложений. Как, например, хорошо известные классические примеры из теории игр и оптимального управления [1]. В последние годы задачи СТ стали популярными и в нескольких других отношениях. Можно отметить ветвь работ, посвященных решению негладких задач путем их переформулирования в виде задачи СТ [2; 3], а также применение таких подходов к задачам обработки изображений [4; 5]. Однако в первую очередь интересны приложения седловых задач в машинном обучении. Конечно, прежде всего здесь стоит упомянуть о GAN. В классической формулировке [6] обучение этих моделей является минимаксной задачей.

Довольно большая часть прикладных задач, в том числе задачи машинного обучения, являются стохастическими, поэтому естественно сосредоточиться на случае, когда невыгодно (или даже невозможно) вычислить полное значение градиента и когда вместо этого используются некоторые стохастические оценки. Например, для функции из (2), имеющей следующий вид:  $f(x, y) = \mathbb{E}_{p_x \sim D_x, p_y \sim D_y} [f_{p_x, p_y}(x, y)]$ . В частности, в случае GAN,  $f$  есть функция потерь, а переменные  $x$  и  $y$  интерпретируются как относящиеся к двум моделям:  $x$  — параметры дискриминатора, а  $y$  — генератора,  $p_x$  — обучающий пример из реального набора данных,  $p_y$  — случайный вектор, который генератор использует для создания поддельных копий реального набора данных. Стандартное предположение статистического обучения состоит в том, что распределение данных  $D_x$  неизвестно, а потому полный градиент  $\nabla_x f(x, y)$  не может быть вычислен, тогда как можно легко вычислить градиент для некоторых отдельных данных. Возвращаясь к основной проблеме (1) этой статьи, мы интерпретируем сказанное выше следующим образом: предположим, что мы не имеем доступа к “честному”

$F(z)$ , а только к некоторой несмещенной стохастической оценке  $F(z, \xi)$ :

$$F(z) = \mathbb{E}_{\xi \sim P} [F(z, \xi)]. \quad (3)$$

В данной работе нас также будет интересовать другая стохастическая постановка задачи (1), это тот случай, когда значение  $F$  — это среднее значение большого числа операторов:

$$F(z) = \frac{1}{M} \sum_{m=1}^M F_m(z). \quad (4)$$

Такие постановки возникают в результате применения подхода интегрирования Монте–Карло. Например, пусть для (2) имеется  $M$  частей набора данных, тогда можно вычислить градиенты  $\nabla_x f_m(x, y)$ ,  $\nabla_y f_m(x, y)$  для каждой из этих частей, тогда как вычисление полных градиентов  $\nabla_x f(x, y)$ ,  $\nabla_y f(x, y)$  будет очень дорогим и потребует много времени. Отсюда кажется естественным выбирать случайный индекс  $m$  части набора данных на каждой итерации и учитывать градиенты только на ней. Этот подход обычно используется при практическом решении задач обучения.

Приведенный выше взгляд на (4) справедлив в случае, когда у нас есть только одно устройство для вычислений. Однако в случае распределенного обучения можно просто обмениваться данными между устройствами, и тогда каждому устройству будет соответствовать свой  $F_m$ . В такой постановке задачу также можно рассматривать в виде (4), но мы предпочтём переписать ее следующим образом:

$$F(Z) = \Phi(Z) + \lambda \cdot (Z - \bar{Z}), \quad (5)$$

где вектора  $\Phi(Z) = [F_1^T(z_1), \dots, F_M^T(z_M)] \in \mathbb{R}^{Md}$ ,  $Z = [z_1^T, \dots, z_M^T] \in \mathbb{R}^{Md}$  и  $\bar{Z} = [\bar{z}^T, \dots, \bar{z}^T] \in \mathbb{R}^{Md}$  для  $\bar{z} = \frac{1}{M} \sum_{m=1}^M z_m$ . Легко проверить, что задача минимизации  $\min_{x_1, \dots, x_M} \left[ \sum_{m=1}^M f_m(x_m) + \frac{\lambda}{2} \sum_{m=1}^M \|x_m - \bar{x}\|^2 \right]$  соответствует

ВН с оператором (5). Аналогично, задача СТ

$$\min_{x_1, \dots, x_M} \max_{y_1, \dots, y_M} \left[ \sum_{m=1}^M f_m(x_m, y_m) + \frac{\lambda}{2} \sum_{m=1}^M \|x_m - \bar{x}\|^2 - \frac{\lambda}{2} \sum_{m=1}^M \|y_m - \bar{y}\|^2 \right]$$

также соответствует ВН с оператором (5) (как построить ВН из задачи минимизации и задачи нахождения седловой точки описано выше). Разобравшись в интуиции, переходим к интерпретации:  $z_1, \dots, z_M$  — локальные модели на каждом устройстве,  $\bar{z}$  — глобальная модель, полученная усреднением всех локальных моделей. Понятно, что нужно выбрать параметр регуляризации  $\lambda$  достаточно большим, чтобы решения задач (4) и (5) были примерно одинаковыми:  $z_m \approx \bar{z}$ . Но имеет ли смысл брать  $\lambda$  малым (например, когда  $\lambda = 0$ , мы имеем просто локальные модели)? Оказывается, да: такая постановка задачи с переменной  $\lambda$ , значение которой может варьироваться, породила целую тенденцию в персонализированном федеративном обучении — федеративное обучение со смешиванием [7; 8]. Постановка (5) будет последней из формулировок задачи (1), которые будут исследоваться в рамках предлагаемого унифицированного анализа.

### 1.1. Экстраградиентный метод

Стохастический градиентный спуск по-прежнему является основным методом минимизации, он используется для большого количества задач машинного обучения, несмотря на наличие более современных методов. Основным же методом решения гладких вариационных неравенств является экстраградиентный метод [9]. В простом виде методы такого рода записываются следующим образом:

$$z^{k+1/2} = \text{prox}_{\gamma h}(z^k - \gamma g^k), \quad z^{k+1} = \text{prox}_{\gamma h}(z^k - \gamma g^{k+1/2}), \quad (6)$$

где  $\text{prox}_{\gamma h}(z) = \arg \min_x \{ \gamma h(x) + \frac{1}{2} \|z - x\|^2 \}$ . В классическом, детерминированном, варианте  $g^k = F(z^k)$  и  $g^{k+1/2} = F(z^{k+1/2})$ . Этот метод оп-

тимален с точностью до численных констант для гладких монотонных и сильно монотонных вариационных неравенств [10]. Причем на практике этот метод показывает себя лучше, чем спуск с итерацией обычного вида  $z^{k+1} = z^k - \gamma g^k$ . Более того, известно, что метод без дополнительного шага расходится для наиболее распространенных билинейных задач. Следовательно, вычисление  $z^{k+1/2}$  является ключевым. В этой работе, для нашего унифицированного анализа, мы используем метод (6) с несколько более сложной структурой:

$$\begin{aligned}\bar{z}^k &= \tau z^k + (1 - \tau)w^k, \\ z^{k+1/2} &= \text{prox}_{\gamma h}(\bar{z}^k - \gamma g^k), \\ z^{k+1} &= \text{prox}_{\gamma h}(\bar{z}^k - \gamma g^{k+1/2}),\end{aligned}\tag{7}$$

где  $w^{k+1} = z^{k+1}$  с вероятностью  $(1 - \tau)$  или  $w^k$ , иначе. Легко увидеть, что при  $\tau = 0$  верно  $w^k = z^k$ , и метод (7) в точности соответствует (6). Метод (7) был впервые предложен в [11].

## 1.2. От минимизации к ВН

SGD используется для решения задач минимизации с середины прошлого века [12] и с того времени был расширен огромным количеством различных модификаций. Это методы уменьшения дисперсии [13], квантования [14] и координатные методы [15] и т.д. Различные варианты SGD см. в [16]. В отличие от задач минимизации, вариационные неравенства и седловые задачи не имеют такого широкого набора теоретических результатов, хотя базовый метод для задачи ВН и более сложен и дает широкий простор для творчества. Но в то же время развитие тех же идей, что и для задач минимизации, для ВН происходит значительно медленнее, в том числе и из-за того, что ВН более общи и сложны в теоретическом анализе.

Далее мы перечислим основные достижения в области решения ВН касательно конструирования методов подобных уже существующим методам минимизации.

- **Базовые методы.** Как отмечалось ранее, базовым методом решения ВН является экстраградиентный метод [9]; еще более общая его версия называется Mirror Prox [3]. Анализ в стохастическом случае с ограниченной дисперсией шума описан в [10]. Стоит обратить внимание на интересные модификации этого метода: с одним дополнительным вызовом оракула [17] и с повторный вызовом [18]. Можно также причислить следующие классические методы, отличающиеся по структуре от экстраградиентного: [19; 20].

- **Редукция дисперсии.** Направление разработки методов редукции дисперсии для задач ВН и СТ развивалось начиная с работы [21], где представлен метод, основанный на сильно выпукло-сильно вогнутых седлах (сильно-монотонных ВН). Также в [22] был предложен метод для сильно монотонных ВН. Наконец, стоит выделить работу [11], которая пересекается с прошлыми результатами или повторяет их, предоставляя методы редукции дисперсии для монотонных и сильно монотонных ВН. Отметим, что приведенные выше методы сильно отличаются друг от друга, более того, они далеки от классической редукции дисперсии для задач минимизации.

- **Покомпонентные и квантизованные методы.** Покомпонентные методы для задач СТ и ВН изучены не слишком хорошо. Можно выделить работы, посвященные конкретным покомпонентным методам для каких-то определенных классов задач СТ [23; 24], а также работы по безградиентным методам [25]. Методы же с квантизацией, специализированные для задач СТ, до сих пор совсем не предлагались.



- **Локальные методы.** Направление разработки локальных методов изучено совсем не в той мере [26; 27], как это сделано для задач минимизации. В данной работе мы обращаем внимание не на детерминированные методы типа Local SGD, а на рандомизированные, которые могут быть применены для решения задачи (5), как например описанные в работе [7]. Суть этих методов в том, что мы с некоторой вероятностью вызываем и делаем шаг только по оракулу для  $\Phi(Z)$ , иначе обращаемся к  $\lambda(Z - \bar{Z})$ . Вызов  $\Phi(Z)$  соответствует локальной итерации, а вызов  $\lambda(Z - \bar{Z})$  – коммуникации.

### 1.3. Наш вклад

- **Унифицированный анализ.** В данной работе предлагается унифицированный теоретический анализ для методов типа (6) и (7) в сильно монотонном и монотонном случаях. Анализ основан на параметризованных предположениях, поэтому он позволяет легко конструировать и анализировать огромное количество новых методов.

- **Улучшенные оценки для существующих методов.** В исходном анализе Past ES из работы [17] в стохастическом сильно монотонном случае достигается сублинейная скорость сходимости, тогда как мы можем гарантировать линейную скорость в детерминированном члене и сублинейную в стохастическом члене. Более того, мы предоставляем оценки для Past ES и в стохастическом монотонном случае. Кроме того, мы покрываем результаты для некоторых других существующих методов или немного их обобщаем.

- **5 новых методов.** Но более важным, чем обобщение других результатов, является получение новых надежных методов. В отличие от работы [16], к моменту написания которой большинство анализируемых там мето-

дов уже были описаны в других работах, в нашей работе более половины методов являются новыми. Это покомпонентные методы, методы с квантизацией, методы с сэмплированием по важности, локальные рандомизированные методы.

- **Эксперименты.** Мы предоставляем сравнение методов на примере практической задачи, в котором демонстрируем, что предлагаемые нами новые методы могут превосходить по эффективности существующие. Эксперименты проводятся на искусственной билинейной задаче и, в некоторых случаях, на GAN.

## 2. Полученные результаты

Для начала введем основные определения. Мы используем аннотацию  $\langle x, y \rangle := \sum_{i=1}^n x_i y_i$  для скалярного произведения векторов  $x, y \in \mathbb{R}^d$ . Оно порождает  $\ell_2$ -норму в пространстве  $\mathbb{R}^d$  в таком виде:  $\|x\| := \sqrt{\langle x, x \rangle}$ .

### 2.1. Основные предположения

Далее нам потребуются два основных предположения. Наше первое предположение относится к монотонности оператора  $F$  из (1) и сильной выпуклости  $g$ . В частности, мы рассматриваем строго монотонный и монотонный случаи. С точки зрения задачи поиска седловой точки это соответствует сильно выпукло-сильно вогнутому и выпукло-вогнутому случаям.

**Предположение 1. (СМ) Сильная монотонность/сильная выпуклость.** *Существуют неотрицательные  $\mu_F, \mu_h$  такие, что  $\mu_h + \mu_F > 0$  и следующие неравенства верны для всех  $z_1, z_2 \in \mathcal{Z}$ :*

$$\langle F(z_1) - F(z_2), z_1 - z_2 \rangle \geq \mu_F \|z_1 - z_2\|^2$$

$$h(z_1) - h(z_2) - \langle \nabla h(z_2), z_1 - z_2 \rangle \geq \frac{\mu_h}{2} \|z_1 - z_2\|^2.$$

**(M) Монотонность/выпуклость.** Для всех  $z_1, z_2 \in \mathcal{Z}$ :

$$\langle F(z_1) - F(z_2), z_1 - z_2 \rangle \geq 0, \quad h(z_1) - h(z_2) - \langle \nabla h(z_2), z_1 - z_2 \rangle \geq 0.$$

Первые два случая в Предположении (1) более чем стандартны. Третий же (не монотонный случай) рассматривает некоторый класс “хороших” немонотонных операторов (не выпуклых-не вогнутых седел). Подробности см., например, в [28].

Второе предположение ключевое и позволяет нам рассматривать разные методы для решения ВН в унифицированном виде. Суть этого предположения проста, аналогично с [16], мы вводим неравенства для основных членов, которые необходимо оценить при анализе экстра-градиентных методов.

**Предположение 2.** Пусть последовательности  $\{z^k\}$  и  $\{w^k\}$  были получены в результате случайных итераций (7). Предположим, что стохастические операторы  $g^{k+1/2}$  несмещены для всех  $k$

$$\mathbb{E} \left[ g^{k+1/2} \mid z^{k+1/2} \right] = F(z^{k+1/2}), \quad (10)$$

Далее, предположим, что существуют неотрицательные  $A, B, C, E, D_1, D_2, D_3$  и  $\rho \in (0; 1]$  и рандомная последовательность  $\{\sigma_k\}$  (может быть нулевой), что выполняются следующие неравенства:

$$\mathbb{E} \left[ \|g^{k+1/2} - g^k\|^2 \right] \leq A \mathbb{E} \left[ \|z^{k+1/2} - w^k\|^2 \right] + B \mathbb{E} [\sigma_k^2] + D_1, \quad (11)$$

$$\mathbb{E} [\sigma_{k+1}^2] \leq (1 - \rho) \mathbb{E} [\sigma_k^2] + C \mathbb{E} \left[ \|z^{k+1/2} - w^k\|^2 \right] + D_2,$$

$$\mathbb{E} \left[ \|g^{k+1/2} - F(z^{k+1/2})\|^2 \right] \leq E \mathbb{E} \left[ \|z^{k+1/2} - w^k\|^2 \right] + D_3.$$

Похожие неравенства могут быть найдены в анализе методов (6) и (7), см. [10; 11]. Это первая работа, которая рассматривает вариационные неравенства в такой общности.

## 2.2. Унифицированная теорема

Мы готовы представить основной теоретический результат данной статьи. Для начала, введем функцию Ляпунова, с помощью которой будем анализировать сходимость:

$$V^k = \tau \|z^{k+1} - z^*\|^2 + \|w^{k+1} - z^*\|^2 + T\gamma^2 \sigma_k^2,$$

где константа  $T > 0$ . Этот критерий используется в сильно-монотонном случае.

Для монотонного случая используется другой критерий - функция зазора [3; 10]:

$$\text{Gap}(z) = \max_{u \in \mathcal{C}} [\langle F(u), z - u \rangle + h(z) - h(u)].$$

Здесь максимум берется не по всему множеству  $\mathcal{Z}$  (как в классической версии), а по  $\mathcal{C}$  - компактному подмножеству множества  $\mathcal{Z}$ . Таким образом мы можем рассматривать неограниченные множества  $\mathcal{Z}$ . Это допустимо, так как такой вариант критерия верен, если решение  $z^*$  лежит в  $\mathcal{C}$ , см. [20].

**Теорема 1.** Пусть выполнено Предположение 2. Тогда, если дополнительно выполняется одно из условий Предположения 1, верны следующие оценки

- для сильно-монотонного/сильно-выпуклого случаев с

$$\gamma \leq \min \left\{ \frac{\sqrt{1-\tau}}{2\sqrt{2A+TC}}; \frac{1-\tau}{4(\mu_F+\mu_h)} \right\} \text{ и } T \geq \frac{4B}{\rho}:$$

$$\mathbb{E}[V_K] \leq \max \left\{ \left( 1 - \gamma \cdot \frac{\mu_F + \mu_h}{16} \right)^{K-1}; \left( 1 - \frac{\rho}{2} \right)^{K-1} \right\} V_0 + \frac{\gamma^2(2D_1 + TD_2)}{\min \left\{ \gamma \cdot \frac{\mu_F + \mu_h}{16}; \frac{\rho}{2} \right\}};$$

- для монотонного/выпуклого случаев с  $\gamma \leq \frac{\sqrt{1-\tau}}{2\sqrt{2A+TC+E}}$  и  $T \geq \frac{2B}{\rho}$

$$\mathbb{E} [\text{Gap}(\bar{z}^K)] \leq \frac{8 \max_{u \in \mathcal{C}} [\|z^0 - u\|^2] + 4\gamma^2 T \sigma_0^2}{\gamma K} + \gamma(7D_1 + 3TD_2 + D_3),$$

$$\text{где } \bar{z}^K = \frac{1}{K} \sum_{k=0}^{K-1} z^{k+1/2}.$$

Метод (7) в условиях Предположения 2 сходится линейно в сильно-монотонном случае и сублинейно ( $\sim \frac{1}{K}$ ) в монотонном случае до определенного радиуса осцилляции сходимости, который зависит от второго члена в Теореме 1. Формально, можно добиться сходимости по этому члену за фиксированное число шагов. Для этого необходимо правильно выбрать шаг  $\gamma$  [29]. Для некоторых методов мы используем это, чтобы получить классические оценки сходимости – можно посмотреть об этом подробнее в следующем разделе. Оптимальный выбор шага для каждого метода находится в приложении, соответствующем этому методу.

Далее приведем доказательство теоремы 1. Для начала докажем лемму.

**Лемма 1.** Пусть  $h$   $\mu_h$  – сильно выпуклая и  $z^+ = \text{prox}_{\gamma h}(z)$ . Тогда для всех  $x \in \mathbb{R}^d$  справедливо следующее неравенство:

$$\langle z^+ - z, x - z^+ \rangle \geq \gamma \left( h(z^+) - h(x) + \frac{\mu_h}{2} \|z^+ - x\|^2 \right).$$

**Доказательство:** Мы используем  $\gamma\mu$ -сильно выпуклость функции  $\gamma h$  (9):

$$\gamma (h(x) - h(z^+)) - \langle \gamma \nabla h(z^+), x - z^+ \rangle \geq \frac{\gamma \mu_h}{2} \|x - z^+\|^2.$$

Вместе с определением  $\text{prox}$  и необходимым условием оптимума:  $\gamma \nabla h(z^+) = z - z^+$ , это завершает доказательство.

Доказательство Теоремы 1:

**Доказательство:** По Лемме 1 для  $z^{k+1/2} = \text{prox}_{\gamma h}(\bar{z}^k - \gamma g^k)$  and  $z^{k+1} = \text{prox}_{\gamma h}(\bar{z}^k - \gamma g^{k+1/2})$  для  $x = u$  получаем

$$\begin{aligned} \langle z^{k+1} - \bar{z}^k + \gamma g^{k+1/2}, u - z^{k+1} \rangle &\geq \gamma \left( h(z^{k+1}) - h(u) + \frac{\mu h}{2} \|z^{k+1} - u\|^2 \right), \\ \langle z^{k+1/2} - \bar{z}^k + \gamma g^k, z^{k+1} - z^{k+1/2} \rangle \\ &\geq \gamma \left( h(z^{k+1/2}) - h(z^{k+1}) + \frac{\mu h}{2} \|z^{k+1} - z^{k+1/2}\|^2 \right). \end{aligned}$$

Далее суммируем два неравенства и производим некоторые перестановки:

$$\begin{aligned} &\langle z^{k+1} - \bar{z}^k, u - z^{k+1} \rangle + \langle z^{k+1/2} - \bar{z}^k, z^{k+1} - z^{k+1/2} \rangle \\ &+ \gamma \langle g^{k+1/2} - g^k, z^{k+1/2} - z^{k+1} \rangle + \gamma \langle g^{k+1/2}, u - z^{k+1/2} \rangle \\ &\geq \gamma \left( h(z^{k+1/2}) - h(u) + \frac{\mu h}{2} \|z^{k+1} - z^{k+1/2}\|^2 + \frac{\mu h}{2} \|z^{k+1} - u\|^2 \right). \end{aligned}$$

Умножая на 2 и используя определение  $\bar{z}^k$  из (7), получаем

$$\begin{aligned} &2\tau \langle z^{k+1} - z^k, u - z^{k+1} \rangle + 2(1 - \tau) \langle z^{k+1} - w^k, u - z^{k+1} \rangle \\ &+ 2\tau \langle z^{k+1/2} - z^k, z^{k+1} - z^{k+1/2} \rangle + 2(1 - \tau) \langle z^{k+1/2} - w^k, z^{k+1} - z^{k+1/2} \rangle \\ &+ 2\gamma \langle g^{k+1/2} - g^k, z^{k+1/2} - z^{k+1} \rangle + 2\gamma \langle g^{k+1/2}, u - z^{k+1/2} \rangle \\ &\geq 2\gamma \left( h(z^{k+1/2}) - h(u) + \frac{\mu h}{2} \|z^{k+1} - z^{k+1/2}\|^2 + \frac{\mu h}{2} \|z^{k+1} - u\|^2 \right). \end{aligned}$$

Для первой и второй строки используем выражение  $2\langle a, b \rangle = \|a + b\|^2 - \|a\|^2 - \|b\|^2$ , и получаем

$$\begin{aligned} &\tau (\|z^k - u\|^2 - \|z^{k+1} - z^k\|^2 - \|z^{k+1} - u\|^2) \\ &+ (1 - \tau) (\|w^k - u\|^2 - \|z^{k+1} - w^k\|^2 - \|z^{k+1} - u\|^2) \\ &+ \tau (\|z^{k+1} - z^k\|^2 - \|z^{k+1/2} - z^k\|^2 - \|z^{k+1} - z^{k+1/2}\|^2) \\ &+ (1 - \tau) (\|z^{k+1} - w^k\|^2 - \|z^{k+1/2} - w^k\|^2 - \|z^{k+1} - z^{k+1/2}\|^2) \\ &+ 2\gamma \langle g^{k+1/2} - g^k, z^{k+1/2} - z^{k+1} \rangle + 2\gamma \langle g^{k+1/2}, u - z^{k+1/2} \rangle \\ &\geq 2\gamma \left( h(z^{k+1/2}) - h(u) + \frac{\mu h}{2} \|z^{k+1} - z^{k+1/2}\|^2 + \frac{\mu h}{2} \|z^{k+1} - u\|^2 \right). \end{aligned}$$

Небольшая перестановка дает

$$\begin{aligned}
(1 + \gamma\mu_h)\|z^{k+1} - u\|^2 &\leq \tau\|z^k - u\|^2 + (1 - \tau)\|w^k - u\|^2 \\
&- \tau\|z^{k+1/2} - z^k\|^2 - (1 - \tau)\|z^{k+1/2} - w^k\|^2 \\
&- (1 + \gamma\mu_h)\|z^{k+1} - z^{k+1/2}\|^2 \\
&+ 2\gamma\langle g^{k+1/2} - g^k, z^{k+1/2} - z^{k+1} \rangle - 2\gamma\langle g^{k+1/2}, z^{k+1/2} - u \rangle \\
&- 2\gamma\left(h(z^{k+1/2}) - h(u)\right).
\end{aligned}$$

Из простого факта:  $2\langle a, b \rangle \leq \eta\|a\|^2 + \frac{1}{\eta}\|b\|^2$  с  $a = g^{k+1/2} - g^k, b = z^{k+1/2} - z^{k+1}, \eta = 2\gamma$ , следует

$$\begin{aligned}
(1 + \gamma\mu_h)\|z^{k+1} - u\|^2 &\leq \tau\|z^k - u\|^2 + (1 - \tau)\|w^k - u\|^2 \\
&- \tau\|z^{k+1/2} - z^k\|^2 - (1 - \tau)\|z^{k+1/2} - w^k\|^2 \\
&- \left(\frac{1}{2} + \gamma\mu_h\right)\|z^{k+1} - z^{k+1/2}\|^2 + 2\gamma^2\|g^{k+1/2} - g^k\|^2 \\
&- 2\gamma\langle g^{k+1/2}, z^{k+1/2} - u \rangle - 2\gamma\left(h(z^{k+1/2}) - h(u)\right). \tag{12}
\end{aligned}$$

Далее, рассмотрим разные случаи теоремы. Начнем с **СИЛЬНО МОНОТОННОГО/ВЫПУКЛОГО СЛУЧАЯ**. Заменяем  $u = z^*$ , возьмем полное математическое ожидание и получим

$$\begin{aligned}
(1 + \gamma\mu_h)\mathbb{E}\left[\|z^{k+1} - z^*\|^2\right] &\leq \tau\mathbb{E}\left[\|z^k - z^*\|^2\right] + (1 - \tau)\mathbb{E}\left[\|w^k - z^*\|^2\right] \\
&- \tau\mathbb{E}\left[\|z^{k+1/2} - z^k\|^2\right] - (1 - \tau)\mathbb{E}\left[\|z^{k+1/2} - w^k\|^2\right] \\
&- \left(\frac{1}{2} + \gamma\mu_h\right)\mathbb{E}\left[\|z^{k+1} - z^{k+1/2}\|^2\right] + 2\gamma^2\mathbb{E}\left[\|g^{k+1/2} - g^k\|^2\right] \\
&- 2\gamma\mathbb{E}\left[\langle g^{k+1/2}, z^{k+1/2} - z^* \rangle + h(z^{k+1/2}) - h(z^*)\right] \\
&= \tau\mathbb{E}\left[\|z^k - z^*\|^2\right] + (1 - \tau)\mathbb{E}\left[\|w^k - z^*\|^2\right] \\
&- (1 - \tau)\mathbb{E}\left[\|z^{k+1/2} - w^k\|^2\right] \\
&- \left(\frac{1}{2} + \gamma\mu_h\right)\mathbb{E}\left[\|z^{k+1} - z^{k+1/2}\|^2\right] + 2\gamma^2\mathbb{E}\left[\|g^{k+1/2} - g^k\|^2\right]
\end{aligned}$$

$$- 2\gamma\mathbb{E} \left[ \langle \mathbb{E} \left[ g^{k+1/2} \mid z^{k+1/2} \right], z^{k+1/2} - z^* \rangle + h(z^{k+1/2}) - h(z^*) \right].$$

Далее применим Предположение 2, а именно, (10) и (11):

$$\begin{aligned} (1 + \gamma\mu_h)\mathbb{E} [\|z^{k+1} - z^*\|^2] &\leq \tau\mathbb{E} [\|z^k - z^*\|^2] + (1 - \tau)\mathbb{E} [\|w^k - z^*\|^2] \\ &\quad - \tau\mathbb{E} [\|z^{k+1/2} - z^k\|^2] - (1 - \tau)\mathbb{E} [\|z^{k+1/2} - w^k\|^2] \\ &\quad - \left(\frac{1}{2} + \gamma\mu_h\right)\mathbb{E} [\|z^{k+1} - z^{k+1/2}\|^2] \\ &\quad + 2\gamma^2 \left( A\mathbb{E} [\|z^{k+1/2} - w^k\|^2] + B\mathbb{E} [\sigma_k^2] + D_1 \right) \\ &\quad - 2\gamma\mathbb{E} \left[ \langle F(z^{k+1/2}), z^{k+1/2} - z^* \rangle + h(z^{k+1/2}) - h(z^*) \right] \\ &= \tau\mathbb{E} [\|z^k - z^*\|^2] + (1 - \tau)\mathbb{E} [\|w^k - z^*\|^2] - \tau\mathbb{E} [\|z^{k+1/2} - z^k\|^2] \\ &\quad - \left(\frac{1}{2} + \gamma\mu_h\right)\mathbb{E} [\|z^{k+1} - z^{k+1/2}\|^2] \\ &\quad - ((1 - \tau) - 2\gamma^2 A)\mathbb{E} [\|z^{k+1/2} - w^k\|^2] + 2\gamma^2 B\mathbb{E} [\sigma_k^2] \\ &\quad - 2\gamma\mathbb{E} \left[ \langle F(z^{k+1/2}), z^{k+1/2} - z^* \rangle + h(z^{k+1/2}) - h(z^*) \right] + 2\gamma^2 D_1. \end{aligned}$$

Свойство решения (1) дает

$$\begin{aligned} (1 + \gamma\mu_h)\mathbb{E} [\|z^{k+1} - z^*\|^2] &\leq \tau\mathbb{E} [\|z^k - z^*\|^2] + (1 - \tau)\mathbb{E} [\|w^k - z^*\|^2] \\ &\quad - \tau\mathbb{E} [\|z^{k+1/2} - z^k\|^2] - \left(\frac{1}{2} + \gamma\mu_h\right)\mathbb{E} [\|z^{k+1} - z^{k+1/2}\|^2] \\ &\quad - ((1 - \tau) - 2\gamma^2 A)\mathbb{E} [\|z^{k+1/2} - w^k\|^2] + 2\gamma^2 B\mathbb{E} [\sigma_k^2] \\ &\quad - 2\gamma\mathbb{E} \left[ \langle F(z^{k+1/2}) - F(z^*), z^{k+1/2} - z^* \rangle \right] + 2\gamma^2 D_1. \end{aligned}$$

И по Предположению 1 в сильно монотонном случае получим

$$\begin{aligned} (1 + \gamma\mu_h)\mathbb{E} [\|z^{k+1} - z^*\|^2] &\leq \tau\mathbb{E} [\|z^k - z^*\|^2] + (1 - \tau)\mathbb{E} [\|w^k - z^*\|^2] \\ &\quad - \tau\mathbb{E} [\|z^{k+1/2} - z^k\|^2] - \left(\frac{1}{2} + \gamma\mu_h\right)\mathbb{E} [\|z^{k+1} - z^{k+1/2}\|^2] \\ &\quad - ((1 - \tau) - 2\gamma^2 A)\mathbb{E} [\|z^{k+1/2} - w^k\|^2] + 2\gamma^2 B\mathbb{E} [\sigma_k^2] \\ &\quad - 2\gamma\mu_F\mathbb{E} [\|z^{k+1/2} - z^*\|^2] + 2\gamma^2 D_1. \end{aligned}$$



С другой стороны

$$\mathbb{E} [\|w^{k+1} - z^*\|^2] = (1 - \tau)\mathbb{E} [\|z^{k+1} - z^*\|^2] + \tau\mathbb{E} [\|w^k - z^*\|^2].$$

Суммируя два предыдущих неравенства:

$$\begin{aligned} & \tau\mathbb{E} [\|z^{k+1} - z^*\|^2] + \mathbb{E} [\|w^{k+1} - z^*\|^2] \leq \tau\mathbb{E} [\|z^k - z^*\|^2] \\ & + \mathbb{E} [\|w^k - z^*\|^2] - \tau\mathbb{E} [\|z^{k+1/2} - z^k\|^2] - \gamma\mu_h\mathbb{E} [\|z^{k+1} - z^*\|^2] \\ & - \left(\frac{1}{2} + \gamma\mu_h\right)\mathbb{E} [\|z^{k+1} - z^{k+1/2}\|^2] \\ & - ((1 - \tau) - 2\gamma^2 A)\mathbb{E} [\|z^{k+1/2} - w^k\|^2] \\ & + 2\gamma^2 B\mathbb{E} [\sigma_k^2] + 2\gamma^2 D_1 - 2\gamma\mu_F\mathbb{E} [\|z^{k+1/2} - z^*\|^2]. \end{aligned}$$

Добавив  $\mathbb{E} [\gamma^2 T \sigma_{k+1}^2]$ , мы получим функцию Ляпунова с левой стороны:

$$\begin{aligned} \mathbb{E} [V_{k+1}] &= \tau\mathbb{E} [\|z^{k+1} - z^*\|^2] + \mathbb{E} [\|w^{k+1} - z^*\|^2] + \mathbb{E} [\gamma^2 T \sigma_{k+1}^2] \\ &\leq \tau\mathbb{E} [\|z^k - z^*\|^2] + \mathbb{E} [\|w^k - z^*\|^2] - \tau\mathbb{E} [\|z^{k+1/2} - z^k\|^2] \\ &\quad - 2\gamma\mu_F\mathbb{E} [\|z^{k+1/2} - z^*\|^2] - \gamma\mu_h\mathbb{E} [\|z^{k+1} - z^*\|^2] \\ &\quad - \left(\frac{1}{2} + \gamma\mu_h\right)\mathbb{E} [\|z^{k+1} - z^{k+1/2}\|^2] \\ &\quad - ((1 - \tau) - 2\gamma^2 A)\mathbb{E} [\|z^{k+1/2} - w^k\|^2] + 2\gamma^2 B\mathbb{E} [\sigma_k^2] \\ &\quad + \mathbb{E} [\gamma^2 T \sigma_{k+1}^2] + 2\gamma^2 D_1. \end{aligned}$$

С Предположением 2 для  $\sigma_{k+1}$  получаем

$$\begin{aligned} \mathbb{E} [V_{k+1}] &\leq \tau\mathbb{E} [\|z^k - z^*\|^2] + \mathbb{E} [\|w^k - z^*\|^2] \\ &\quad + \left(1 - \rho + \frac{2B}{T}\right)\mathbb{E} [\gamma^2 T \sigma_k^2] \\ &\quad - \tau\mathbb{E} [\|z^{k+1/2} - z^k\|^2] - \gamma\mu_h\mathbb{E} [\|z^{k+1} - z^*\|^2] \\ &\quad - \left(\frac{1}{2} + \gamma\mu_h\right)\mathbb{E} [\|z^{k+1} - z^{k+1/2}\|^2] \\ &\quad - ((1 - \tau) - 2\gamma^2 A - \gamma^2 TC)\mathbb{E} [\|z^{k+1/2} - w^k\|^2] \end{aligned}$$

$$- 2\gamma\mu_F \mathbb{E} \left[ \|z^{k+1/2} - z^*\|^2 \right] + \gamma^2(2D_1 + TD_2).$$

Используем  $-\|z^{k+1} - z^*\|^2 \leq -\frac{1}{2}\|z^{k+1/2} - z^*\|^2 + \|z^{k+1} - z^{k+1/2}\|^2$ , что дает:

$$\begin{aligned} \mathbb{E} [V_{k+1}] &\leq \tau \mathbb{E} [\|z^k - z^*\|^2] + \mathbb{E} [\|w^k - z^*\|^2] \\ &\quad + \left(1 - \rho + \frac{2B}{T}\right) \mathbb{E} [\gamma^2 T \sigma_k^2] - \tau \mathbb{E} [\|z^{k+1/2} - z^k\|^2] \\ &\quad - ((1 - \tau) - 2\gamma^2 A - \gamma^2 TC) \mathbb{E} [\|z^{k+1/2} - w^k\|^2] \\ &\quad - \left(\frac{1}{2} + \gamma\mu_h\right) \mathbb{E} [\|z^{k+1} - z^{k+1/2}\|^2] \\ &\quad - \gamma \left(2\mu_F + \frac{\mu_h}{2}\right) \tau \mathbb{E} [\|z^{k+1/2} - z^*\|^2] \\ &\quad - \gamma \left(2\mu_F + \frac{\mu_h}{2}\right) \cdot (1 - \tau) \mathbb{E} [\|z^{k+1/2} - z^*\|^2] \\ &\quad + \gamma^2(2D_1 + TD_2). \end{aligned}$$

Из простых фактов:  $\|z^{k+1/2} - z^*\|^2 \geq \frac{1}{2}\|z^k - z^*\|^2 - \|z^{k+1/2} - z^k\|^2$  и  $\|z^{k+1/2} - z^*\|^2 \geq \frac{1}{2}\|w^k - z^*\|^2 - \|z^{k+1/2} - w^k\|^2$ , следует

$$\begin{aligned} \mathbb{E} [V_{k+1}] &\leq \tau \mathbb{E} [\|z^k - z^*\|^2] + \mathbb{E} [\|w^k - z^*\|^2] + \left(1 - \rho + \frac{2B}{T}\right) \mathbb{E} [\gamma^2 T \sigma_k^2] \\ &\quad - \left((1 - \tau) - 2\gamma^2 A - \gamma^2 TC - \gamma \left(2\mu_F + \frac{\mu_h}{2}\right) \cdot (1 - \tau)\right) \mathbb{E} [\|z^{k+1/2} - w^k\|^2] \\ &\quad - \left(\frac{1}{2} + \gamma\mu_h\right) \mathbb{E} [\|z^{k+1} - z^{k+1/2}\|^2] \\ &\quad - \left(1 - \gamma \left(2\mu_F + \frac{\mu_h}{2}\right)\right) \tau \mathbb{E} [\|z^{k+1/2} - z^k\|^2] \\ &\quad - \gamma \left(\mu_F + \frac{\mu_h}{4}\right) \tau \mathbb{E} [\|z^k - z^*\|^2] \\ &\quad - \gamma \left(\mu_F + \frac{\mu_h}{4}\right) \cdot (1 - \tau) \mathbb{E} [\|w^k - z^*\|^2] + \gamma^2(2D_1 + TD_2). \end{aligned} \tag{13}$$

Далее работаем с предпоследней строкой (13):

$$\begin{aligned} &-\gamma \left(\mu_F + \frac{\mu_h}{4}\right) \tau \mathbb{E} [\|z^k - z^*\|^2] - \gamma \left(\mu_F + \frac{\mu_h}{4}\right) \cdot (1 - \tau) \mathbb{E} [\|w^k - z^*\|^2] \\ &= -\frac{\gamma}{2} \left(\mu_F + \frac{\mu_h}{4}\right) \tau \mathbb{E} [\|z^k - z^*\|^2] \\ &\quad - \frac{\gamma}{2} \left(\mu_F + \frac{\mu_h}{4}\right) \tau \mathbb{E} [\|z^k - z^*\|^2] \end{aligned}$$

$$\begin{aligned}
& -\gamma \left( \mu_F + \frac{\mu_h}{4} \right) \cdot (1 - \tau) \mathbb{E} [\|w^k - z^*\|^2] \\
\leq & -\frac{\gamma}{2} \left( \mu_F + \frac{\mu_h}{4} \right) \tau \mathbb{E} [\|z^k - z^*\|^2] - \frac{\gamma}{4} \left( \mu_F + \frac{\mu_h}{4} \right) \tau \mathbb{E} [\|w^k - z^*\|^2] \\
& + \frac{\gamma}{2} \left( \mu_F + \frac{\mu_h}{4} \right) \tau \mathbb{E} [\|z^k - w^k\|^2] \\
& -\gamma \left( \mu_F + \frac{\mu_h}{4} \right) \cdot (1 - \tau) \mathbb{E} [\|w^k - z^*\|^2] \\
\leq & -\frac{\gamma}{4} \left( \mu_F + \frac{\mu_h}{4} \right) \tau \mathbb{E} [\|z^k - z^*\|^2] - \frac{\gamma}{4} \left( \mu_F + \frac{\mu_h}{4} \right) \mathbb{E} [\|w^k - z^*\|^2] \\
& + \frac{\gamma}{2} \left( \mu_F + \frac{\mu_h}{4} \right) \tau \mathbb{E} [\|z^k - w^k\|^2] \\
\leq & -\frac{\gamma}{4} \left( \mu_F + \frac{\mu_h}{4} \right) \tau \mathbb{E} [\|z^k - z^*\|^2] - \frac{\gamma}{4} \left( \mu_F + \frac{\mu_h}{4} \right) \mathbb{E} [\|w^k - z^*\|^2] \\
& + \gamma \left( \mu_F + \frac{\mu_h}{4} \right) \tau \mathbb{E} [\|z^{k+1/2} - z^k\|^2] \\
& + \gamma \left( \mu_F + \frac{\mu_h}{4} \right) \tau \mathbb{E} [\|z^{k+1/2} - w^k\|^2].
\end{aligned}$$

Подставив в (13), получим

$$\begin{aligned}
\mathbb{E} [V_{k+1}] & \leq \tau \mathbb{E} [\|z^k - z^*\|^2] + \mathbb{E} [\|w^k - z^*\|^2] \\
& + \left( 1 - \rho + \frac{2B}{T} \right) \mathbb{E} [\gamma^2 T \sigma_k^2] \\
& - \left( (1 - \tau) - 2\gamma^2 A - \gamma^2 TC - \gamma \left( 2\mu_F + \frac{\mu_h}{2} \right) \right) \mathbb{E} [\|z^{k+1/2} - w^k\|^2] \\
& - \left( \frac{1}{2} + \gamma \mu_h \right) \mathbb{E} [\|z^{k+1} - z^{k+1/2}\|^2] \\
& - \left( 1 - 3\gamma \left( \mu_F + \frac{\mu_h}{4} \right) \right) \tau \mathbb{E} [\|z^{k+1/2} - z^k\|^2] \\
& - \frac{\gamma}{4} \left( \mu_F + \frac{\mu_h}{4} \right) \tau \mathbb{E} [\|z^k - z^*\|^2] \\
& - \frac{\gamma}{4} \left( \mu_F + \frac{\mu_h}{4} \right) \mathbb{E} [\|w^k - z^*\|^2] + \gamma^2 (2D_1 + TD_2). \tag{14}
\end{aligned}$$

Остается только выбрать  $\gamma \leq \min \left\{ \frac{\sqrt{1-\tau}}{2\sqrt{2A+TC}}; \frac{1-\tau}{4(\mu_F+\mu_h)} \right\}$  и  $T \geq \frac{4B}{\rho}$  и получить

$$\begin{aligned}
\mathbb{E} [V_{k+1}] & \leq \left( 1 - \frac{\gamma}{4} \left( \mu_F + \frac{\mu_h}{4} \right) \right) (\tau \mathbb{E} [\|z^k - z^*\|^2] + \mathbb{E} [\|w^k - z^*\|^2]) \\
& + \left( 1 - \frac{\rho}{2} \right) \mathbb{E} [\gamma^2 T \sigma_k^2] + \gamma^2 (2D_1 + TD_2),
\end{aligned}$$

и в результате:

$$\begin{aligned} \mathbb{E}[V_{k+1}] &\leq \max \left\{ \left( 1 - \gamma \cdot \frac{\mu_F + \mu_h}{16} \right); \left( 1 - \frac{\rho}{2} \right) \right\} \cdot \mathbb{E}[V_k] \\ &\quad + \gamma^2(2D_1 + TD_2). \end{aligned}$$

Выполнение рекурсивных переходов завершает доказательство.

Далее рассмотрим **Монотонный/выпуклый случай** ( $\mu_h = 0$ ,  $\mu_F = 0$ ). Начнем с (12) с дополнительным обозначением  $\text{gap}(z^{k+1/2}, u) = \langle F(z^{k+1/2}), z^{k+1/2} - u \rangle + h(z^{k+1/2}) - h(u)$ :

$$\begin{aligned} 2\gamma \cdot \text{gap}(z^{k+1/2}, u) + \|z^{k+1} - u\|^2 &\leq \tau \|z^k - u\|^2 + (1 - \tau) \|w^k - u\|^2 \\ &\quad - \tau \|z^{k+1/2} - z^k\|^2 - (1 - \tau) \|z^{k+1/2} - w^k\|^2 \\ &\quad + 2\gamma^2 \|g^{k+1/2} - g^k\|^2 \\ &\quad - 2\gamma \langle g^{k+1/2} - F(z^{k+1/2}), z^{k+1/2} - u \rangle. \end{aligned}$$

Добавив к обоим частям  $\|w^{k+1} - u\|^2$  и произведя некоторые перестановки, получим

$$\begin{aligned} 2\gamma \cdot \text{gap}(z^{k+1/2}, u) &\leq [\tau \|z^k - u\|^2 + \|w^k - u\|^2] \\ &\quad - [\tau \|z^{k+1} - u\|^2 + \|w^{k+1} - u\|^2] - \tau \|w^k - u\|^2 - \\ &\quad - (1 - \tau) \|z^{k+1} - u\|^2 + \|w^{k+1} - u\|^2 \\ &\quad - \tau \|z^{k+1/2} - z^k\|^2 - (1 - \tau) \|z^{k+1/2} - w^k\|^2 \\ &\quad + 2\gamma^2 \|g^{k+1/2} - g^k\|^2 \\ &\quad - 2\gamma \langle g^{k+1/2} - F(z^{k+1/2}), z^{k+1/2} - u \rangle. \end{aligned}$$

Просуммируем по  $k = 0, \dots, K - 1$ , возьмем максимум от обеих частей по  $z \in \mathcal{C}$ , далее возьмем математическое ожидание и получим

$$2\gamma \cdot \mathbb{E} \left[ \max_{u \in \mathcal{C}} \sum_{k=0}^{K-1} \text{gap}(z^{k+1/2}, u) \right] \leq \max_{u \in \mathcal{C}} [\tau \|z^0 - u\|^2 + \|w^0 - u\|^2]$$

$$\begin{aligned}
& + \mathbb{E} \left[ \max_{u \in \mathcal{C}} \sum_{k=0}^{K-1} [-\tau \|w^k - u\|^2 - (1 - \tau) \|z^{k+1} - u\|^2 + \|w^{k+1} - u\|^2] \right] \\
& - \sum_{k=0}^{K-1} \left[ \tau \mathbb{E} \left[ \|z^{k+1/2} - z^k\|^2 \right] + (1 - \tau) \mathbb{E} \left[ \|z^{k+1/2} - w^k\|^2 \right] \right] \\
& - 2\gamma^2 \mathbb{E} \left[ \|g^{k+1/2} - g^k\|^2 \right] \\
& + 2\gamma \mathbb{E} \left[ \max_{u \in \mathcal{C}} \sum_{k=0}^{K-1} \left[ \langle g^{k+1/2} - F(z^{k+1/2}), u - z^{k+1/2} \rangle \right] \right].
\end{aligned}$$

Используя Предположение 2 для  $\mathbb{E} [\|g^{k+1/2} - g^k\|^2]$ , получаем

$$\begin{aligned}
2\gamma \cdot \mathbb{E} \left[ \max_{u \in \mathcal{C}} \sum_{k=0}^{K-1} \text{gap}(z^{k+1/2}, u) \right] & \leq \max_{u \in \mathcal{C}} [\tau \|z^0 - u\|^2 + \|w^0 - u\|^2] \\
& + \mathbb{E} \left[ \max_{u \in \mathcal{C}} \sum_{k=0}^{K-1} [-\tau \|w^k - u\|^2 - (1 - \tau) \|z^{k+1} - u\|^2 + \|w^{k+1} - u\|^2] \right] \\
& - \sum_{k=0}^{K-1} \left[ \tau \mathbb{E} \left[ \|z^{k+1/2} - z^k\|^2 \right] + (1 - \tau) \mathbb{E} \left[ \|z^{k+1/2} - w^k\|^2 \right] \right] \\
& - 2\gamma^2 \left( A \mathbb{E} \left[ \|z^{k+1/2} - w^k\|^2 \right] + B \mathbb{E} [\sigma_k^2] + D_1 \right) \\
& + 2\gamma \mathbb{E} \left[ \max_{u \in \mathcal{C}} \sum_{k=0}^{K-1} \left[ \langle g^{k+1/2} - F(z^{k+1/2}), u - z^{k+1/2} \rangle \right] \right].
\end{aligned}$$

Добавим и вычтем  $\sum_{k=0}^{K-1} [\gamma^2 T \mathbb{E} [\sigma_{k+1}^2]]$  и применим Предположение 2 для  $\sigma_k$ , что дает

$$\begin{aligned}
2\gamma \cdot \mathbb{E} \left[ \max_{u \in \mathcal{C}} \sum_{k=0}^{K-1} \text{gap}(z^{k+1/2}, u) \right] & \leq \max_{u \in \mathcal{C}} [\tau \|z^0 - u\|^2 + \|w^0 - u\|^2] \\
& + \sum_{k=0}^{K-1} [\gamma^2 T \mathbb{E} [\sigma_{k+1}^2]] - \sum_{k=0}^{K-1} [\gamma^2 T \mathbb{E} [\sigma_{k+1}^2]] \\
& + \mathbb{E} \left[ \max_{u \in \mathcal{C}} \sum_{k=0}^{K-1} [-\tau \|w^k - u\|^2 - (1 - \tau) \|z^{k+1} - u\|^2 + \|w^{k+1} - u\|^2] \right] \\
& - \sum_{k=0}^{K-1} \left[ \tau \mathbb{E} \left[ \|z^{k+1/2} - z^k\|^2 \right] + (1 - \tau - 2\gamma^2 A) \mathbb{E} \left[ \|z^{k+1/2} - w^k\|^2 \right] \right]
\end{aligned}$$

$$\begin{aligned}
& + 2\gamma^2 K D_1 + \sum_{k=0}^{K-1} \left[ 2\gamma^2 B \mathbb{E} [\sigma_k^2] \right] \\
& + 2\gamma \mathbb{E} \left[ \max_{u \in \mathcal{C}} \sum_{k=0}^{K-1} \left[ \langle g^{k+1/2} - F(z^{k+1/2}), u - z^{k+1/2} \rangle \right] \right] \\
\leq & \max_{u \in \mathcal{C}} \left[ \tau \|z^0 - u\|^2 + \|w^0 - u\|^2 \right] \\
& + \sum_{k=0}^{K-1} \left[ \gamma^2 T \left( (1 - \rho) \mathbb{E} [\sigma_k^2] + C \mathbb{E} [\|z^{k+1/2} - w^k\|^2] + D_2 \right) \right] \\
& - \sum_{k=0}^{K-1} \left[ \gamma^2 T \mathbb{E} [\sigma_{k+1}^2] \right] \\
& + \mathbb{E} \left[ \max_{u \in \mathcal{C}} \sum_{k=0}^{K-1} \left[ -\tau \|w^k - u\|^2 - (1 - \tau) \|z^{k+1} - u\|^2 + \|w^{k+1} - u\|^2 \right] \right] \\
& - \sum_{k=0}^{K-1} \left[ \tau \mathbb{E} [\|z^{k+1/2} - z^k\|^2] + (1 - \tau - 2\gamma^2 A) \mathbb{E} [\|z^{k+1/2} - w^k\|^2] \right] \\
& + 2\gamma^2 K D_1 + \sum_{k=0}^{K-1} \left[ 2\gamma^2 B \mathbb{E} [\sigma_k^2] \right] \\
& + 2\gamma \mathbb{E} \left[ \max_{u \in \mathcal{C}} \sum_{k=0}^{K-1} \left[ \langle g^{k+1/2} - F(z^{k+1/2}), u - z^{k+1/2} \rangle \right] \right] \\
= & \max_{u \in \mathcal{C}} \left[ \tau \|z^0 - u\|^2 + \|w^0 - u\|^2 \right] \\
& + \sum_{k=0}^{K-1} \left[ \gamma^2 T \left( 1 + \frac{2B}{T} - \rho \right) \mathbb{E} [\sigma_k^2] \right] - \sum_{k=0}^{K-1} \left[ \gamma^2 T \mathbb{E} [\sigma_{k+1}^2] \right] \\
& + \mathbb{E} \left[ \max_{u \in \mathcal{C}} \sum_{k=0}^{K-1} \left[ -\tau \|w^k - u\|^2 - (1 - \tau) \|z^{k+1} - u\|^2 + \|w^{k+1} - u\|^2 \right] \right] \\
& - \sum_{k=0}^{K-1} \left[ \tau \mathbb{E} [\|z^{k+1/2} - z^k\|^2] \right] \\
& + (1 - \tau - \gamma^2(2A + TC)) \mathbb{E} [\|z^{k+1/2} - w^k\|^2] \\
& + \gamma^2 K(2D_1 + TD_2) \\
& + 2\gamma \mathbb{E} \left[ \max_{u \in \mathcal{C}} \sum_{k=0}^{K-1} \left[ \langle g^{k+1/2} - F(z^{k+1/2}), u - z^{k+1/2} \rangle \right] \right].
\end{aligned}$$

С  $\gamma \leq \frac{\sqrt{1-\tau}}{\sqrt{2A+TC}}$  и  $T \geq \frac{2B}{\rho}$  получим

$$\begin{aligned}
2\gamma \cdot \mathbb{E} \left[ \max_{u \in \mathcal{C}} \sum_{k=0}^{K-1} \text{gap}(z^{k+1/2}, u) \right] &\leq \max_{u \in \mathcal{C}} [\tau \|z^0 - u\|^2 + \|w^0 - u\|^2] + \gamma^2 T \sigma_0^2 \\
&+ \mathbb{E} \left[ \max_{u \in \mathcal{C}} \sum_{k=0}^{K-1} [-\tau \|w^k - u\|^2 - (1-\tau) \|z^{k+1} - u\|^2 + \|w^{k+1} - u\|^2] \right] \\
&+ 2\gamma \mathbb{E} \left[ \max_{u \in \mathcal{C}} \sum_{k=0}^{K-1} [\langle g^{k+1/2} - F(z^{k+1/2}), u - z^{k+1/2} \rangle] \right] \\
&+ \gamma^2 K(2D_1 + TD_2). \tag{15}
\end{aligned}$$

Для того, чтобы завершить доказательство, надо оценить члены в последних двух строках. Начнем с  $\mathbb{E} \left[ \max_{u \in \mathcal{C}} \sum_{k=0}^{K-1} \langle F(z^{k+1/2}) - g^{k+1/2}, z^{k+1/2} - u \rangle \right]$ . Определим последовательность  $v$ :  $v^0 = z^0$ ,  $v^{k+1} = \text{prox}_{\gamma h}(v^k - \gamma \delta_k)$  с  $\delta^k = F(z^{k+1/2}) - g^{k+1/2}$ . Тогда получим

$$\sum_{k=0}^{K-1} \langle \delta^k, z^{k+1/2} - u \rangle = \sum_{k=0}^{K-1} \langle \delta^k, z^{k+1/2} - v^k \rangle + \sum_{k=0}^{K-1} \langle \delta^k, v^k - u \rangle. \tag{16}$$

По определению  $v^{k+1}$  (свойство прох), для всех  $z \in \mathcal{Z}$

$$\langle v^{k+1} - v^k + \gamma \delta^k, z - v^{k+1} \rangle \geq 0.$$

Переписав этого неравенство, получим

$$\begin{aligned}
\langle \gamma \delta^k, v^k - z \rangle &\leq \langle \gamma \delta^k, v^k - v^{k+1} \rangle + \langle v^{k+1} - v^k, z - v^{k+1} \rangle \\
&\leq \langle \gamma \delta^k, v^k - v^{k+1} \rangle + \frac{1}{2} \|v^k - z\|^2 \\
&\quad - \frac{1}{2} \|v^{k+1} - z\|^2 - \frac{1}{2} \|v^k - v^{k+1}\|^2 \\
&\leq \frac{\gamma^2}{2} \|\delta^k\|^2 + \frac{1}{2} \|v^k - v^{k+1}\|^2 + \frac{1}{2} \|v^k - z\|^2 - \frac{1}{2} \|v^{k+1} - z\|^2 \\
&\quad - \frac{1}{2} \|v^k - v^{k+1}\|^2 \\
&= \frac{\gamma^2}{2} \|\delta^k\|^2 + \frac{1}{2} \|v^k - z\|^2 - \frac{1}{2} \|v^{k+1} - z\|^2.
\end{aligned}$$

Вместе с (16) это дает

$$\begin{aligned}
\sum_{k=0}^{K-1} \langle \delta^k, z^{k+1/2} - u \rangle &\leq \sum_{k=0}^{K-1} \langle \delta^k, z^{k+1/2} - v^k \rangle \\
&\quad + \frac{1}{\gamma} \sum_{k=0}^{K-1} \left( \frac{\gamma^2}{2} \|\delta^k\|^2 + \frac{1}{2} \|v^k - u\|^2 - \frac{1}{2} \|v^{k+1} - u\|^2 \right) \\
&\leq \sum_{k=0}^{K-1} \langle \delta^k, z^{k+1/2} - v^k \rangle + \frac{\gamma}{2} \sum_{k=0}^{K-1} \|\delta^k\|^2 + \frac{1}{2\gamma} \|v^0 - u\|^2.
\end{aligned}$$

Берем максимум по  $u$  и получаем

$$\begin{aligned}
\max_{u \in \mathcal{C}} \sum_{k=0}^{K-1} \langle \delta^k, z^{k+1/2} - u \rangle &\leq \sum_{k=0}^{K-1} \langle \delta^k, z^{k+1/2} - v^k \rangle + \frac{1}{2\gamma} \max_{u \in \mathcal{C}} \|v^0 - u\|^2 \\
&\quad + \frac{\gamma}{2} \sum_{k=0}^{K-1} \|F(z^{k+1/2}) - g^{k+1/2}\|^2.
\end{aligned}$$

Берем полное математическое ожидание и получаем

$$\begin{aligned}
\mathbb{E} \left[ \max_{u \in \mathcal{C}} \sum_{k=0}^{K-1} \langle \delta^k, z^{k+1/2} - u \rangle \right] &\leq \mathbb{E} \left[ \sum_{k=0}^{K-1} \langle \delta^k, z^{k+1/2} - v^k \rangle \right] \\
&\quad + \frac{\gamma}{2} \sum_{k=0}^{K-1} \mathbb{E} \left[ \|F(z^{k+1/2}) - g^{k+1/2}\|^2 \right] + \frac{1}{2\gamma} \max_{u \in \mathcal{C}} \|v^0 - u\|^2 \\
&= \mathbb{E} \left[ \sum_{k=0}^{K-1} \langle \mathbb{E} [F(z^{k+1/2}) - g^{k+1/2} \mid z^{k+1/2} - v^k], z^{k+1/2} - v^k \rangle \right] \\
&\quad + \frac{\gamma}{2} \sum_{k=0}^{K-1} \mathbb{E} \left[ \|F(z^{k+1/2}) - g^{k+1/2}\|^2 \right] + \frac{1}{2\gamma} \max_{u \in \mathcal{C}} \|v^0 - u\|^2 \\
&= \frac{\gamma}{2} \sum_{k=0}^{K-1} \mathbb{E} \left[ \|F(z^{k+1/2}) - g^{k+1/2}\|^2 \right] + \frac{1}{2\gamma} \max_{u \in \mathcal{C}} \|v^0 - u\|^2. \tag{17}
\end{aligned}$$

Далее оценим

$$\mathbb{E} \left[ \max_{u \in \mathcal{C}} \sum_{k=0}^{K-1} [-\tau \|w^k - u\|^2 - (1 - \tau) \|z^{k+1} + u\|^2 + \|w^{k+1} - u\|^2] \right],$$

для этого заметим, что

$$\mathbb{E} \left[ \max_{u \in \mathcal{C}} \sum_{k=0}^{K-1} [-\tau \|w^k - u\|^2 - (1 - \tau) \|z^{k+1} - u\|^2 + \|w^{k+1} - u\|^2] \right]$$



$$\begin{aligned}
&= \mathbb{E} \left[ \max_{u \in \mathcal{C}} \sum_{k=0}^{K-1} \left[ -2 \langle (1-\tau)z^{k+1} + \tau w^k - w^{k+1}, u \rangle \right. \right. \\
&\quad \left. \left. - (1-\tau)\|z^{k+1}\|^2 - \tau\|w^k\|^2 + \|w^{k+1}\|^2 \right] \right] \\
&= \mathbb{E} \left[ \max_{u \in \mathcal{C}} \sum_{k=0}^{K-1} \left[ -2 \langle (1-\tau)z^{k+1} + \tau w^k - w^{k+1}, u \rangle \right] \right] \\
&\quad + \mathbb{E} \left[ \sum_{k=0}^{K-1} - (1-\tau)\|z^{k+1}\|^2 - \tau\|w^k\|^2 + \|w^{k+1}\|^2 \right].
\end{aligned}$$

По определению  $w^{k+1}$ :  $\mathbb{E} [(1-\tau)\|z^{k+1}\|^2 + \tau\|w^k\|^2 - \|w^{k+1}\|^2] = 0$ , тогда

$$\begin{aligned}
\mathbb{E} \left[ \max_{u \in \mathcal{C}} \sum_{k=0}^{K-1} \left[ -\tau\|w^k - u\|^2 - (1-\tau)\|z^{k+1} - u\|^2 + \|w^{k+1} - u\|^2 \right] \right] \\
&= 2\mathbb{E} \left[ \max_{u \in \mathcal{C}} \sum_{k=0}^{K-1} \langle (1-\tau)z^{k+1} + \tau w^k - w^{k+1}, -u \rangle \right] \\
&= 2\mathbb{E} \left[ \max_{u \in \mathcal{C}} \sum_{k=0}^{K-1} \langle (1-\tau)z^{k+1} + \tau w^k - w^{k+1}, u \rangle \right].
\end{aligned}$$

Далее можно провести рассуждения аналогично цепочке рассуждений для (17):

$$\begin{aligned}
&\mathbb{E} \left[ \max_{u \in \mathcal{C}} \sum_{k=0}^{K-1} \left[ \tau\|w^k - u\|^2 + (1-\tau)\|z^{k+1} - u\|^2 - \|w^{k+1} - u\|^2 \right] \right] \\
&\leq \sum_{k=0}^{K-1} \mathbb{E} [\|(1-\tau)z^{k+1} + \tau w^k - w^{k+1}\|^2] + \max_{u \in \mathcal{C}} \|v^0 - u\|^2 \\
&= \sum_{k=0}^{K-1} \mathbb{E} [\|\mathbb{E}_{w^{k+1}}[w^{k+1}] - w^{k+1}\|^2] + \max_{u \in \mathcal{C}} \|v^0 - u\|^2 \\
&= \sum_{k=0}^{K-1} \mathbb{E} [-\|\mathbb{E}_{w^{k+1}}[w^{k+1}]\|^2 + \mathbb{E}_{w^{k+1}}\|w^{k+1}\|^2] + \max_{u \in \mathcal{C}} \|v^0 - u\|^2 \\
&= \sum_{k=0}^{K-1} \mathbb{E} [-\|(1-\tau)z^{k+1} + \tau w^k\|^2 + (1-\tau)\|z^{k+1}\|^2 + \tau\|w^k\|^2] \\
&\quad + \max_{u \in \mathcal{C}} \|v^0 - u\|^2 = \sum_{k=0}^{K-1} \tau(1-\tau)\mathbb{E} [\|z^{k+1} - w^k\|^2] + \max_{u \in \mathcal{C}} \|v^0 - u\|^2 \quad (18)
\end{aligned}$$

Подставив (17) и (18) в (15) получим

$$\begin{aligned}
2\gamma \cdot \mathbb{E} \left[ \max_{u \in \mathcal{C}} \sum_{k=0}^{K-1} \text{gap}(z^{k+1/2}, u) \right] &\leq \max_{u \in \mathcal{C}} [(2 + \tau) \|z^0 - u\|^2 + \|w^0 - u\|^2] \\
&+ \gamma^2 T \sigma_0^2 \\
&+ \sum_{k=0}^{K-1} \left[ \tau(1 - \tau) \mathbb{E} [\|z^{k+1} - w^k\|^2] + \gamma^2 \mathbb{E} [\|F(z^{k+1/2}) - g^{k+1/2}\|^2] \right] \\
&+ \gamma^2 K(2D_1 + TD_2).
\end{aligned}$$

Предположение 2 для  $\mathbb{E} [\|F(z^{k+1/2}) - g^{k+1/2}\|^2]$  дает

$$\begin{aligned}
2\gamma \cdot \mathbb{E} \left[ \max_{u \in \mathcal{C}} \sum_{k=0}^{K-1} \text{gap}(z^{k+1/2}, u) \right] &\leq \max_{u \in \mathcal{C}} [(2 + \tau) \|z^0 - u\|^2 + \|w^0 - u\|^2] \\
&+ \sum_{k=0}^{K-1} \left[ \tau(1 - \tau) \mathbb{E} [\|z^{k+1} - w^k\|^2] + \gamma^2 E \mathbb{E} [\|z^{k+1/2} - w^k\|^2] \right] \\
&+ \gamma^2 T \sigma_0^2 + \gamma^2 K(2D_1 + TD_2 + D_3).
\end{aligned}$$

С  $\gamma \leq \frac{\sqrt{1-\tau}}{\sqrt{E}}$  приходим к

$$\begin{aligned}
2\gamma \cdot \mathbb{E} \left[ \max_{u \in \mathcal{C}} \sum_{k=0}^{K-1} \text{gap}(z^{k+1/2}, u) \right] &\leq \max_{u \in \mathcal{C}} [(2 + \tau) \|z^0 - u\|^2 + \|w^0 - u\|^2] \\
&+ (1 - \tau) \sum_{k=0}^{K-1} \left[ \mathbb{E} [\|z^{k+1} - w^k\|^2] + \mathbb{E} [\|z^{k+1/2} - w^k\|^2] \right] \\
&+ \gamma^2 T \sigma_0^2 + \gamma^2 K(2D_1 + TD_2 + D_3) \\
&\leq \max_{u \in \mathcal{C}} [(2 + \tau) \|z^0 - u\|^2 + \|w^0 - u\|^2] + \gamma^2 T \sigma_0^2 \\
&+ 3(1 - \tau) \sum_{k=0}^{K-1} \left[ \mathbb{E} [\|z^{k+1} - z^{k+1/2}\|^2] + \mathbb{E} [\|z^{k+1/2} - w^k\|^2] \right] \\
&+ \gamma^2 K(2D_1 + TD_2 + D_3).
\end{aligned}$$

Вернемся к (14) с  $\mu_h = 0$ ,  $\mu_F = 0$ ,  $T \geq \frac{2B}{\rho}$ ,  $\gamma \leq \frac{\sqrt{1-\tau}}{2\sqrt{2A+TC}}$  и получим

$$\mathbb{E} [V_{k+1}] \leq \mathbb{E} [V_k] - ((1 - \tau) - 2\gamma^2 A - \gamma^2 TC) \mathbb{E} [\|z^{k+1/2} - w^k\|^2]$$

$$\begin{aligned}
& -\frac{1}{2}\mathbb{E}\left[\|z^{k+1} - z^{k+1/2}\|^2\right] + \gamma^2(2D_1 + TD_2) \\
& \leq \mathbb{E}[V_k] - \frac{(1-\tau)}{2}\mathbb{E}\left[\|z^{k+1/2} - w^k\|^2\right] \\
& \quad - \frac{(1-\tau)}{2}\mathbb{E}\left[\|z^{k+1} - z^{k+1/2}\|^2\right] + \gamma^2(2D_1 + TD_2).
\end{aligned}$$

Следовательно, подставляя это, подходим к концу доказательства:

$$\begin{aligned}
2\gamma \cdot \mathbb{E}\left[\max_{u \in \mathcal{C}} \sum_{k=0}^{K-1} \text{gap}(z^{k+1/2}, u)\right] & \leq \max_{u \in \mathcal{C}} \left[(2+\tau)\|z^0 - u\|^2 + \|w^0 - u\|^2\right] \\
& + \gamma^2 T \sigma_0^2 + 6 \sum_{k=0}^{K-1} \left[\mathbb{E}[V_k] - \mathbb{E}[V_{k+1}] + \gamma^2(2D_1 + TD_2)\right] \\
& + \gamma^2 K(2D_1 + TD_2 + D_3) \\
& \leq \max_{u \in \mathcal{C}} \left[(2+7\tau)\|z^0 - u\|^2 + 7\|w^0 - u\|^2\right] + 7\gamma^2 T \sigma_0^2 \\
& + \gamma^2 K(14D_1 + 7TD_2 + D_3) \\
& \leq \max_{u \in \mathcal{C}} \left[16\|z^0 - u\|^2\right] + 7\gamma^2 T \sigma_0^2 + \gamma^2 K(14D_1 + 7TD_2 + D_3).
\end{aligned}$$

Остается немного подкорректировать критерий сходимости на монотонность  $F$  и неравенство Йенсена для выпуклых функций:

$$\begin{aligned}
& \mathbb{E}\left[\max_{u \in \mathcal{C}} \sum_{k=0}^{K-1} \text{gap}(z^{k+1/2}, u)\right] \\
& = \mathbb{E}\left[\max_{u \in \mathcal{C}} \sum_{k=0}^{K-1} \left[\langle F(z^{k+1/2}), z^{k+1/2} - u \rangle + h(z^{k+1/2}) - h(u)\right]\right] \\
& \geq \mathbb{E}\left[\max_{u \in \mathcal{C}} \sum_{k=0}^{K-1} \left[\langle F(u), z^{k+1/2} - u \rangle + h(z^{k+1/2}) - h(u)\right]\right] \\
& \geq \mathbb{E}\left[K \cdot \max_{u \in \mathcal{C}} \left[\langle F(u), \bar{z}^K - u \rangle + h(\bar{z}^K) - h(u)\right]\right] \\
& = K \cdot \mathbb{E}\left[\text{Gap}(\bar{z}^K)\right],
\end{aligned}$$

где мы используем  $\bar{z}^K = \frac{1}{K} \sum_{k=0}^{K-1} z^{k+1/2}$ . Что приводит к

$$\mathbb{E}\left[\text{Gap}(\bar{z}^K)\right] \leq \frac{8 \max_{u \in \mathcal{C}} \left[\|z^0 - u\|^2\right] + 4\gamma^2 T \sigma_0^2}{\gamma K} + \gamma(7D_1 + 3TD_2 + D_3).$$

### 2.3. Анализ для различных методов

В этом разделе устанавливаем связь между единым анализом и конкретными методами, удовлетворяющими предположению 2. В скобках указаны разделы Приложения, где представлен псевдокод соответствующего метода, а также его анализ при предположении 2, основанный на применении теоремы 1. В большинстве случаев анализируются операторы, удовлетворяющие следующему условию:

**Предположение 3.**  $F(z)$  — ограниченно-липшицева с константами  $L$  и  $D$ , то есть для любых  $z_1, z_2 \in \mathcal{Z}$  верно

$$\|F(z_1) - F(z_2)\|^2 \leq L^2 \|z_1 - z_2\|^2 + D^2.$$

Заметим, что для  $D = 0$  это эквивалентно определению липшицевости. При  $D > 0$ , это предположение покрывает случай, когда оператор не является липшицевым, но ограничен.

- **Существующие методы (A.1 - A.3).** Прежде всего хотелось бы упомянуть методы, которые соответствуют нашему параметризованному предположению. Это, конечно, классический экстраградиентный [10] для задачи (1) + (3). Далее, отметим методы с однократным вызовом оракула [17]; отличие этих методов от классического экстраградиентного в том, что на каждой итерации они вычисляют новое значение оператора  $F$  лишь один раз. Например, этого можно добиться, используя значение  $F$  с предыдущей итерации:  $g^k = F(z^{k-1/2})$ ,  $g^{k+1/2} = F(z^{k+1/2})$  (в экстраградиентном методе имеем:  $g^k = F(z^k)$ ,  $g^{k+1/2} = F(z^{k+1/2})$ ). Вариант метода редукции дисперсии [11], специализированный для задач решения ВН (1) + (4) также удовлетворяет условиям предлагаемого анализа.

- **Ко-коэрцитивность.** Это предположение аналогично липшицевости

оператора:

$$\|F(z_1) - F(z_2)\|^2 \leq l \langle F(z_1) - F(z_2), z_1 - z_2 \rangle.$$

Легко видеть, что  $l$ -коэрцитивный оператор также является  $l$ -липшицевым (обратное, вообще говоря, неверно). Более того, если  $F$  — градиент выпуклой функции, то  $l$ -липшицевость и  $l$ -коэрцитивность эквивалентны. В литературе имеется анализ некоторых методов (например, метода редукции дисперсии [22]) с этим дополнительным допущением. Довольно легко проанализировать многие методы решения ВН с предположением ко-коэрцитивности. Мы также могли бы построить унифицированную теорию вокруг него и так перенести многие методы минимизации в контекст ВН. Но основная проблема предположения о ко-коэрцитивности состоит в том, что это свойство не выполняется для самой распространенной, билинейной, задачи. Поэтому такой анализ будет справедлив только для минимизации, а это уже сделано в работе [16].

- **Coord-ES** для (1) + (3) (A.4). Наш первый новый метод позволяет работать не с полным оператором  $F$ , а выбирать его случайную координату (координаты) и делать шаг только вдоль нее. Методы этого типа называются координатными [30]. С помощью таких методов можно произвести более тщательный поиск решения — выбрать направления, в которых оператор изменяется в большей степени, и проделывать больше шагов в этих направлениях [31]. Также координатный метод очень близок к безградиентным методам [25], которые актуальны, когда мы работаем с функциями в соответствии с моделью черного ящика, и не можем вычислить оператор  $F$ /градиент.

- **Quant-ES** для (1) + (3) (A.5). Суть **Quant-ES** заключается в исполь-

зовании так называемого оператора квантизации:

$$\mathbb{E}Q(x) = x, \quad \mathbb{E}\|Q(x)\|^2 = \omega\|x\|, \quad \text{для любых } x.$$

Такие операторы могут быть рандомизированными или детерминированными, с большим или малым параметром  $\omega$  [14; 32], но все они имеют одну и ту же функцию - сжать вектор  $x$ . Методы с квантизацией популярны с точки зрения распределенной оптимизации, поскольку основным узким местом там является коммуникация, а сжатие позволяет передавать меньше информации и, следовательно, выигрывать в этом отношении. Мы представляем метод для вариационных неравенств, который может использовать квантизованный оператор.

- **QVR-ES** for (1) + (4) (A.6). QVR-ES сочетает методы редукции дисперсии и квантизации, т.е. сначала мы выбираем случайную функцию с номером  $m$  из  $M$  вариантов, а затем также квантизуем ее. В простейшем виде это выглядит так:  $Q(F_m(z))$  – в нашем методе это делается немного в другом виде, но суть остается той же. Этот метод красочно демонстрирует гибкость нашего подхода и возможность создания различных комбинаций методов с использованием параметризованного предположения 2.

- **IS-ES** (1) + (4) (A.7). В этом случае мы рассматриваем задачу более общую, чем (1) + (4). Здесь мы предполагаем, что мы не вызываем функции случайно и равномерно от 1 до  $M$ . Теперь каждый оператор  $F_m$  имеет свой вес, в зависимости от которого его можно вызывать чаще или реже.

- **Local-ES** for (1) + (5) (A.8). Этот метод относится к так называемым локальным методам, которые делают ряд локальных обновлений между периодическими коммуникациями. Наш метод является рандомизированным [7] и основан на методе из предыдущего абзаца, и также использует технику сэмплирования по важности.

### 3. Заключение

#### Список литературы

1. *Facchinei F., Pang J.* Finite-Dimensional Variational Inequalities and Complementarity Problems. — Springer New York, 2007. — (Springer Series in Operations Research and Financial Engineering). — URL: [https://books.google.ru/books?id=1X%5C\\_7Rce3%5C\\_Q0C](https://books.google.ru/books?id=1X%5C_7Rce3%5C_Q0C).
2. *Nesterov Y.* Smooth minimization of non-smooth functions // Mathematical programming. — 2005. — Т. 103, № 1. — С. 127–152.
3. *Nemirovski A.* Prox-method with rate of convergence  $O(1/t)$  for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems // SIAM Journal on Optimization. — 2004. — Т. 15, № 1. — С. 229–251.
4. *Chambolle A., Pock T.* A first-order primal-dual algorithm for convex problems with applications to imaging // Journal of mathematical imaging and vision. — 2011. — Т. 40, № 1. — С. 120–145.
5. *Esser E., Zhang X., Chan T. F.* A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science // SIAM Journal on Imaging Sciences. — 2010. — Т. 3, № 4. — С. 1015–1046.
6. Generative Adversarial Networks / I. J. Goodfellow [и др.]. — 2014. — arXiv: [1406.2661](https://arxiv.org/abs/1406.2661) [stat.ML].
7. *Hanzely F., Richtárik P.* Federated learning of a mixture of global and local models // arXiv preprint arXiv:2002.05516. — 2020.
8. Lower bounds and optimal algorithms for personalized federated learning / F. Hanzely [и др.] // arXiv preprint arXiv:2010.02372. — 2020.

9. *Korpelevich G. M.* The extragradient method for finding saddle points and other problems // . — 1976.
10. *Juditsky A., Nemirovski A., Tauvel C.* Solving variational inequalities with stochastic mirror-prox algorithm // Stochastic Systems. — 2011. — T. 1, № 1. — C. 17—58.
11. *Alacaoglu A., Malitsky Y.* Stochastic Variance Reduction for Variational Inequality Methods // arXiv preprint arXiv:2102.08352. — 2021.
12. *Robbins H., Monro S.* A Stochastic Approximation Method // The Annals of Mathematical Statistics. — 1951. — T. 22, № 3. — C. 400—407. — URL: <https://doi.org/10.1214/aoms/1177729586>.
13. *Johnson R., Zhang T.* Accelerating stochastic gradient descent using predictive variance reduction // Advances in neural information processing systems. — 2013. — T. 26. — C. 315—323.
14. QSGD: Communication-efficient SGD via gradient quantization and encoding / D. Alistarh [и др.] // Advances in Neural Information Processing Systems. — 2017. — C. 1709—1720.
15. *Hanzely F., Mishchenko K., Richtárik P.* SEGA: Variance reduction via gradient sketching // arXiv preprint arXiv:1809.03054. — 2018.
16. *Gorbunov E., Hanzely F., Richtárik P.* A unified theory of SGD: Variance reduction, sampling, quantization and coordinate descent // International Conference on Artificial Intelligence and Statistics. — PMLR. 2020. — C. 680—690.
17. On the convergence of single-call stochastic extra-gradient methods / Y.-G. Hsieh [и др.]. — 2019. — arXiv: [1908.08465](https://arxiv.org/abs/1908.08465) [math.OC].



18. Revisiting stochastic extragradient / K. Mishchenko [и др.] // International Conference on Artificial Intelligence and Statistics. — PMLR. 2020. — С. 4573—4582.
19. *Tseng P.* A Modified Forward-Backward Splitting Method for Maximal Monotone Mappings // SIAM Journal on Control and Optimization. — 2000. — Т. 38, № 2. — С. 431—446. — eprint: <https://doi.org/10.1137/S0363012998338806>. — URL: <https://doi.org/10.1137/S0363012998338806>.
20. *Nesterov Y.* Dual extrapolation and its applications to solving variational inequalities and related problems // Mathematical Programming. — 2007. — Т. 109, № 2. — С. 319—344.
21. *Palaniappan B., Bach F.* Stochastic Variance Reduction Methods for Saddle-Point Problems // Advances in Neural Information Processing Systems. Т. 29 / под ред. D. Lee [и др.]. — Curran Associates, Inc., 2016. — URL: <https://proceedings.neurips.cc/paper/2016/file/1aa48fc4880bb0c9b8a1Paper.pdf>.
22. Reducing noise in gan training with variance reduced extragradient / T. Chavdarova [и др.] // arXiv preprint arXiv:1904.08598. — 2019.
23. *Sidford A., Tian K.* Coordinate methods for accelerating regression and faster approximate maximum flow // 2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS). — IEEE. 2018. — С. 922—933.
24. Coordinate methods for matrix games / Y. Carmon [и др.] // arXiv preprint arXiv:2009.08447. — 2020.
25. Zeroth-Order Algorithms for Smooth Saddle-Point Problems / A. Sadiev [и др.] // arXiv preprint arXiv:2009.09908. — 2020.

26. *Deng Y., Mahdavi M.* Local Stochastic Gradient Descent Ascent: Convergence Analysis and Communication Efficiency // International Conference on Artificial Intelligence and Statistics. — PMLR. 2021. — C. 1387—1395.
27. *Beznosikov A., Samokhin V., Gasnikov A.* Distributed Saddle-Point Problems: Lower Bounds, Optimal Algorithms and Federated GANs // arXiv preprint arXiv:2010.13112. — 2021.
28. *Dang C. D., Lan G.* On the convergence properties of non-euclidean extragradient methods for variational inequalities with generalized monotone operators // Computational Optimization and applications. — 2015. — T. 60, № 2. — C. 277—310.
29. *Stich S. U.* Unified optimal analysis of the (stochastic) gradient method // arXiv preprint arXiv:1907.04232. — 2019.
30. *Wright S. J.* Coordinate descent algorithms // Mathematical Programming. — 2015. — T. 151, № 1. — C. 3—34.
31. *Nesterov Y.* Efficiency of coordinate descent methods on huge-scale optimization problems // SIAM Journal on Optimization. — 2012. — T. 22, № 2. — C. 341—362.
32. On Biased Compression for Distributed Learning / A. Beznosikov [и др.] // arXiv preprint arXiv:2002.12410. — 2020.
33. *Barratt S., Sharma R.* A note on the inception score // arXiv preprint arXiv:1801.01973. — 2018.
34. *Radford A., Metz L., Chintala S.* Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. — 2016. — arXiv: [1511.06434 \[cs.LG\]](https://arxiv.org/abs/1511.06434).

35. *Mirza M., Osindero S.* Conditional generative adversarial nets // arXiv preprint arXiv:1411.1784. — 2014.

## А. Анализ для различных методов

### А.1. Экстраградиентный метод

Начнём с простейшего случая в рамках (1) + (3) - стохастического с равномерно ограниченным шумом [10]:

$$F(z) = \mathbb{E} [F(z, \xi)], \quad \mathbb{E} [\|F(z, \xi) - F(z)\|^2] \leq \sigma^2,$$

где  $z$  и  $\xi$  независимые. Для него может быть применён

---

**Algorithm 1** Экстраградиентный метод (Extra Step)

---

**Параметры:** Размер шага  $\gamma$ ,  $K$ .

**Инициализация:** Выбрать  $z^0 \in \mathcal{Z}$ .

**for**  $k = 0, 1, 2, \dots, K - 1$

    Выбрать случайные  $\xi^k, \xi^{k+1/2}$ ,

$z^{k+1/2} = \text{prox}_{\gamma h}(z^k - \gamma F(z^k, \xi^k))$ ,

$z^{k+1} = \text{prox}_{\gamma h}(z^k - \gamma F(z^{k+1/2}, \xi^{k+1/2}))$ .

**end for**

---

Заметим, что в этом алгоритме  $\tau = 0$ , и следовательно  $w^k = z^k$  для любого  $k$ . Также мы полагаем  $\sigma_k = 0$ . Следующая лемма определяет константы и Предположения 2:

**Лемма 2.** Предположим, что  $F$  ограничено-липшицева с константами  $L$  и  $D$  (Предположение 3), тогда  $g^k$  и  $g^{k+1}$  из Алгоритма 1 удовлетворяют Предположению 2 с константами  $A = 3L^2$ ,  $D_1 = 3D^2 + 6\sigma^2$ ,  $D_3 = \sigma^2$ .

**Доказательство:** Легко убедиться, что  $g^{k+1/2}$  несмещённо. Далее,

$$\begin{aligned}
\mathbb{E} \left[ \|g^{k+1/2} - g^k\|^2 \right] &= \mathbb{E} \left[ \|F(z^{k+1/2}, \xi^{k+1/2}) - F(z^k, \xi^k)\|^2 \right] \\
&\leq 3\mathbb{E} \left[ \|F(z^{k+1/2}) - F(z^k)\|^2 \right] \\
&\quad + 3\mathbb{E} \left[ \|F(z^{k+1/2}, \xi^{k+1/2}) - F(z^{k+1/2})\|^2 \right] \\
&\quad + 3\mathbb{E} \left[ \|F(z^k, \xi^k) - F(z^k)\|^2 \right] \\
&\leq 3L^2\mathbb{E} \left[ \|z^{k+1/2} - z^k\|^2 \right] + 3D^2 + 6\sigma^2,
\end{aligned}$$

и наконец,

$$\mathbb{E} \left[ \|g^{k+1/2} - F(z^{k+1/2})\|^2 \right] = \mathbb{E} \left[ \|F(z^{k+1/2}, \xi^{k+1/2}) - F(z^{k+1/2})\|^2 \right] \leq \sigma^2,$$

**Следствие 1.1.** *Предположим, что  $F$  ограничено-липшицева с константами  $L$  и  $D$ . Тогда Extra Step*

- в сильно-монотонном случае с  $\gamma \leq \min \left\{ \frac{1}{6L}; \frac{1}{4(\mu_F + \mu_h)} \right\}$  удовлетворяет

$$\mathbb{E} \left[ \|z^K - z^*\|^2 \right] \leq \left( 1 - \gamma \cdot \frac{\mu_F + \mu_h}{16} \right)^{K-1} \cdot \|z^0 - z^*\|^2 + \frac{96\gamma(D^2 + 2\sigma^2)}{\mu_F + \mu_h},$$

- в монотонном случае  $\gamma \leq \frac{1}{3L}$  удовлетворяет

$$\mathbb{E} \left[ \text{Gap}(\bar{z}^K) \right] \leq \frac{8 \max_{u \in \mathcal{C}} [\|z^0 - u\|^2]}{\gamma K} + \gamma(21D^2 + 43\sigma^2).$$

**Замечание.** Этот анализ покрывает гладкий случай при  $D = 0$ . Чтобы получить оценки для негладкого, но ограниченного оператора  $F$ , достаточно взять  $L = 0$  и положить  $\frac{1}{L} = +\infty$ .

При правильном выборе  $\gamma$  (см., например, [29]), можно прийти к следующим оценкам скорости сходимости: • в сильно-монотонном случае

$$\begin{aligned}
\mathbb{E} \left[ \|z^K - z^*\|^2 \right] &= \tilde{O} \left( \exp \left( -\frac{(\mu_F + \mu_h)(K-1)}{96L} \right) \cdot \|z^0 - z^*\|^2 \right. \\
&\quad \left. + \frac{(D^2 + \sigma^2)}{(\mu_F + \mu_h)^2(K-1)} \right),
\end{aligned}$$

- в монотонном случае

$$\mathbb{E} [\text{Gap}(\bar{z}^K)] = \mathcal{O} \left( \frac{L \max_{u \in \mathcal{C}} [\|z^0 - u\|^2]}{K} + \frac{(D + \sigma) \cdot \max_{u \in \mathcal{C}} [\|z^0 - u\|]}{\sqrt{K}} \right).$$

## A.2. Экстраградиентный метод без дополнительного вызова оракула

Здесь мы также рассматриваем постановку, аналогичную рассматриваемой в предыдущем разделе: (1) + (3). Однако теперь рассматриваем модификацию метода **Extra Step**.

---

**Algorithm 2** Экстраградиентный метод без дополнительного вызова оракула (Past-ES)

---

**Параметры:** Размер шага  $\gamma$ ,  $K$ .

**Инициализация:** Выбрать  $z^0 \in \mathcal{Z}$ .

**for**  $k = 0, 1, 2, \dots, K - 1$

    Выбрать случайно  $\xi_{k+1/2}$ ,

$$z^{k+1/2} = \text{prox}_{\gamma h}(z^k - \gamma F(z^{k-1/2}, \xi_{k-1/2})),$$

$$z^{k+1} = \text{prox}_{\gamma h}(z^k - \gamma F(z^{k+1/2}, \xi_{k+1/2})),$$

**end for**

---

**Лемма 3.** Предположим, что  $F$  ограниченно-липшицев с константами  $L$  и  $M$  (Предположение 3), тогда  $g^k$  и  $g^{k+1}$  из Алгоритма 2 удовлетворяют Предположению 2 с константами  $\rho = \frac{1}{3}$ ,  $B = 3$ ,  $C = 2L^2$ ,  $D_1 = 6\sigma^2$ ,  $D_2 = 4D^2 + 12\sigma^2$ ,  $D_3 = \sigma^2$ .

**Доказательство:** Положим  $\sigma_k^2 = \|F(z^{k-1/2}) - F(z^{k+1/2})\|^2$

$$\begin{aligned} \mathbb{E} [\sigma_k^2] &\leq 2\mathbb{E} [\|F(z^k) - F(z^{k+1/2})\|^2] + 2\mathbb{E} [\|F(z^{k-1/2}) - F(z^k)\|^2] \\ &\leq 2L^2\mathbb{E} [\|z^k - z^{k+1/2}\|^2] + 2L^2\mathbb{E} [\|z^{k-1/2} - z^k\|^2] + 4D^2 \end{aligned}$$

$$\begin{aligned}
&= 2L^2\mathbb{E} \left[ \|z^k - z^{k+1/2}\|^2 \right] + 4D^2 \\
&\quad + 2L^2\mathbb{E} \left[ \|z^{k-1} - \gamma F(z^{k-1/2}, \xi_k) - z^{k-1} + \gamma F(z^{k-3/2}, \xi_{k-1})\|^2 \right] \\
&= 2L^2\mathbb{E} \left[ \|z^k - z^{k+1/2}\|^2 \right] \\
&\quad + 2L^2\gamma^2\mathbb{E} \left[ \|F(z^{k-1/2}, \xi_k) - F(z^{k-3/2}, \xi_{k-1})\|^2 \right] + 4D^2 \\
&\leq 2L^2\mathbb{E} \left[ \|z^k - z^{k+1/2}\|^2 \right] \\
&\quad + 6L^2\gamma^2\mathbb{E} \left[ \|F(z^{k-1/2}) - F(z^{k-3/2})\|^2 \right] + 4D^2 + 12\sigma^2 \\
&\leq 2L^2\mathbb{E} \left[ \|z^k - z^{k+1/2}\|^2 \right] + \frac{2}{3}\mathbb{E} [\sigma_{k-1}^2] + 4D^2 + 12\sigma^2,
\end{aligned}$$

если положить  $\gamma \leq \frac{1}{3L}$ . Следовательно,

$$\begin{aligned}
\mathbb{E} \left[ \|g^{k+1/2} - g^k\|^2 \right] &= \mathbb{E} \left[ \|F(z^{k+1/2}, \xi_{k+1}) - F(z^{k-1/2}, \xi_k)\|^2 \right] \\
&\leq 3\mathbb{E} [\sigma_k^2] + 6\sigma^2,
\end{aligned}$$

и наконец,

$$\mathbb{E} \left[ \|g^{k+1/2} - F(z^{k+1/2})\|^2 \right] = \mathbb{E} \left[ \|F(z^{k+1/2}, \xi^{k+1/2}) - F(z^{k+1/2})\|^2 \right] \leq \sigma^2.$$

**Следствие 1.2.** *Предположим, что  $F$  ограниченно-липицев с константами  $L$  и  $D$ . Тогда Past-ES*

- *в сильно-монотонном случае с  $T = 36$  и  $\gamma \leq \min \left\{ \frac{1}{12L\sqrt{2}}; \frac{1}{4(\mu_F + \mu_h)} \right\}$  удовлетворяет*

$$\begin{aligned}
\mathbb{E} \left[ \|z^K - z^*\|^2 \right] &\leq \left( 1 - \gamma \cdot \frac{\mu_F + \mu_h}{16} \right)^{K-1} \cdot \|z^0 - z^*\|^2 \\
&\quad + \frac{192\gamma(37\sigma^2 + 12D^2)}{\mu_F + \mu_h},
\end{aligned}$$

- *в монотонном случае с  $T = 18$  и  $\gamma \leq \frac{1}{12L\sqrt{2}}$  удовлетворяет*

$$\mathbb{E} \left[ \text{Gap}(\bar{z}^K) \right] \leq \frac{8 \max_{u \in C} [\|z^0 - u\|^2] + 72\gamma^2\sigma_0^2}{\gamma K} + \gamma(216D^2 + 691\sigma^2).$$

Для следующего метода, мы рассматриваем постановку оптимизации оператора вида суммы: (1) +(4).

### А.3. Экстраградиентный метод с редукцией дисперсии

---

**Algorithm 3** Экстраградиентный метод с редукцией дисперсии (VR-ES)

---

**Параметры:** Размер шага  $\gamma$ ,  $K$ .

**Инициализация:** Выбрать  $z^0 = w^0 \in \mathcal{Z}$ .

**for**  $k = 0, 1, 2, \dots, K - 1$

$$\bar{z}^k = \tau z^k + (1 - \tau)w^k$$

Выбрать равномерно случайно  $m_k \in 1 \dots M$ ,

$$z^{k+1/2} = \text{prox}_{\gamma h}(\bar{z}^k - \gamma F(w^k)),$$

$$z^{k+1} = \text{prox}_{\gamma h}(\bar{z}^k - \gamma(F_{m_k}(z^{k+1/2}) - F_{m_k}(w^k) + F(w^k))),$$

$$w^{k+1} = \begin{cases} z^{k+1}, & \text{с вероятностью } 1 - \tau \\ w^k, & \text{с вероятностью } \tau \end{cases}$$

**end for**

---

Мы полагаем  $\sigma_k = 0$ . Следующая лемма даёт значения констант для Предположения 2:

**Лемма 4.** *Предположим, что каждый  $F_{m_k}$  и сам  $F$  ограниченно-липшицевы с константами  $L$  и  $D$  (Предположение 3), тогда  $g^k$  и  $g^{k+1}$  из Алгоритма 3 удовлетворяют Предположению 2 с константами  $A = L^2$ ,  $D_1 = D^2$ ,  $E = 4L^2$ ,  $D_3 = 4D^2$ .*

**Доказательство:** Легко убедиться, что  $g^{k+1/2}$  несмещённо. Далее,

$$\begin{aligned} & \mathbb{E} \left[ \|g^{k+1/2} - g^k\|^2 \right] \\ &= \mathbb{E} \left[ \|F_{m_k}(z^{k+1/2}) - F_{m_k}(w^k) + F(w^k) - F(w^k)\|^2 \right] \\ &= \mathbb{E} \left[ \|F_{m_k}(z^{k+1/2}) - F_{m_k}(w^k)\|^2 \right] \\ &\leq L^2 \mathbb{E} \left[ \|z^{k+1/2} - w^k\|^2 \right] + D^2, \end{aligned}$$



и наконец,

$$\begin{aligned}
& \mathbb{E} \left[ \|g^{k+1/2} - F(z^{k+1/2})\|^2 \right] \\
&= \mathbb{E} \left[ \|F_{m_k}(z^{k+1/2}) - F_{m_k}(w^k) + F(w^k) - F(z^{k+1/2})\|^2 \right] \\
&\leq 2\mathbb{E} \left[ \|F_{m_k}(z^{k+1/2}) - F_{m_k}(w^k)\|^2 \right] + 2\mathbb{E} \left[ \|F(z^{k+1/2}) - F(w^k)\|^2 \right] \\
&\leq 4L^2\mathbb{E} \left[ \|w^k - z^{k+1/2}\|^2 \right] + 4D^2,
\end{aligned}$$

**Следствие 1.3.** *Предположим, что каждый  $F_{m_k}$  и сам  $F$  ограниченно-липшицев с константами  $L$  и  $D$ . Тогда VR-ES*

• *в сильно монотонном случае с  $\gamma \leq \min \left\{ \frac{\sqrt{1-\tau}}{2\sqrt{2}L}; \frac{1-\tau}{4(\mu_F+\mu_h)} \right\}$  удовлетворяет*

$$\begin{aligned}
& \mathbb{E} \left[ \tau \|z^{k+1} - z^*\|^2 + \|w^{k+1} - z^*\|^2 \right] \\
&\leq \left( 1 - \gamma \cdot \frac{\mu_F + \mu_h}{16} \right)^{K-1} \cdot (\tau \|z^0 - z^*\|^2 + \|w^0 - z^*\|^2) \\
&\quad + \frac{32\gamma D^2}{\mu_F + \mu_h},
\end{aligned}$$

• *в монотонном случае с  $\gamma \leq \frac{\sqrt{1-\tau}}{2\sqrt{6}L}$  удовлетворяет*

$$\mathbb{E} \left[ \text{Gap}(\bar{z}^K) \right] \leq \frac{8 \max_{u \in \mathcal{C}} [\|z^0 - u\|^2]}{\gamma K} + 11\gamma D^2.$$

На каждой итераций, мы вычисляем лишь 1 оператор из  $M$ . Но в момент обновления  $w^k$ , необходимо вычислить все  $M$  операторов в новой точке  $w^k$ . На основании этого, мы можем выбрать оптимальное значение для  $\tau$  следующим образом:

$$(1 - \tau)M \sim \tau \quad \Rightarrow \quad \tau = \frac{M}{M + 1}.$$

#### **A.4. Покомпонентный экстраградиентный метод**

Вернёмся назад и снова рассмотрим наиболее общую постановку без конечных сумм: (1).

---

**Algorithm 4** Покомпонентный экстраградиентный метод (Coord-ES)

---

**Параметры:** Размер шага  $\gamma$ ,  $K$ .

**Инициализация:** Выбрать  $z^0 = w^0 \in \mathcal{Z}$ .

**for**  $k = 0, 1, 2, \dots, K - 1$

$$\bar{z}^k = \tau z^k + (1 - \tau)w^k$$

Выбрать равномерно случайно  $i_k \in 1 \dots d$ ,

$$z^{k+1/2} = \text{prox}_{\gamma h}(\bar{z}^k - \gamma F(w^k)),$$

$$z^{k+1} = \text{prox}_{\gamma h}(\bar{z}^k - \gamma(d[F(z^{k+1/2})]_{i_k} e_{i_k} - d[F(w^k)]_{i_k} e_{i_k} + F(w^k))),$$

$$w^{k+1} = \begin{cases} z^{k+1}, & \text{с вероятностью } 1 - \tau \\ w^k, & \text{с вероятностью } \tau \end{cases}$$

**end for**

---

Положим  $\sigma_k = 0$ . Следующая лемма даёт значения констант для Предположения 2:

**Лемма 5.** *Предположим, что  $F$  ограничено-липицев с константами  $L$  and  $D$  (Предположение 3), тогда  $g^k$  и  $g^{k+1}$  из Алгоритма 4 удовлетворяют Предположению 2 с константами  $A = dL^2$ ,  $D_1 = dD^2$ ,  $E = 2(d + 1)L^2$ ,  $D_3 = 2(d + 1)D^2$ .*

**Доказательство:** Легко убедиться, что  $g^{k+1/2}$  несмещённо. Далее,

$$\begin{aligned} & \mathbb{E} \left[ \|g^{k+1/2} - g^k\|^2 \right] \\ &= \mathbb{E} \left[ \|d[F(z^{k+1/2})]_{i_k} e_{i_k} - d[F(w^k)]_{i_k} e_{i_k} + F(w^k) - F(w^k)\|^2 \right] \\ &= \mathbb{E} \left[ \|d[F(z^{k+1/2}) - F(w^k)]_{i_k} e_{i_k}\|^2 \right] \\ &\leq d\mathbb{E} \left[ \|F(z^{k+1/2}) - F(w^k)\|^2 \right] \leq dL^2\mathbb{E} \left[ \|z^{k+1/2} - w^k\|^2 \right] + dD^2, \end{aligned}$$

и наконец,

$$\mathbb{E} \left[ \|g^{k+1/2} - F(z^{k+1/2})\|^2 \right]$$

$$\begin{aligned}
&= \mathbb{E} \left[ \|d[F(z^{k+1/2})]_{i_k} e_{i_k} - d[F(w^k)]_{i_k} e_{i_k} + F(w^k) - F(z^{k+1/2})\|^2 \right] \\
&\leq 2\mathbb{E} \left[ \|d[F(z^{k+1/2})]_{i_k} e_{i_k}\|^2 \right] + 2\mathbb{E} \left[ \|F(z^{k+1/2}) - F(w^k)\|^2 \right] \\
&\leq 2(d+1)L^2\mathbb{E} \left[ \|w^k - z^{k+1/2}\|^2 \right] + 2(d+1)D^2,
\end{aligned}$$

**Следствие 1.4.** *Предположим, что  $F$  ограничено-липшицев с константами  $L$  и  $D$ . Тогда Coord-ES*

• *в сильно монотонном случае с  $\gamma \leq \min \left\{ \frac{\sqrt{1-\tau}}{2\sqrt{2dL}}; \frac{1-\tau}{4(\mu_F + \mu_h)} \right\}$  удовлетворяет*

$$\begin{aligned}
&\mathbb{E} \left[ \tau \|z^{k+1} - z^*\|^2 + \|w^{k+1} - z^*\|^2 \right] \\
&\leq \left( 1 - \gamma \cdot \frac{\mu_F + \mu_h}{16} \right)^{K-1} \cdot (\tau \|z^0 - z^*\|^2 + \|w^0 - z^*\|^2) \\
&\quad + \frac{32\gamma d D^2}{\mu_F + \mu_h},
\end{aligned}$$

• *в монотонном случае с  $\gamma \leq \frac{\sqrt{1-\tau}}{2L\sqrt{4d+2}}$  удовлетворяет*

$$\mathbb{E} \left[ \text{Gap}(\bar{z}^K) \right] \leq \frac{8 \max_{u \in \mathcal{C}} [\|z^0 - u\|^2]}{\gamma K} + \gamma(9d+2)D^2.$$

Оптимальное значение  $\tau = \frac{d}{d+1}$ .

### A.5. Квантизованный экстраградиентный метод

В этом разделе рассматривается метод, использующий квантизованный оператор.

**Определение 1.** *(Квантизация).  $Q(x)$  называется квантизацией вектора  $x \in \mathbb{R}^d$ , если*

$$\mathbb{E}Q(x) = x, \quad \mathbb{E}\|Q(x)\|^2 = \omega\|x\|.$$

для некоторого  $\omega$ .

---

**Algorithm 5** Квантизованный экстраградиентный метод (Quant-ES)

---

**Параметры:** Размер шага  $\gamma$ ,  $K$ .

**Инициализация:** Выбрать  $z^0 = w^0 \in \mathcal{Z}$ .

**for**  $k = 0, 1, 2, \dots, K - 1$

$$\bar{z}^k = \tau z^k + (1 - \tau)w^k$$

$$z^{k+1/2} = \text{prox}_{\gamma h}(\bar{z}^k - \gamma F(w^k)),$$

$$z^{k+1} = \text{prox}_{\gamma h}(\bar{z}^k - \gamma Q(F(z^{k+1/2}) - F(w^k)) + \gamma F(w^k)),$$

$$w^{k+1} = \begin{cases} z^{k+1}, & \text{с вероятностью } 1 - \tau \\ w^k, & \text{с вероятностью } \tau \end{cases}$$

**end for**

---

**Лемма 6.** *Предположим, что  $F$  ограничено-липшицев с константами  $L$  и  $D$  (Предположение 3), тогда  $g^k$  и  $g^{k+1}$  из Алгоритма 5 удовлетворяют Предположению 2 с константами  $A = \omega L^2$ ,  $D_1 = \omega D^2$ ,  $E = 2(\omega + 1)L^2$ ,  $D_3 = 2(\omega + 1)D^2$ .*

**Доказательство:** Легко убедиться, что  $g^{k+1/2}$  несмещённо. Далее,

$$\begin{aligned} \mathbb{E} \left[ \|g^{k+1/2} - g^k\|^2 \right] &= \mathbb{E} \left[ \|Q(F(z^{k+1/2}) - F(w^k)) + F(w^k) - F(w^k)\|^2 \right] \\ &= \mathbb{E} \left[ \|Q(F(z^{k+1/2}) - F(w^k))\|^2 \right] \\ &\leq \omega \mathbb{E} \left[ \|F(z^{k+1/2}) - F(w^k)\|^2 \right] \\ &\leq \omega L^2 \mathbb{E} \left[ \|z^{k+1/2} - w^k\|^2 \right] + \omega D^2 \end{aligned}$$

и наконец,

$$\begin{aligned} &\mathbb{E} \left[ \|g^{k+1/2} - F(z^{k+1/2})\|^2 \right] \\ &= \mathbb{E} \left[ \|Q(F(z^{k+1/2}) - F(w^k)) + F(w^k) - F(z^{k+1/2})\|^2 \right] \\ &\leq 2\mathbb{E} \left[ \|Q(F(z^{k+1/2}) - F(w^k))\|^2 \right] + 2\mathbb{E} \left[ \|F(w^k) - F(z^{k+1/2})\|^2 \right] \\ &\leq 2(\omega + 1)L^2 \mathbb{E} \left[ \|w^k - z^{k+1/2}\|^2 \right] + 2(\omega + 1)D^2, \end{aligned}$$

**Следствие 1.5.** *Предположим, что  $F$  ограниченно-липшицев с константами  $L$  и  $D$ . Тогда Quant-ES*

• *в сильно монотонном случае с  $\gamma \leq \min \left\{ \frac{\sqrt{1-\tau}}{2L\sqrt{2\omega}}; \frac{1-\tau}{4(\mu_F+\mu_h)} \right\}$  удовлетворяет*

$$\begin{aligned} & \mathbb{E} [\tau \|z^{k+1} - z^*\|^2 + \|w^{k+1} - z^*\|^2] \\ & \leq \left( 1 - \gamma \cdot \frac{\mu_F + \mu_h}{16} \right)^{K-1} \cdot (\tau \|z^0 - z^*\|^2 + \|w^0 - z^*\|^2) \\ & \quad + \frac{32\gamma\omega D^2}{\mu_F + \mu_h}, \end{aligned}$$

• *в монотонном случае с  $\gamma \leq \frac{\sqrt{1-\tau}}{2L\sqrt{4\omega+2}}$  удовлетворяет*

$$\mathbb{E} [Gap(\bar{z}^K)] \leq \frac{8 \max_{u \in \mathcal{C}} [\|z^0 - u\|^2]}{\gamma K} + \gamma(9\omega + 2)D^2.$$

Рассмотрим случай  $D = 0$ . Квантизация требуется чтобы сжать информацию, при этом  $\omega$  выступает здесь как коэффициент сжатия, то есть мы передаём в  $\omega$  раз меньше информации чем если бы мы использовали не квантизованный оператор. Однако раз в  $1/(1-\tau)$  итераций (когда мы обновляем  $w^k$ ) необходимо вычислить именно не квантизованный оператор. Основываясь на этом, можно выбрать оптимальное значение для  $\tau$  следующим образом:

$$(1 - \tau) \sim \tau \cdot \frac{1}{\omega} \quad \Rightarrow \quad \tau = \frac{\omega}{\omega + 1}.$$

### ***А.6. Квантизованный экстраградиентный метод с редукцией дисперсии***

Далее мы совмещаем идеи квантизации и редукции дисперсии для случая задачи (1) + (4) с оператором вида конечной суммы.

Положим  $\sigma_k = 0$ . Следующая лемма даёт значения констант для Предположения 2:

---

**Algorithm 6** Квантизованный экстраградиентный метод с редукцией дисперсии

---

**Параметры:** Размер шага  $\gamma$ ,  $K$ .

**Инициализация:** Выбрать  $z^0 = w^0 \in \mathcal{Z}$ .

**for**  $k = 0, 1, 2, \dots, K - 1$

    Выбрать равномерно случайно  $m_k \in 1 \dots M$ ,

$$\bar{z}^k = \tau z^k + (1 - \tau)w^k$$

$$z^{k+1/2} = \text{prox}_{\gamma h}(\bar{z}^k - \gamma F(w^k)),$$

$$z^{k+1} = \text{prox}_{\gamma h}(\bar{z}^k - \gamma Q(F_{m_k}(z^{k+1/2}) - F_{m_k}(w^k)) + \gamma F(w^k)),$$

$$w^{k+1} = \begin{cases} z^{k+1}, & \text{с вероятностью } 1 - \tau \\ w^k, & \text{с вероятностью } \tau \end{cases}$$

**end for**

---

**Лемма 7.** *Предположим, что каждый  $F_{m_k}$  и сам  $F$  ограничено-липшицевы с константами  $L$  и  $D$  (Предположение 3), тогда  $g^k$  и  $g^{k+1}$  из Алгоритма 6 удовлетворяют Предположению 2 с  $A = \omega L^2$ ,  $D_1 = \omega D^2$ ,  $E = 2(\omega + 1)L^2$ ,  $D_3 = 2(\omega + 1)D^2$ .*

**Доказательство:** Легко убедиться, что  $g^{k+1/2}$  несмещённо. Далее,

$$\begin{aligned} \mathbb{E} \left[ \|g^{k+1/2} - g^k\|^2 \right] &= \mathbb{E} \left[ \|Q(F_{m_k}(z^{k+1/2}) - F_{m_k}(w^k)) + F(w^k) - F(w^k)\|^2 \right] \\ &= \mathbb{E} \left[ \|Q(F_{m_k}(z^{k+1/2}) - F_{m_k}(w^k))\|^2 \right] \\ &\leq \omega \mathbb{E} \left[ \|F_{m_k}(z^{k+1/2}) - F_{m_k}(w^k)\|^2 \right] \\ &\leq \omega L^2 \mathbb{E} \left[ \|w^k - z^{k+1/2}\|^2 \right] + \omega D^2, \end{aligned}$$

и наконец,

$$\begin{aligned} &\mathbb{E} \left[ \|g^{k+1/2} - F(z^{k+1/2})\|^2 \right] \\ &= \mathbb{E} \left[ \|Q(F_{m_k}(z^{k+1/2}) - F_{m_k}(w^k)) + F(w^k) - F(z^{k+1/2})\|^2 \right] \end{aligned}$$

$$\begin{aligned}
&\leq 2\mathbb{E} \left[ \|Q(F_{m_k}(z^{k+1/2}) - F_{m_k}(w^k))\|^2 \right] + 2\mathbb{E} \left[ \|F(w^k) - F(z^{k+1/2})\|^2 \right] \\
&\leq 2(\omega + 1)L^2\mathbb{E} \left[ \|w^k - z^{k+1/2}\|^2 \right] + 2(\omega + 1)D^2,
\end{aligned}$$

**Следствие 1.6.** *Предположим, что каждый  $F_{m_k}$  и сам  $F$  ограниченно-липищевы с константами  $L$  и  $D$ . Тогда Квантизованный экстраградиентный метод с редуkcией дисперсии*

- *в сильно монотонном случае с  $\gamma \leq \min \left\{ \frac{\sqrt{1-\tau}}{2L\sqrt{2\omega}}; \frac{1-\tau}{4(\mu_F + \mu_h)} \right\}$  удовлетворяет*

$$\begin{aligned}
&\mathbb{E} \left[ \tau \|z^{k+1} - z^*\|^2 + \|w^{k+1} - z^*\|^2 \right] \\
&\leq \left( 1 - \gamma \cdot \frac{\mu_F + \mu_h}{16} \right)^{K-1} \cdot (\tau \|z^0 - z^*\|^2 + \|w^0 - z^*\|^2) \\
&\quad + \frac{32\gamma\omega D^2}{\mu_F + \mu_h},
\end{aligned}$$

- *в монотонном случае с  $\gamma \leq \frac{\sqrt{1-\tau}}{2L\sqrt{4\omega+2}}$  удовлетворяет*

$$\mathbb{E} \left[ \text{Gap}(\bar{z}^K) \right] \leq \frac{8 \max_{u \in \mathcal{C}} [\|z^0 - u\|^2]}{\gamma K} + \gamma(9\omega + 2)D^2.$$

Оптимальное значение  $\tau$  здесь то же, что и в предыдущем разделе.

### ***А.7. Экстраградиентный метод с сэмплированием по важности***

Здесь мы рассматриваем более общий случай, нежели просто случай конечной суммы. Теперь каждая функция имеет свой вес  $p_m$ . А именно, рассмотрим дискретную случайную переменную  $\eta$  вида

$$\mathbb{P}(\eta = m) = p_m, \quad \sum_{m=1}^M p_m = 1.$$

На каждом шаге, мы обращаемся к  $F_\eta$ . Веса/вероятности  $p_m$  могут быть заданы априори, или же мы можем задать их самостоятельно: например, имеет смысл выбрать  $p_m = \frac{L_m}{\sum_m L_m}$ .

---

**Algorithm 7** Экстраградиентный метод с сэмплированием по важности

---

**Параметры:** Размер шага  $\gamma$ ,  $K$ .

**Инициализация:** Выбрать  $z^0 = w^0 \in \mathcal{Z}$ .

**for**  $k = 0, 1, 2, \dots, K - 1$

    Выбрать  $\eta_k$ ,

$$\bar{z}^k = \tau z^k + (1 - \tau)w^k$$

$$z^{k+1/2} = \text{prox}_{\gamma h}(\bar{z}^k - \gamma F(w^k)),$$

$$z^{k+1} = \text{prox}_{\gamma h}(\bar{z}^k - \gamma \cdot \frac{1}{p_{\eta_k}} \cdot (F_{\eta_k}(z^{k+1/2}) - F_{\eta_k}(w^k)) + \gamma F(w^k)),$$

$$w^{k+1} = \begin{cases} z^{k+1}, & \text{с вероятностью } 1 - \tau \\ w^k, & \text{с вероятностью } \tau \end{cases}$$

**end for**

---

**Лемма 8.** Предположим, что каждый  $F_m$  ограниченно-липшицев с константами  $L_m$  и  $D_m$  (Предположение 3), как и  $F$ , соответственно с константами  $L$  и  $D$ , тогда  $g^k$  и  $g^{k+1}$  из Алгоритма 7 удовлетворяют Предположению 2 с константами  $A = \sum_{m=1}^M \frac{L_m^2}{p_m}$ ,  $D_1 = \sum_{m=1}^M \frac{D_m^2}{p_m}$ ,  $E = 2 \left( \sum_{m=1}^M \frac{L_m^2}{p_m} + L^2 \right)$ ,  $D_3 = 2 \left( \sum_{m=1}^M \frac{D_m^2}{p_m} + D^2 \right)$ .

**Доказательство:** Легко убедиться, что  $g^{k+1/2}$  несмещённо. Далее,

$$\begin{aligned} & \mathbb{E} \left[ \|g^{k+1/2} - g^k\|^2 \right] \\ &= \mathbb{E} \left[ \left\| \frac{1}{p_{\eta_k}} (F_{\eta_k}(z^{k+1/2}) - F_{\eta_k}(w^k)) + F(w^k) - F(w^k) \right\|^2 \right] \\ &= \mathbb{E} \left[ \left\| \frac{1}{p_{\eta_k}} (F_{\eta_k}(z^{k+1/2}) - F_{\eta_k}(w^k)) \right\|^2 \right] \\ &= \mathbb{E} \sum_{m=1}^M \frac{1}{p_m} \left[ \|F_m(z^{k+1/2}) - F_m(w^k)\|^2 \right] \\ &\leq \sum_{m=1}^M \frac{L_m^2}{p_m} \mathbb{E} \left[ \|z^{k+1/2} - w^k\|^2 \right] + \sum_{m=1}^M \frac{D_m^2}{p_m}, \end{aligned}$$

и наконец,



$$\begin{aligned}
& \mathbb{E} \left[ \|g^{k+1/2} - F(z^{k+1/2})\|^2 \right] \\
&= \mathbb{E} \left[ \left\| \frac{1}{p_{\eta_k}} (F_{m_k}(z^{k+1/2}) - F_{m_k}(w^k)) + F(w^k) - F(z^{k+1/2}) \right\|^2 \right] \\
&\leq 2\mathbb{E} \left[ \left\| \frac{1}{p_{\eta_k}} (F_{m_k}(z^{k+1/2}) - F_{m_k}(w^k)) \right\|^2 \right] + 2\mathbb{E} \left[ \|F(w^k) - F(z^{k+1/2})\|^2 \right] \\
&\leq 2 \left( \sum_{m=1}^M \frac{L_m^2}{p_m} + L^2 \right) \mathbb{E} \left[ \|w^k - z^{k+1/2}\|^2 \right] + 2 \left( \sum_{m=1}^M \frac{D_m^2}{p_m} + D^2 \right),
\end{aligned}$$

**Следствие 1.7.** *Предположим, что каждый  $F_m$  ограниченно-липшицев с константами  $L_m$  и  $D_m$  (Предположение 3) и  $F$  с константами  $L$  и  $D$ .*

*Тогда Экстраградиентный метод с сэмплированием по важности*

• *в сильно монотонном случае с  $\gamma \leq \min \left\{ \frac{\sqrt{1-\tau}}{2\sqrt{2} \sum_{m=1}^M \frac{L_m^2}{p_m}}; \frac{1-\tau}{4(\mu_F + \mu_h)} \right\}$  удовлетворяет*

$$\begin{aligned}
& \mathbb{E} \left[ \tau \|z^{k+1} - z^*\|^2 + \|w^{k+1} - z^*\|^2 \right] \\
&\leq \left( 1 - \gamma \cdot \frac{\mu_F + \mu_h}{16} \right)^{K-1} \cdot (\tau \|z^0 - z^*\|^2 + \|w^0 - z^*\|^2) \\
&\quad + \frac{32\gamma}{\mu_F + \mu_h} \cdot \sum_{m=1}^M \frac{D_m^2}{p_m},
\end{aligned}$$

• *в монотонном случае с  $\gamma \leq \frac{\sqrt{1-\tau}}{2\sqrt{2L^2+4} \sum_{m=1}^M \frac{L_m^2}{p_m}}$  удовлетворяет*

$$\mathbb{E} \left[ \text{Gap}(\bar{z}^K) \right] \leq \frac{8 \max_{u \in \mathcal{C}} [\|z^0 - u\|^2]}{\gamma K} + \gamma \left( 9 \sum_{m=1}^M \frac{D_m^2}{p_m} + 2D^2 \right).$$

### A.8. Локальный экстраградиентный метод

Этот метод предназначен для распределённой задачи (1)+(5). Суть метода заключается в переключении между локальными итерациями и усреднением со значением на сервере. С вероятностью  $\tau$  производится локальный шаг, с вероятностью  $1 - \tau$  – шаг коммуникации.

Здесь  $Z_{\text{avg}} = [\bar{z}^T, \dots, \bar{z}^T] \in \mathbb{R}^{Md}$  с  $\bar{z} = \frac{1}{M} \sum_{m=1}^M z_m$  (и то же для  $W_{\text{avg}}$ ).

---

**Algorithm 8** Рандомизированный локальный экстраградиентный метод

---

**Параметры:** Размер шага  $\gamma$ ,  $K$ , вероятность  $p$ .

**Инициализация:** Выбрать  $z^0 = w^0 \in \mathcal{Z}$ ,  $z_m^0 = z^0$  для всех  $m$ .

**for**  $k = 0, 1, 2, \dots$

$$\bar{Z}^k = \tau Z^k + (1 - \tau)W^k,$$

$$Z^{k+1/2} = \text{prox}_{\gamma h}(\bar{Z}^k - \gamma \cdot (\Phi(W^k) + \lambda \cdot (W^k - W_{\text{avg}}^k))),$$

$$G(Z) = \begin{cases} \frac{1}{\tau} \cdot \Phi(Z), & \text{с вероятностью } \tau \\ \frac{1}{1-\tau} \cdot \lambda \cdot (Z - Z_{\text{avg}}), & \text{с вероятностью } 1 - \tau \end{cases},$$

$$Z^{k+1} = \text{prox}_{\gamma h}(\bar{Z}^k - \gamma \cdot (G(Z^{k+1/2}) - G(W^k) + \Phi(W^k) + \lambda \cdot (W^k - W_{\text{avg}}^k))),$$

$$W^{k+1} = \begin{cases} Z^{k+1}, & \text{с вероятностью } 1 - \tau \\ W^k, & \text{с вероятностью } \tau \end{cases}$$

**end for**

---

На самом деле, анализ для этого метода уже был проделан в предыдущем разделе. Действительно, здесь мы имеем два оператора:  $\Phi(Z)$  и  $\lambda \cdot (Z - Z_{\text{avg}})$ , являющихся  $L$  и  $\lambda$  гладкими, соответственно.

Если выбрать  $\tau = \frac{L}{L+\lambda}$ , то верно

**Следствие 1.8.** *Рандомизированный локальный экстраградиентный метод*

- в сильно монотонном случае с  $\gamma \leq \min \left\{ \frac{\sqrt{\lambda}}{2\sqrt{2}(L+\lambda)^{3/2}}; \frac{\sqrt{\lambda}}{4(\mu_F + \mu_h)\sqrt{L+\lambda}} \right\}$

удовлетворяет

$$\begin{aligned} & \mathbb{E} [\tau \|z^{k+1} - z^*\|^2 + \|w^{k+1} - z^*\|^2] \\ & \leq \left( 1 - \gamma \cdot \frac{\mu_F + \mu_h}{16} \right)^{K-1} \cdot (\tau \|z^0 - z^*\|^2 + \|w^0 - z^*\|^2) \\ & \quad + \frac{32\gamma}{\mu_F + \mu_h} \cdot \sum_{m=1}^M \frac{D_m^2}{p_m}, \end{aligned}$$

- в монотонном случае с  $\gamma \leq \frac{\sqrt{\lambda}}{2\sqrt{6}(L+\lambda)^{3/2}}$  удовлетворяет

$$\mathbb{E} [Gap(\bar{z}^K)] \leq \frac{8 \max_{u \in \mathcal{C}} [\|z^0 - u\|^2]}{\gamma K} + \gamma \left( 9 \sum_{m=1}^M \frac{D_m^2}{p_m} + 2D^2 \right).$$

Отсюда, можно получить оценку для числа локальных шагов и числа коммуникаций:

- в сильно монотонном случае:

$$\mathcal{O} \left( \frac{\sqrt{\lambda(\lambda + L)}}{\mu} \log \frac{1}{\varepsilon} \right) \text{ комм.} \quad \text{и} \quad \mathcal{O} \left( \frac{\sqrt{L(\lambda + L)}}{\mu} \log \frac{1}{\varepsilon} \right) \text{ локальных шагов.}$$

- в монотонном случае

$$\mathcal{O} \left( \frac{\sqrt{\lambda(\lambda + L)} \cdot \Omega^2}{\varepsilon} \right) \text{ комм.} \quad \text{и} \quad \mathcal{O} \left( \frac{\sqrt{L(\lambda + L)} \cdot \Omega^2}{\varepsilon} \right) \text{ локальных шагов.}$$

Отметим, что эти оценки являются достаточно хорошими при малом значении  $\lambda$ .

## В. Эксперименты

### В.1. Генеративно-состязательные сети

В качестве примера оптимизации мини-максного целевого функционала при различных подходах оптимизации была предложена задача оптимизации генеративно-состязательных сетей (GAN). GAN — это структура для оценки генеративных моделей с помощью состязательного процесса, в котором мы одновременно обучаем две модели: генеративную модель  $G$ , которая фиксирует распределение данных, и дискриминативную модель  $D$ , которая оценивает вероятность того, что пример пришел из обучающей выборки, а не из  $G$ .  $D(G(z))$  это вероятность (скалярная), что выход генератора  $G$  это реальное изображение. Согласно статье Гудфеллоу [6], в

этой задаче  $D$  пытается максимизировать вероятность того, что он правильно классифицирует реальные и фейковые изображения ( $\log(D(x))$ ), а  $G$  пытается минимизировать вероятность того, что  $D$  будет предсказывать его выходные данные как фейковые ( $\log(1 - D(G(z)))$ ). Из статьи функция потерь GAN

$$\min_{\theta_G} \max_{\theta_D} \mathbb{E}_{x \sim \mathbb{P}_{data}(x)} [\log(D(x))] + \mathbb{E}_{z \sim \mathbb{P}_z(z)} [\log(1 - D(z))]$$

Целью наших экспериментов не было получение наилучших результатов для генеративных сетей или подходом к формулировке задачи для GAN. В наших экспериментах мы подтверждаем работоспособность SVRG, а также методов квантизации для обучения GAN.

**Данные, модель и оптимизаторы.** Для проведения экспериментов было предложено использовать датасет CIFAR. Он содержит 50000 и 10000 изображений в обучающей и валидационной выборках соответственно, равномерно распределенных по 10 классам. Для каждого оптимизационного подхода в качестве оптимизатора использовался Adam. В ходе проведения эксперимента, был проведен анализ гиперпараметров каждого из оптимизаторов(генератора и дискриминатора):

$\beta_1$	.5	.9	.99
$\beta_2$	.99	.999	
скорость обучения	$.9 \cdot 10^{-4}$	$1 \cdot 10^{-4}$	$2 \cdot 10^{-4}$

Таблица 1. Набор анализируемых гиперпараметров

По результатам проведенного анализа сходимости (проверялись результаты работы после 100 эпох обучения), были выбраны следующие параметры:  $\beta_1 = .9$ ,  $\beta_2 = .999$  и коэффициент скорости обучения =  $2 \cdot 10^{-4}$  для

обоих методов оптимизации. Размер набора элементов, на котором производилось одна итерация вычисления градиента, составлял 64.

Для проверки качества выдаваемых изображений, была использована метрика Inception score [33].

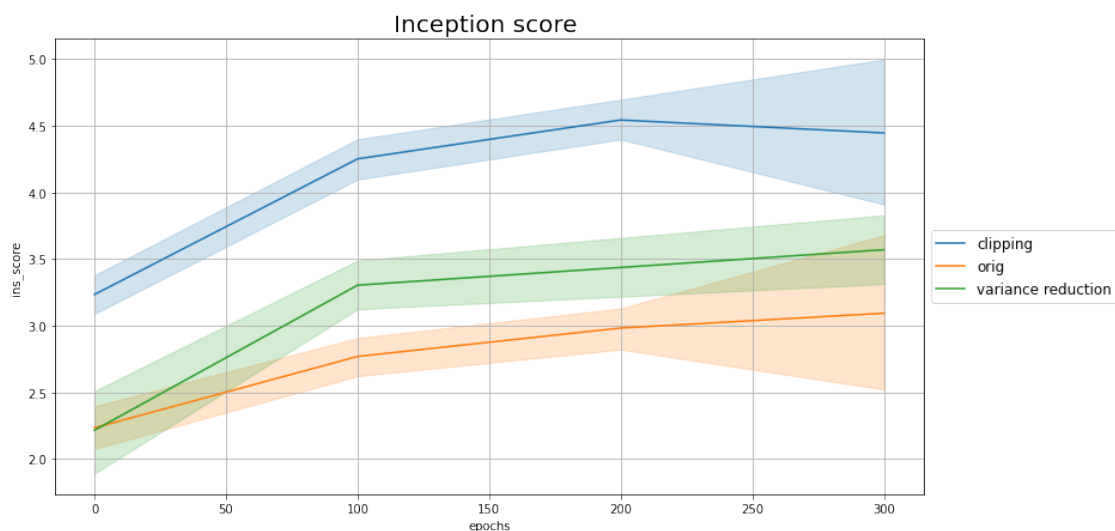
$$\text{IS}(G) = \exp(\mathbb{E}_{x \sim p_g} D_{KL}(p(y|\mathbf{x})||p(y)))$$

где  $x \sim p_g$  означает что изображение  $\mathbf{x}$  сгенерировано из распределения  $p_g$ ,  $D_{KL}(p||q)$  – расстояние Кульбака-Лейбнера между двумя распределениями  $p, q$ .

Данная метрика позволяет автоматически оценивать качество картинок, генерируемых моделью. в ходе исследования было показано, что данная метрика хорошо коррелирует с результатами оценки человека на сгенерированных изображениях CIFAR10. данная метрика использует нейронную сеть Inception v3, предобученную на датасете ImageNet, и собирает статистику выходных данных работы сети, примененной к сгенерированным изображениям. Для экспериментов было предложено использовать архитектуру DCGan [34] с подходом Conditional [35]. Особенность данной архитектуры заключается в том что мы можем обучить модель сэмплировать изображения из определенного распределения, сэмплы из которого будут схожи с элементами из желаемого распределения обучающей выборки.

Основной целью этого эксперимента была возможность не переобучать определенные архитектуры под разные подходы оптимизации. В процессе обучения генератор и дискриминатор имели одинаковое количество шагов оптимизации.

**Результаты.** Посмотрите на рисунки представленные на фигурах 1 и 2. Согласно результатам, полученным в рамках экспериментов, можно заметить, что подходы, предложенные в данной статье, позволяют достигнуть



Фиг. 1. Результаты работы модели, полученных при ее потимизации с помощью метода Adam разными способами: подход с использованием квантизации/клиппинга 70% всех градиентов модели, подход с использованием редукции дисперсии, оригинальный подход [6]

лучших результатов, по сравнению с оригинальными.

- Метод редукции дисперсии позволял оптимизировать функционал точнее, чем оригинальный способ, на протяжении всей эпохи оптимизации, однако, приближаясь к конечным итерациям каждой эпохи обучения, полный градиент достаточно сильно отличался от направления градиента, получаемого на последних итерациях, что вносило неточность при оптимизации, что привело к переобучению генератора.

- Метод с использованием квантизации/клиппинга (случайным образом обнуляется 70% от всего градиента модели как в генераторе так и в дискриминаторе, что соответствует оператору сжатия  $\text{Rand70\%}$ ), позволил предотвратить переобучение генератора и дискриминатора на ранних стадиях, что в свою очередь позволило достичь лучших результатов.

**Подводя итоги**, все вышеупомянутые подходы позволили получить результаты лучше полученных при оригинальном подходе обучения гене-



Фиг. 2. Эти изображения были сгенерированы архитектурой conditional DCGan, которая обучалась с использованием разных подходов: (а)- исходная оптимизация с использованием Adam 1, (б) оптимизация с занулением 70% всех значений градиентов, (с) оптимизация со стохастической редукцией дисперсии градиентного спуска.

ративных моделей и могут быть применены для оптимизации GAN, но квантизация позволяет быстро получить наилучшие результаты при сохранении вычислительных ресурсов.



## В.2. Полицейский vs. Грабитель

Рассмотрим задачу Полицейский vs. Грабитель, чтобы сравнить производительность некоторых методов, представленных в нашей статье. В этой задаче мы рассматриваем квадратный город размером 200 на 200 клеток. В каждом городе есть дом и будка милиции. Грабитель выбирает один дом для ограбления, полицейский выбирает будку, в которой будет дежурить. Задача заключается в нахождении оптимальной смешанной стратегии для противоборствующих игроков: грабителя и полицейского в игре, где цель грабителя – атаковать выбранный дом  $i$  с максимальным благосостоянием  $w_i$ , а цель полицейского – выбрать оптимальный пост и поймать грабителя, чтобы предотвратить нанесенный им максимальный ожидаемый убыток. Мы предполагаем, что вероятность поймать вора для определенного дома  $i$  и поста  $j$  равна  $\exp(-\theta \cdot d(i, j))$  для функции расстояния  $d$ , которая вводится выше. Мы можем сформулировать описанную постановку как задачу о поиски билинейной седловой точки:

$$\min_{x \in \Delta(n^2)} \max_{y \in \Delta(n^2)} f(x, y) := \frac{1}{n} \sum_{k=1}^n y^\top A^{(k)} x, \quad (\text{П.19})$$

для  $x$  и  $y$ , которые являются векторами вероятности выбора какого-либо дома и поста соответственно, и для матриц

$$A_{ij}^{(k)} = w_i^{(k)} \cdot (1 - \exp(-\theta \cdot d(i, j))),$$

где благосостояние  $w^{(k)}$  и функция расстояния  $d$  определяются следующим образом (эти выражения легко понять, если представить  $i$  как сплюсненную координату на игровом поле размера  $n \times n$ ,  $i(x, y) = x \cdot n + y$ , график  $w$  представляет собой пирамиду с центром в центре этого поля и  $d$  евклидово расстояние на ней):

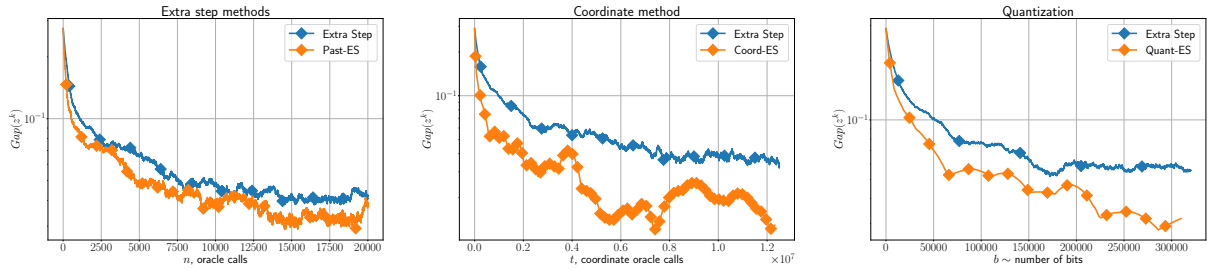
$$w_i = 1 - \frac{2}{n} \cdot \min \{ |[i/n] - n/2|, |i \bmod n - n/2| \},$$



$$w_i^{(k)} = w_i \cdot (1 + \xi^{(k)}) \text{ for } \xi^{(k)} \sim \mathcal{U}(0, \sigma),$$

$$d(i, j) = \sqrt{([\![i/n] - [j/n]\!]^2 + (i \bmod n - j \bmod n)^2)}.$$

Для экспериментов были выбраны параметры  $\theta = 0.6, n = 25, \sigma = 3$ . Coord-ES, Quant-ES и Past-ES методы использует такой же значение  $\gamma$ , как Extra Step метод, которое является оптимальным для последнего. Мы сравниваем Past метод по количеству обращений к оракулу  $F$ , метод с квантизацией по количеству используемых бит, координатный метод по количеству используемых координат.



(a) Extra Step and Past-ES

(b) Extra Step and Coord-ES

(c) Extra Step and Quant-ES

Фиг. 3. Сходимость методов с дополнительным шагом для задачи Полицейский vs. Грабитель (П.19).