ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2023, том 63, № 2, с. 189–217

ОБЩИЕ ЧИСЛЕННЫЕ МЕТОДЫ

УДК 519.85

УНИФИЦИРОВАННЫЙ АНАЛИЗ МЕТОДОВ РЕШЕНИЯ ВАРИАЦИОННЫХ НЕРАВЕНСТВ: РЕДУКЦИЯ ДИСПЕРСИИ, СЭМПЛИРОВАНИЕ, КВАНТИЗАЦИЯ И ПОКОМПОНЕНТНЫЙ СПУСК[1)]

© 2023 г. А. Н. Безносиков[1], А. В. Гасников[1,2,3,*], К. Э. Зайнуллина[1], А. Ю. Масловский[1], Д. А. Пасечнюк[1,2]

[1] 141701 Долгопрудный, М.о., Институтский пер., 9, Московский физико-технический институт (национальный исследовательский университет), Россия

[2] 127051 Москва, Большой Каретный пер., 19, стр. 1, Институт проблем передачи информации им. А.А. Харкевича, Россия

[3] 385000 Республика Адыгея, Майкоп, ул. Первомайская, 208, Кавказский математический центр Адыгейского государственного университета, Россия

*****e-mail: gasnikov@yandex.ru*

Поступила в редакцию 28.01.2022 г.
Переработанный вариант 28.01.2022 г.
Принята к публикации 14.10.2022 г.

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 экстраградиентного метода, являющегося стандартным для решения вариационных 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 др. 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.

DOI: 10.31857/S0044466923020059, **EDN:** BMKMVF

1. ВВЕДЕНИЕ

Основная постановка задачи, рассматриваемой в настоящей статье, - решение вариационного неравенства (ВН) – имеет следующий вид:

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

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

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

[1)]Работа выполнена при финансовой поддержке Минобрнауки РФ (госзадание) № 075-00337-20-03 (проект 0714-2020-0005).

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 минимаксную задачу

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

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, причем решение (x^*, y^*) вариационного неравенства тогда является глобальной седловой точкой:

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

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. Это связано со сложностью задачи ВН: многие техники из минимизации еще 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 (см. [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 (см. [2], [3]), а также применение таких подходов к задачам обработки изображений (см. [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 вид $f(x, y) = \mathbb{E}_{p_x\sim D_x, p_y\sim D_y}\left[f_{p_x,p_y}(x, y)\right]$. 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, интерпретируем сказанное выше следующим образом: предположим, что мы не имеем доступа к “честному” $F(z)$, а только к некоторой несмещенной стохастической оценке $F(z,\xi)$:

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

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

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

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 предпочтем переписать ее следующим образом:

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

где векторы $\Phi(Z) = [F_1^T(z_1),\dots,F_M^T(z_M)] \in \mathbb{R}^{Md}$, $Z = [z_1^T,\dots,z_M^T] \in \mathbb{R}^{Md}$ 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 $\min_{x_1,\dots x_M}\left[\sum_{m=1}^M f_m(x_m) + \right.$ $\left. + \frac{\lambda}{2}\sum_{m=1}^M \|x_m - \bar{x}\|^2\right]$ 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 λ достаточно большим, чтобы решения задач (4) и (5) были примерно одинаковыми: $z_m \approx \bar{z}$. Но имеет ли смысл брать λ малым (например, когда $\lambda = 0$, мы имеем просто локальные модели)? Оказывается, да: такая постановка задачи с переменной λ, значение которой может варьироваться, породила целую тенденцию в персонализированном федеративном обучении – федеративное обучение со смешиванием (см. [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 (см. [9]). В простом виде методы такого рода записываются следующим образом:

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

где $\mathrm{prox}_{\gamma h}(z) = \arg\min_x\left\{\gamma h(x) + \frac{1}{2}\|z - x\|^2\right\}$. В классическом, детерминированном, варианте $g^k = F(z^k)$ и $g^{k+1/2} = F(z^{k+1/2})$. Этот метод оптимален с точностью до численных констант для гладких монотонных и сильно монотонных вариационных неравенств (см. [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 настоящей работе для нашего унифицированного анализа мы используем метод (6) с несколько более сложной структурой:

$$\bar{z}^k = \tau z^k + (1-\tau)w^k, \quad z^{k+1/2} = \mathrm{prox}_{\gamma h}(\bar{z}^k - \gamma g^k), \quad z^{k+1} = \mathrm{prox}_{\gamma h}(\bar{z}^k - \gamma g^{k+1/2}), \tag{7}$$

где $w^{k+1} = z^{k+1}$ с вероятностью $(1-\tau)$ или w^k – иначе. Видно, 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

Метод SGD используется для решения задач минимизации с середины прошлого века (см. [12]) и с того времени был расширен огромным количеством различных модификаций. Это методы уменьшения дисперсии (см. [13]), квантования (см. [14]), координатные методы (см. [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 (см. [9]); еще более общая его версия называется Mirror Prox (см. [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 (см. [17]) и с повторным вызовом (см. [18]). Можно также причислить классические методы, отличающиеся по структуре от экстраградиентного из [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 (см. [23], [24]), а также работы по безградиентным методам (см. [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 (см. [26], [27]), как это сделано для задач минимизации. В настоящей работе мы обращаем внимание не на детерминированные методы типа Local SGD, а на рандомизированные, которые могут быть применены для решения задачи (5), например, как описанные в [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 настоящей 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 [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.

Пять новых методов. Более важным, чем обобщение других результатов, является получение новых надежных методов. В отличие от [16], к моменту написания которой большинство анализируемых там методов 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 ℓ_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* μ_F, μ_h *такие, что* $\mu_h + \mu_F > 0$ *и следующие неравенства верны для всех* $z_1, z_2 \in \mathcal{Z}$*:*

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

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

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

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

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* $g^{k+1/2}$ *не смещены для всех* k*:*

$$\mathbb{E}\left[g^{k+1/2} \middle| z^{k+1/2}\right] = F(z^{k+1/2}). \tag{10}$$

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:

$$\begin{aligned}
\mathbb{E}\left[\left\|g^{k+1/2} - g^k\right\|^2\right] &\leq A\mathbb{E}\left[\left\|z^{k+1/2} - w^k\right\|^2\right] + B\mathbb{E}\left[\sigma_k^2\right] + D_1,\\
\mathbb{E}\left[\sigma_{k+1}^2\right] &\leq (1-\rho)\mathbb{E}\left[\sigma_k^2\right] + C\mathbb{E}\left[\left\|z^{k+1/2} - w^k\right\|^2\right] + D_2,\\
\mathbb{E}\left[\left\|g^{k+1/2} - F(z^{k+1/2})\right\|^2\right] &\leq E\mathbb{E}\left[\left\|z^{k+1/2} - w^k\right\|^2\right] + D_3.
\end{aligned} \tag{11}$$

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:

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

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

Для монотонного случая используется другой критерий – функция зазора (см. [3], [10]):

$$\mathrm{Gap}(z) = \max_{u \in \mathscr{C}} \left[\langle F(u), z - u\rangle + h(z) - h(u)\right].$$

Здесь максимум берется не по всему множеству $\mathscr{Z}$ (как в классической версии), а по $\mathscr{C}$ – компактному подмножеству множества $\mathscr{Z}$. Таким образом, мы можем рассматривать неограниченные множества $\mathscr{Z}$. Это допустимо, так как такой вариант критерия верен, если решение z^* лежит в $\mathscr{C}$ (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* $\gamma \le \min\left\{\dfrac{\sqrt{1-\tau}}{2\sqrt{2A + TC}}; \dfrac{1-\tau}{4(\mu_F + \mu_h)}\right\}$ *и* $T \ge \dfrac{4B}{\rho}$:

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

- *для монотонного/выпуклого случаев с* $\gamma \le \dfrac{\sqrt{1-\tau}}{2\sqrt{2A + TC + E}}$ *и* $T \ge \dfrac{2B}{\rho}$

$$\mathbb{E}\left[\mathrm{Gap}(\overline{z}^K)\right] \le \frac{8\max\limits_{u \in \mathscr{C}}\left[\left\|z^0 - u\right\|^2\right] + 4\gamma^2 T\sigma_0^2}{\gamma K} + \gamma(7D_1 + 3TD_2 + D_3),$$

где $\overline{z}^K = \dfrac{1}{K}\sum_{k=0}^{K-1} z^{k+1/2}$.

Метод (7) в условиях предположения 2 сходится линейно в сильно монотонном случае и сублинейно ($\sim 1/K$) 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 γ (см. [28]). Для некоторых методов мы используем это, чтобы получить классические оценки сходимости (см. подробнее об этом в п. 2.3). 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 μ_h *– сильно выпуклая и* $z^+ = \mathrm{prox}_{\gamma h}(z)$. *Тогда для всех* $x \in \mathbb{R}^d$ *справедливо следующее неравенство:*

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

Доказательство. Мы используем $\gamma\mu$-сильно выпуклость функции γh (9):

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

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

Доказательство теоремы 1. По лемме 1 для $z^{k+1/2} = \mathrm{prox}_{\gamma h}(\overline{z}^k - \gamma g^k)$ и $z^{k+1} = \mathrm{prox}_{\gamma h}(\overline{z}^k - \gamma g^{k+1/2})$ для $x = u$ получаем

$$\left\langle z^{k+1} - \overline{z}^k + \gamma g^{k+1/2}, u - z^{k+1}\right\rangle \ge \gamma\left(h(z^{k+1}) - h(u) + \frac{\mu_h}{2}\left\|z^{k+1} - u\right\|^2\right),$$

$$\left\langle z^{k+1/2} - \overline{z}^k + \gamma g^k, z^{k+1} - z^{k+1/2}\right\rangle \ge \gamma\left(h(z^{k+1/2}) - h(z^{k+1}) + \frac{\mu_h}{2}\left\|z^{k+1} - z^{k+1/2}\right\|^2\right).$$

Далее суммируем два неравенства и проводим некоторые перестановки:

$$\left\langle z^{k+1} - \overline{z}^k, u - z^{k+1}\right\rangle + \left\langle z^{k+1/2} - \overline{z}^k, z^{k+1} - z^{k+1/2}\right\rangle + \gamma\left\langle g^{k+1/2} - g^k, z^{k+1/2} - z^{k+1}\right\rangle +$$
$$+ \gamma\left\langle g^{k+1/2}, u - z^{k+1/2}\right\rangle \geq \gamma\left(h(z^{k+1/2}) - h(u) + \frac{\mu_h}{2}\left\|z^{k+1} - z^{k+1/2}\right\|^2 + \frac{\mu_h}{2}\left\|z^{k+1} - u\right\|^2\right).$$

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

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

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

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

Небольшая перестановка дает

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

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

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

Далее рассмотрим разные случаи теоремы. Начнем с сильно монотонного/выпуклого случая. Заменим $u = z^*$, возьмем полное математическое ожидание и получим

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

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

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

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

Свойство решения (1) дает

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

И по предположению 1 в сильно монотонном случае получим

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

С другой стороны,

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

Суммируя два предыдущих неравенства, имеем

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

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

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

С предположением 2 для σ_{k+1} получаем

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

Используем $-\left\|z^{k+1}-z^*\right\|^2 \le -\frac{1}{2}\left\|z^{k+1/2}-z^*\right\|^2 + \left\|z^{k+1}-z^{k+1/2}\right\|^2$, что дает

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

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

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

Далее работаем с предпоследней строкой (13):

$$\begin{aligned}-\gamma\left(\mu_F+\frac{\mu_h}{4}\right)\tau\mathbb{E}\left[\left\|z^k-z^*\right\|^2\right] - \gamma\left(\mu_F+\frac{\mu_h}{4}\right)(1-\tau)\mathbb{E}\left[\left\|w^k-z^*\right\|^2\right] = -\frac{\gamma}{2}\left(\mu_F+\frac{\mu_h}{4}\right)\tau\mathbb{E}\left[\left\|z^k-z^*\right\|^2\right] - \\ - \frac{\gamma}{2}\left(\mu_F+\frac{\mu_h}{4}\right)\tau\mathbb{E}\left[\left\|z^k-z^*\right\|^2\right] - \gamma\left(\mu_F+\frac{\mu_h}{4}\right)(1-\tau)\mathbb{E}\left[\left\|w^k-z^*\right\|^2\right] - \le -\frac{\gamma}{2}\left(\mu_F+\frac{\mu_h}{4}\right)\tau\mathbb{E}\left[\left\|z^k-z^*\right\|^2\right] - \\ - \frac{\gamma}{4}\left(\mu_F+\frac{\mu_h}{4}\right)\tau\mathbb{E}\left[\left\|w^k-z^*\right\|^2\right] + \frac{\gamma}{2}\left(\mu_F+\frac{\mu_h}{4}\right)\tau\mathbb{E}\left[\left\|z^k-w^k\right\|^2\right] - \gamma\left(\mu_F+\frac{\mu_h}{4}\right)(1-\tau)\mathbb{E}\left[\left\|w^k-z^*\right\|^2\right] \le \\ \le -\frac{\gamma}{4}\left(\mu_F+\frac{\mu_h}{4}\right)\tau\mathbb{E}\left[\left\|z^k-z^*\right\|^2\right] - \frac{\gamma}{4}\left(\mu_F+\frac{\mu_h}{4}\right)\mathbb{E}\left[\left\|w^k-z^*\right\|^2\right] + \frac{\gamma}{2}\left(\mu_F+\frac{\mu_h}{4}\right)\tau\mathbb{E}\left[\left\|z^k-w^k\right\|^2\right] \le \\ \le -\frac{\gamma}{4}\left(\mu_F+\frac{\mu_h}{4}\right)\tau\mathbb{E}\left[\left\|z^k-z^*\right\|^2\right] - \frac{\gamma}{4}\left(\mu_F+\frac{\mu_h}{4}\right)\mathbb{E}\left[\left\|w^k-z^*\right\|^2\right] + \gamma\left(\mu_F+\frac{\mu_h}{4}\right)\tau\mathbb{E}\left[\left\|z^{k+1/2}-z^k\right\|^2\right] + \\ + \gamma\left(\mu_F+\frac{\mu_h}{4}\right)\tau\mathbb{E}\left[\left\|z^{k+1/2}-w^k\right\|^2\right].\end{aligned} \tag{14}$$

Подставив (14) в (13), получим

$$\begin{aligned}\mathbb{E}[V_{k+1}] \le \tau\mathbb{E}\left[\left\|z^k-z^*\right\|^2\right] + \mathbb{E}\left[\left\|w^k-z^*\right\|^2\right] + \left(1-\rho+\frac{2B}{T}\right)\mathbb{E}\left[\gamma^2 T\sigma_k^2\right] - \\ - \left((1-\tau)-2\gamma^2 A-\gamma^2 TC-\gamma\left(2\mu_F+\frac{\mu_h}{2}\right)\right)\mathbb{E}\left[\left\|z^{k+1/2}-w^k\right\|^2\right] - \left(\frac{1}{2}+\gamma\mu_h\right)\mathbb{E}\left[\left\|z^{k+1}-z^{k+1/2}\right\|^2\right] - \\ - \left(1-3\gamma\left(\mu_F+\frac{\mu_h}{4}\right)\right)\tau\mathbb{E}\left[\left\|z^{k+1/2}-z^k\right\|^2\right] - \frac{\gamma}{4}\left(\mu_F+\frac{\mu_h}{4}\right)\tau\mathbb{E}\left[\left\|z^k-z^*\right\|^2\right] - \\ - \frac{\gamma}{4}\left(\mu_F+\frac{\mu_h}{4}\right)\mathbb{E}\left[\left\|w^k-z^*\right\|^2\right] + \gamma^2(2D_1+TD_2).\end{aligned} \tag{15}$$

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

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

В результате

$$\mathbb{E}[V_{k+1}] \le \max\left\{\left(1-\gamma\frac{\mu_F+\mu_h}{16}\right);\left(1-\frac{\rho}{2}\right)\right\}\mathbb{E}[V_k]+\gamma^2(2D_1+TD_2).$$

Выполнение рекурсивных переходов завершает доказательство.

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

$$2\gamma\mathrm{gap}(z^{k+1/2},u) + \left\|z^{k+1}-u\right\|^2 \le \tau\left\|z^k-u\right\|^2 + (1-\tau)\left\|w^k-u\right\|^2 - \tau\left\|z^{k+1/2}-z^k\right\|^2 -$$
$$-(1-\tau)\left\|z^{k+1/2}-w^k\right\|^2 + 2\gamma^2\left\|g^{k+1/2}-g^k\right\|^2 - 2\gamma\left\langle g^{k+1/2}-F(z^{k+1/2}), z^{k+1/2}-u\right\rangle.$$

Добавив к обеим частям $\left\|w^{k+1}-u\right\|^2$ и проведя некоторые перестановки, получим

$$2\gamma\mathrm{gap}(z^{k+1/2},u) \le \left[\tau\left\|z^k-u\right\|^2 + \left\|w^k-u\right\|^2\right] - \left[\tau\left\|z^{k+1}-u\right\|^2 + \left\|w^{k+1}-u\right\|^2\right] - \tau\left\|w^k-u\right\|^2 -$$
$$-(1-\tau)\left\|z^{k+1}-u\right\|^2 + \left\|w^{k+1}-u\right\|^2 - \tau\left\|z^{k+1/2}-z^k\right\|^2 - (1-\tau)\left\|z^{k+1/2}-w^k\right\|^2 +$$
$$+2\gamma^2\left\|g^{k+1/2}-g^k\right\|^2 - 2\gamma\left\langle g^{k+1/2}-F(z^{k+1/2}), z^{k+1/2}-u\right\rangle.$$

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

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

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

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

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

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

$$+ (1-\tau-2\gamma^2 A)\mathbb{E}\left[\left\|z^{k+1/2}-w^k\right\|^2\right] + 2\gamma^2 KD_1 + \sum_{k=0}^{K-1}\left[2\gamma^2 B\mathbb{E}\left[\sigma_k^2\right]\right] +$$
$$+ 2\gamma\mathbb{E}\left[\max_{u\in\mathscr{C}}\sum_{k=0}^{K-1}\left[\left\langle g^{k+1/2}-F(z^{k+1/2}),u-z^{k+1/2}\right\rangle\right]\right] \le \max_{u\in\mathscr{C}}\left[\tau\left\|z^0-u\right\|^2+\left\|w^0-u\right\|^2\right] +$$
$$+ \sum_{k=0}^{K-1}\left[\gamma^2 T\left((1-\rho)\mathbb{E}\left[\sigma_k^2\right]+C\mathbb{E}\left[\left\|z^{k+1/2}-w^k\right\|^2\right]+D_2\right)\right] - \sum_{k=0}^{K-1}\left[\gamma^2 T\mathbb{E}\left[\sigma_{k+1}^2\right]\right] +$$
$$+ \mathbb{E}\left[\max_{u\in\mathscr{C}}\sum_{k=0}^{K-1}\left[-\tau\left\|w^k-u\right\|^2-(1-\tau)\left\|z^{k+1}-u\right\|^2+\left\|w^{k+1}-u\right\|^2\right]\right] - \sum_{k=0}^{K-1}\left[\tau\mathbb{E}\left[\left\|z^{k+1/2}-z^k\right\|^2\right]+\right.$$

$$+ (1-\tau-2\gamma^2 A)\mathbb{E}\left[\left\|z^{k+1/2}-w^k\right\|^2\right] + 2\gamma^2 KD_1 + \sum_{k=0}^{K-1}\left[2\gamma^2 B\mathbb{E}\left[\sigma_k^2\right]\right] +$$
$$+ 2\gamma\mathbb{E}\left[\max_{u\in\mathscr{C}}\sum_{k=0}^{K-1}\left[\left\langle g^{k+1/2}-F(z^{k+1/2}),u-z^{k+1/2}\right\rangle\right]\right] = \max_{u\in\mathscr{C}}\left[\tau\left\|z^0-u\right\|^2+\left\|w^0-u\right\|^2\right] +$$
$$+ \sum_{k=0}^{K-1}\left[\gamma^2 T\left(1+\frac{2B}{T}-\rho\right)\mathbb{E}\left[\sigma_k^2\right]\right] - \sum_{k=0}^{K-1}\left[\gamma^2 T\mathbb{E}\left[\sigma_{k+1}^2\right]\right] +$$
$$+ \mathbb{E}\left[\max_{u\in\mathscr{C}}\sum_{k=0}^{K-1}\left[-\tau\left\|w^k-u\right\|^2-(1-\tau)\left\|z^{k+1}-u\right\|^2+\left\|w^{k+1}-u\right\|^2\right]\right] - \sum_{k=0}^{K-1}\left[\tau\mathbb{E}\left[\left\|z^{k+1/2}-z^k\right\|^2\right]+\right.$$

$$+ (1-\tau-\gamma^2(2A+TC))\mathbb{E}\left[\left\|z^{k+1/2}-w^k\right\|^2\right] + \gamma^2 K(2D_1+TD_2) +$$
$$+ 2\gamma\mathbb{E}\left[\max_{u\in\mathscr{C}}\sum_{k=0}^{K-1}\left[\left\langle g^{k+1/2}-F(z^{k+1/2}),u-z^{k+1/2}\right\rangle\right]\right].$$

С $\gamma \le \dfrac{\sqrt{1-\tau}}{\sqrt{2A+TC}}$ и $T \ge \dfrac{2B}{\rho}$ получим

$$2\gamma\mathbb{E}\left[\max_{u\in\mathscr{C}}\sum_{k=0}^{K-1}\operatorname{gap}(z^{k+1/2},u)\right] \le \max_{u\in\mathscr{C}}\left[\tau\left\|z^0-u\right\|^2+\left\|w^0-u\right\|^2\right] + \gamma^2 T\sigma_0^2 +$$
$$+ \mathbb{E}\left[\max_{u\in\mathscr{C}}\sum_{k=0}^{K-1}\left[-\tau\left\|w^k-u\right\|^2-(1-\tau)\left\|z^{k+1}-u\right\|^2+\left\|w^{k+1}-u\right\|^2\right]\right] + \tag{16}$$
$$+ 2\gamma\mathbb{E}\left[\max_{u\in\mathscr{C}}\sum_{k=0}^{K-1}\left[\left\langle g^{k+1/2}-F(z^{k+1/2}),u-z^{k+1/2}\right\rangle\right]\right] + \gamma^2 K(2D_1+TD_2).$$

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

$$\sum_{k=0}^{K-1}\left\langle\delta^k,z^{k+1/2}-u\right\rangle = \sum_{k=0}^{K-1}\left\langle\delta^k,z^{k+1/2}-v^k\right\rangle + \sum_{k=0}^{K-1}\left\langle\delta^k,v^k-u\right\rangle. \tag{17}$$

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

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

Переписав это неравенство, получим

$$\left\langle \gamma\delta^k, v^k - z\right\rangle \le \left\langle \gamma\delta^k, v^k - v^{k+1}\right\rangle + \left\langle v^{k+1} - v^k, z - v^{k+1}\right\rangle \le \left\langle \gamma\delta^k, v^k - v^{k+1}\right\rangle + \frac{1}{2}\left\|v^k - z\right\|^2 -$$
$$- \frac{1}{2}\left\|v^{k+1} - z\right\|^2 - \frac{1}{2}\left\|v^k - v^{k+1}\right\|^2 \le \frac{\gamma^2}{2}\left\|\delta^k\right\|^2 + \frac{1}{2}\left\|v^k - v^{k+1}\right\|^2 + \frac{1}{2}\left\|v^k - z\right\|^2 - \frac{1}{2}\left\|v^{k+1} - z\right\|^2 -$$
$$- \frac{1}{2}\left\|v^k - v^{k+1}\right\|^2 = \frac{\gamma^2}{2}\left\|\delta^k\right\|^2 + \frac{1}{2}\left\|v^k - z\right\|^2 - \frac{1}{2}\left\|v^{k+1} - z\right\|^2.$$

Вместе с (17) это дает

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

Берем максимум по u и получаем

$$\max_{u\in\mathscr{C}}\sum_{k=0}^{K-1}\left\langle \delta^k, z^{k+1/2} - u\right\rangle \le \sum_{k=0}^{K-1}\left\langle \delta^k, z^{k+1/2} - v^k\right\rangle + \frac{1}{2\gamma}\max_{u\in\mathscr{C}}\left\|v^0 - u\right\|^2 + \frac{\gamma}{2}\sum_{k=0}^{K-1}\left\|F(z^{k+1/2}) - g^{k+1/2}\right\|^2.$$

Берем полное математическое ожидание и получаем

$$\mathbb{E}\left[\max_{u\in\mathscr{C}}\sum_{k=0}^{K-1}\left\langle \delta^k, z^{k+1/2} - u\right\rangle\right] \le \mathbb{E}\left[\sum_{k=0}^{K-1}\left\langle \delta^k, z^{k+1/2} - v^k\right\rangle\right] + \frac{\gamma}{2}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|F(z^{k+1/2}) - g^{k+1/2}\right\|^2\right] +$$
$$+ \frac{1}{2\gamma}\max_{u\in\mathscr{C}}\left\|v^0 - u\right\|^2 = \mathbb{E}\left[\sum_{k=0}^{K-1}\left\langle \mathbb{E}\left[F(z^{k+1/2}) - g^{k+1/2}\,|\,z^{k+1/2} - v^k\right], z^{k+1/2} - v^k\right\rangle\right] + \tag{18}$$
$$+ \frac{\gamma}{2}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|F(z^{k+1/2}) - g^{k+1/2}\right\|^2\right] + \frac{1}{2\gamma}\max_{u\in\mathscr{C}}\left\|v^0 - u\right\|^2 = \frac{\gamma}{2}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|F(z^{k+1/2}) - g^{k+1/2}\right\|^2\right] + \frac{1}{2\gamma}\max_{u\in\mathscr{C}}\left\|v^0 - u\right\|^2.$$

Далее оценим

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

для этого заметим, что

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

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

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

Далее можно провести рассуждения аналогично цепочке рассуждений для (18):

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

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

$$+\max_{u\in\mathscr{C}}\left\|v^0-u\right\|^2=\sum_{k=0}^{K-1}\mathbb{E}\left[-\left\|\mathbb{E}_{w^{k+1}}[w^{k+1}]\right\|^2+\mathbb{E}_{w^{k+1}}\left\|w^{k+1}\right\|^2\right]+\max_{u\in\mathscr{C}}\left\|v^0-u\right\|^2= \tag{19}$$

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

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

Подставив (18) и (19) в (16), получим

$$2\gamma\mathbb{E}\left[\max_{u\in\mathscr{C}}\sum_{k=0}^{K-1}\mathrm{gap}(z^{k+1/2},u)\right]\le\max_{u\in\mathscr{C}}\left[(2+\tau)\left\|z^0-u\right\|^2+\left\|w^0-u\right\|^2\right]+\gamma^2T\sigma_0^2+$$

$$+\sum_{k=0}^{K-1}\left[\tau(1-\tau)\mathbb{E}\left[\left\|z^{k+1}-w^k\right\|^2\right]+\gamma^2\mathbb{E}\left[\left\|F(z^{k+1/2})-g^{k+1/2}\right\|^2\right]\right]+\gamma^2K(2D_1+TD_2).$$

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

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

$$+\sum_{k=0}^{K-1}\left[\tau(1-\tau)\mathbb{E}\left[\left\|z^{k+1}-w^k\right\|^2\right]+\gamma^2E\mathbb{E}\left[\left\|z^{k+1/2}-w^k\right\|^2\right]\right]+\gamma^2T\sigma_0^2+\gamma^2K(2D_1+TD_2+D_3).$$

С $\gamma\le\dfrac{\sqrt{1-\tau}}{\sqrt{E}}$ приходим к

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

$$\left.+\mathbb{E}\left[\left\|z^{k+1/2}-w^k\right\|^2\right]\right]+\gamma^2T\sigma_0^2+\gamma^2K(2D_1+TD_2+D_3)\le\max_{u\in\mathscr{C}}\left[(2+\tau)\left\|z^0-u\right\|^2+\left\|w^0-u\right\|^2\right]+ \tag{20}$$

$$+\gamma^2T\sigma_0^2+3(1-\tau)\sum_{k=0}^{K-1}\left[\mathbb{E}\left[\left\|z^{k+1}-z^{k+1/2}\right\|^2\right]+\mathbb{E}\left[\left\|z^{k+1/2}-w^k\right\|^2\right]\right]+\gamma^2K(2D_1+TD_2+D_3).$$

Вернемся к (15) с $\mu_h=0$, $\mu_F=0$, $T\ge\dfrac{2B}{\rho}$, $\gamma\le\dfrac{\sqrt{1-\tau}}{2\sqrt{2A+TC}}$ и получим

$$\mathbb{E}[V_{k+1}]\le\mathbb{E}[V_k]-\left((1-\tau)-2\gamma^2A-\gamma^2TC\right)\mathbb{E}\left[\left\|z^{k+1/2}-w^k\right\|^2\right]-\frac{1}{2}\mathbb{E}\left[\left\|z^{k+1}-z^{k+1/2}\right\|^2\right]+\gamma^2(2D_1+TD_2)\le$$

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

Откуда

$$3(1-\tau)\mathbb{E}\left[\left\|z^{k+1/2}-w^k\right\|^2+\left\|z^{k+1}-z^{k+1/2}\right\|^2\right]\le6\mathbb{E}[V_k-V_{k+1}]+6\gamma^2(2D_1+TD_2). \tag{21}$$

Подставляя (21) в (20), получаем

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

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

$$\mathbb{E}\left[\max_{u\in\mathscr{C}}\sum_{k=0}^{K-1}\mathrm{gap}(z^{k+1/2},u)\right]=\mathbb{E}\left[\max_{u\in\mathscr{C}}\sum_{k=0}^{K-1}\left[\left\langle F(z^{k+1/2}),z^{k+1/2}-u\right\rangle+h(z^{k+1/2})-h(u)\right]\right]\ge$$
$$\ge\mathbb{E}\left[\max_{u\in\mathscr{C}}\sum_{k=0}^{K-1}\left[\left\langle F(u),z^{k+1/2}-u\right\rangle+h(z^{k+1/2})-h(u)\right]\right]\ge$$
$$\ge\mathbb{E}\left[K\max_{u\in\mathscr{C}}\left[\left\langle F(u),\overline{z}^K-u\right\rangle+h(\overline{z}^K)-h(u)\right]\right]=K\mathbb{E}\left[\mathrm{Gap}(\overline{z}^K)\right],$$

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

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

2.3. Анализ для различных методов

Здесь мы 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-липшицев с константами L и D, т.е. для любых $z_1,z_2\in\mathscr{Z}$ верно*

$$\left\|F(z_1)-F(z_2)\right\|^2\le L^2\left\|z_1-z_2\right\|^2+D^2.$$

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 (см. [10]) для задачи (1)–(3). Далее отметим методы с однократным вызовом оракула (см. [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 (см. 11]), специализированный для задач решения ВН (1)–(4) также удовлетворяет условиям предлагаемого анализа.

Ко-коэрцитивность. Это предположение аналогично липшицевости оператора:

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

Видно, 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, см. [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 проделано в [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 (см. [29]). 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 (см. [30]). Также координатный метод очень близок к безградиентным методам (см. [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 использовании так называемого оператора квантизации:

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

Такие операторы могут быть рандомизированными или детерминированными с большим или малым параметром ω (см. [14], [32]), но все они имеют одну и ту же функцию – сжать вектор x. Методы с квантизацией популярны с точки зрения распределенной оптимизации, поскольку основной проблемой 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 вариантом, 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) (5.7). В этом случае мы рассматриваем задачу более общую, чем (1)–(4). Здесь 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) (5.8). Этот метод относится к так называемым локальным методам, которые делают ряд локальных обновлений между периодическими коммуникациями. Наш метод является рандомизированным (см. [7]) и основан на методе из предыдущего абзаца, а также использует технику сэмплирования по важности.

3. ЗАКЛЮЧЕНИЕ

В настоящей работе был рассмотрен экстраградиентный подход, базовый подход для решения вариационных неравенств, а также различные его модификации. Мы представили унифицированный анализ, с помощью которого можно единообразно анализировать семейство методов, основанных на экстраградиенте. Полученный унифицированный анализ позволил улучшить оценки сходимости уже существующих методов, а также представить новые методы и их гарантии сходимости. В будущем хотелось бы усовершенствовать анализ, чтобы в его предположения подпадали методы, которые не могут быть описаны текущей версией анализа.

ПРИЛОЖЕНИЕ

А. АНАЛИЗ ДЛЯ РАЗЛИЧНЫХ МЕТОДОВ

А.1. Экстраградиентный метод

Начнем с простейшего случая в рамках (1)–(3) – стохастического с равномерно ограниченным шумом [10]:

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

где z и ξ независимые. Для него может быть применен следующий метод (Алгоритм 1):

Algorithm 1. Экстраградиентный метод (Extra Step)

Параметры: Размер шага γ, K.

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

for $k = 0,1,\ldots,K-1$

Выбрать случайные ξ^k, $\xi^{k+1/2}$,

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

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

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 ограниченно-липшицев с константами L и D (предположение* 3*), тогда g^k и g^{k+1} из алгоритма* 5.1 *удовлетворяют предположению* 2 *с константами $A = 3L^2$, $D_1 = 3D^2 + 6\sigma^2$, $D_3 = \sigma^2$.*

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

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

и, наконец,

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

Следствие 1. Предположим, что F ограниченно-липшицев с константами L и D. Тогда экстраградиентный метод Extra Step

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

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

- в монотонном случае $\gamma \le 1/3L$ удовлетворяет

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

При правильном выборе γ (*см., например,* [28]), *можно прийти к следующим оценкам скорости сходимости*:

• *в сильно монотонном случае*

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

• *в монотонном случае*

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

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

При правильном выборе γ (см., например, [28]), можно прийти к следующим оценкам скорости сходимости:

• в сильно монотонном случае

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

• в монотонном случае

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

A.2. Экстраградиентный метод без дополнительного вызова оракула

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

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* $\rho = 1/3$, $B = 3$, $C = 2L^2$, $D_1 = 6\sigma^2$, $D_2 = 4D^2 + 12\sigma^2$, $D_3 = \sigma^2$.

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

Параметры: Размер шага γ, K.

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

for $k = 0,1,\ldots,K-1$

Выбрать случайно $\xi_{k+1/2}$,

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

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

end for

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

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

$$+ 2L^2\mathbb{E}\left[\left\|z^{k-1/2} - z^k\right\|^2\right] + 4D^2 = 2L^2\mathbb{E}\left[\left\|z^k - z^{k+1/2}\right\|^2\right] + 4D^2 +$$

$$+ 2L^2\mathbb{E}\left[\left\|z^{k-1} - \gamma F(z^{k-1/2}, \xi_k) - z^{k-1} + \gamma F(z^{k-3/2}, \xi_{k-1})\right\|^2\right] = 2L^2\mathbb{E}\left[\left\|z^k - z^{k+1/2}\right\|^2\right] +$$

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

если положить $\gamma \le 1/(3L)$. Следовательно,

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

и, наконец,

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

Следствие 2. Предположим, что F ограниченно-липшицев с константами L и D. Тогда Past-ES

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

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

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

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

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

A.3. Экстраградиентный метод с редукцией дисперсии

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

Параметры: Размер шага γ, K.

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

for $k = 0,1,\ldots,K-1$

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

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

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

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

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

end for

Мы полагаем $\sigma_k = 0$. Следующая лемма дает 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 $g^{k+1/2}$ несмещенно. Далее

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

и, наконец,

$$\mathbb{E}\left[\left\|g^{k+1/2} - F(z^{k+1/2})\right\|^2\right] = \mathbb{E}\left[\left\|F_{m_k}(z^{k+1/2}) - F_{m_k}(w^k) + F(w^k) - F(z^{k+1/2})\right\|^2\right] \le$$
$$\le 2\mathbb{E}\left[\left\|F_{m_k}(z^{k+1/2}) - F_{m_k}(w^k)\right\|^2\right] + 2\mathbb{E}\left[\left\|F(z^{k+1/2}) - F(w^k)\right\|^2\right] \le 4L^2\mathbb{E}\left[\left\|w^k - z^{k+1/2}\right\|^2\right] + 4D^2.$$

Следствие 3. Предположим, что каждый F_{m_k} и сам F ограниченно-липшицевы с константами L и D. Тогда VR-ES

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

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

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

$$\mathbb{E}\left[\mathrm{Gap}(\overline{z}^K)\right] \le \frac{8\max_{u\in\mathscr{C}}\left[\left\|z^0 - u\right\|^2\right]}{\gamma K} + 11\gamma D^2.$$

На каждой итерации мы вычисляем лишь один 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 τ следующим образом:

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

A.4. Покомпонентный экстраградиентный метод

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

Algorithm 4. Покомпонентный экстраградиентный метод (Coord-ES)

Параметры: Размер шага γ, K.

Инициализация: Выбрать $z^0 = w^0 \in \mathscr{L}$.

for $k = 0,1,\ldots,K-1$

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

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

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

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

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

end for

Положим $\sigma_k = 0$. Следующая лемма дает значения констант для предположения 2.

Лемма 5. *Предположим, что F ограниченно-липшицев 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 $g^{k+1/2}$ несмещенно. Далее

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

и, наконец,

$$\mathbb{E}\left[\left\|g^{k+1/2} - F(z^{k+1/2})\right\|^2\right] = \mathbb{E}\left[\left\|d[F(z^{k+1/2})]_{i_k} e_{i_k} - d[F(w^k)]_{i_k} e_{i_k} + F(w^k) - F(z^{k+1/2})\right\|^2\right] \le$$
$$\le 2\mathbb{E}\left[\left\|d[F(z^{k+1/2}) - F(w^k)]_{i_k} e_{i_k}\right\|^2\right] + 2\mathbb{E}\left[\left\|F(z^{k+1/2}) - F(w^k)\right\|^2\right] \le$$
$$\le 2(d+1)L^2\mathbb{E}\left[\left\|w^k - z^{k+1/2}\right\|^2\right] + 2(d+1)D^2.$$

Следствие 4. Предположим, что F ограниченно-липшицев с константами L и D. Тогда Coord-ES

- в сильно монотонном случае с $\gamma \le \min\left\{\dfrac{\sqrt{1-\tau}}{2\sqrt{2d}L}; \dfrac{1-\tau}{4(\mu_F + \mu_h)}\right\}$ удовлетворяет

$$\mathbb{E}\left[\tau\left\|z^{k+1} - z^*\right\|^2 + \left\|w^{k+1} - z^*\right\|^2\right] \le \left(1 - \gamma\frac{\mu_F + \mu_h}{16}\right)^{K-1}\left(\tau\left\|z^0 - z^*\right\|^2 + \left\|w^0 - z^*\right\|^2\right) + \frac{32\gamma dD^2}{\mu_F + \mu_h},$$

- в монотонном случае с $\gamma \le \dfrac{\sqrt{1-\tau}}{2L\sqrt{4d+2}}$ удовлетворяет

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

Оптимальное значение $\tau = d/(d+1)$.

A.5. Квантизованный экстраградиентный метод

В этом пункте рассматривается метод, использующий квантизованный оператор.

Algorithm 5. Квантизованный экстраградиентный метод (Quant-ES)

Параметры: Размер шага γ, K.

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

for $k = 0,1,\ldots,K-1$

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

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

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

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

end for

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

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

для некоторого ω.

Лемма 6. *Предположим, что F ограниченно-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 $g^{k+1/2}$ несмещенно. Далее

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

и, наконец,

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

Следствие 5. Предположим, что F ограниченно-липшицев с константами L и D. Тогда Quant-ES

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

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

- в монотонном случае с $\gamma \le \dfrac{\sqrt{1-\tau}}{2L\sqrt{4\omega+2}}$ удовлетворяет

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

Рассмотрим случай $D = 0$. Квантизация требуется, чтобы сжать информацию, при этом ω выступает здесь как коэффициент сжатия, т.е. мы передаем в ω 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 τ следующим образом:

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

A.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.

Algorithm 6. Квантизованный экстраградиентный метод с редукцией дисперсии

Параметры: Размер шага γ, K.

Инициализация: Выбрать $z^0 = w^0 \in \mathscr{L}$.

for $k = 0,1,\ldots,K-1$

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

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

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

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

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

end for

Положим $\sigma_k = 0$. Следующая лемма дает значения констант для предположения 2.

Лемма 7. *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 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 $g^{k+1/2}$ несмещенно. Далее

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

$$= \mathbb{E}\left[\left\|Q(F_{m_k}(z^{k+1/2}) - F_{m_k}(w^k))\right\|^2\right] \le \omega\mathbb{E}\left[\left\|F_{m_k}(z^{k+1/2}) - F_{m_k}(w^k)\right\|^2\right] \le \omega L^2\mathbb{E}\left[\left\|w^k - z^{k+1/2}\right\|^2\right] + \omega D^2$$

и, наконец,

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

$$\le 2\mathbb{E}\left[\left\|Q(F_{m_k}(z^{k+1/2}) - F_{m_k}(w^k))\right\|^2\right] + 2\mathbb{E}\left[\left\|F(w^k) - F(z^{k+1/2})\right\|^2\right] \le$$

$$\le 2(\omega+1)L^2\mathbb{E}\left[\left\|w^k - z^{k+1/2}\right\|^2\right] + 2(\omega+1)D^2.$$

Следствие 6. Предположим, что каждый F_{m_k} и сам F ограниченно-липшицевы с константами L и D. Тогда квантизованный экстраградиентный метод с редукцией дисперсии

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

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

- в монотонном случае с $\gamma \le \dfrac{\sqrt{1-\tau}}{2L\sqrt{4\omega+2}}$ удовлетворяет

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

Оптимальное значение τ здесь то же, что и в предыдущем пункте.

A.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 η вида

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

На каждом шаге мы обращаемся к F_η. Веса/вероятности p_m могут быть заданы априори, или же мы можем задать их самостоятельно: например, имеет смысл выбрать $p_m = L_m/\left(\sum_m L_m\right)$.

Algorithm 7. Экстраградиентный метод с сэмплированием по важности

Параметры: Размер шага γ, K.

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

for $k = 0,1,\ldots,K-1$

Выбрать η_k,

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

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

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

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

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 $g^{k+1/2}$ несмещенно. Далее

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

и, наконец,

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

Следствие 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 $\gamma \le \min\left\{\dfrac{\sqrt{1-\tau}}{2\sqrt{2}\sum_{m=1}^{M}\frac{L_m^2}{p_m}}; \dfrac{1-\tau}{4(\mu_F+\mu_h)}\right\}$ удовлетворяет

$$\mathbb{E}\left[\tau\left\|z^{k+1}-z^*\right\|^2+\left\|w^{k+1}-z^*\right\|^2\right]\le$$

$$\le\left(1-\gamma\frac{\mu_F+\mu_h}{16}\right)^{K-1}\left(\tau\left\|z^0-z^*\right\|^2+\left\|w^0-z^*\right\|^2\right)+\frac{32\gamma}{\mu_F+\mu_h}\sum_{m=1}^{M}\frac{D_m^2}{p_m},$$

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

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

A.8. Локальный экстраградиентный метод

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

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

Algorithm 8. Рандомизированный локальный экстраградиентный метод

Параметры: Размер шага γ, K, вероятность p.

Инициализация: Выбрать $z^0=w^0\in\mathscr{Q}$, $z_m^0=z^0$ для всех m.

for $k=0,1,\ldots$

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

$$Z^{k+1/2}=\mathrm{prox}_{\gamma h}\left(\overline{Z}^k-\gamma(\Phi(W^k)+\lambda(W^k-W_{\mathrm{avg}}^k)\right),$$

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

$$Z^{k+1}=\mathrm{prox}_{\gamma h}\left(\overline{Z}^k-\gamma(G(Z^{k+1/2})-G(W^k)+\Phi(W^k)+\lambda(W^k-W_{\mathrm{avg}}^k))\right),$$

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

end for

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

Если выбрать $\tau = L/(L+\lambda)$, то верно

Следствие 8. Рандомизированный локальный экстраградиентный метод

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

$$\mathbb{E}\left[\tau\left\|z^{k+1}-z^*\right\|^2 + \left\|w^{k+1}-z^*\right\|^2\right] \le$$

$$\le \left(1-\gamma\frac{\mu_F+\mu_h}{16}\right)^{K-1}\left(\tau\left\|z^0-z^*\right\|^2 + \left\|w^0-z^*\right\|^2\right) + \frac{32\gamma}{\mu_F+\mu_h}\sum_{m=1}^{M}\frac{D_m^2}{p_m},$$

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

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

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

• в сильно монотонном случае

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

• в монотонном случае

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

Отметим, что эти оценки являются достаточно хорошими при малом значении λ.

В. ЭКСПЕРИМЕНТЫ

В.1. Генеративно-состязательные сети

В качестве примера оптимизации минимаксного 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 есть реальное изображение. Согласно статье Гудфеллоу [6], в этой задаче D пытается максимизировать вероятность того, что он правильно классифицирует реальные и фейковые изображения ($\lg(D(x))$, а G пытается минимизировать вероятность того, что D будет предсказывать его выходные данные как фейковые ($\lg(1-D(G(z)))$. Из статьи функция потерь GAN есть

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

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 50 000 и 10 000 изображений в обучающей и валидационной выборках соответственно, равномерно распределенных по десяти классам. Для каждого оптимизационного подхода в качестве оптимизатора использовался Adam. В ходе эксперимента был проведен анализ гиперпараметроов каждого из оптимизаторов (генератора и дискриминатора) (табл. 1).

По результатам проведенного анализа сходимости (проверялись результаты работы после 100 эпох обучения) были выбраны следующие параметры: $\beta_1 = 0.9$, $\beta_2 = 0.999$ и коэффициент

Таблица 1. Набор анализируемых гиперпараметров

β_1	0.5	0.9	0.99
β_2	00.99	0.999	–
Скорость обучения	0.9×10^{-4}	1×10^{-4}	2×10^{-4}

скорости обучения $= 2\times10^{-4}$ для обоих методов оптимизации. Размер набора элементов, на котором производилась одна итерация вычисления градиента, составлял 64.

Для проверки качества выдаваемых изображений была использована метрика Inception score (см. [32]):

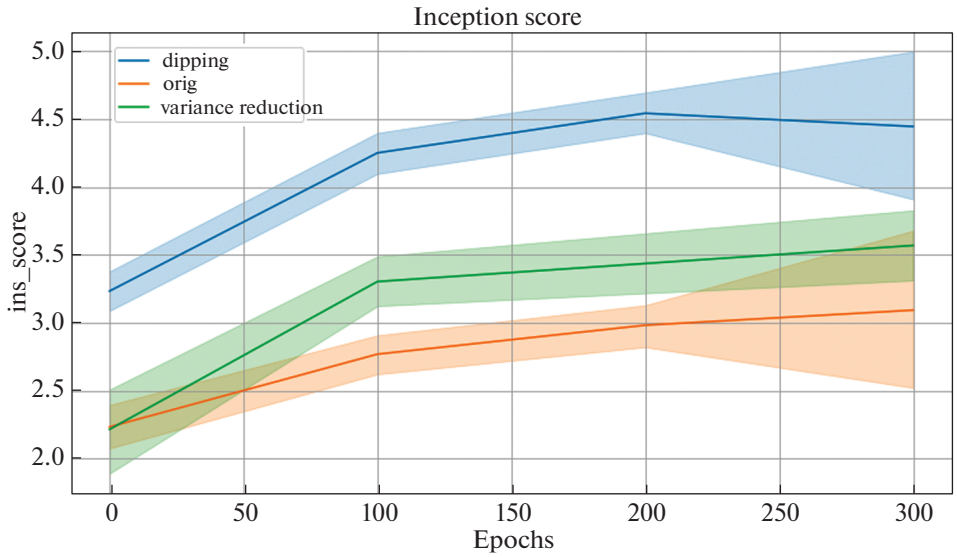
$$\mathbf{IS}(G) = \exp(\mathbb{E}_{x\sim p_g} D_{KL}(p(y|\mathbf{x})||\, p(y))),$$

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 (см. [33]) с подходом Conditional (см. [34]). 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 фиг. 1 и 2. Согласно результатам, полученным в рамках экспериментов, можно заметить, что подходы, предложенные в данной статье, позволяют достигнуть лучших результатов, по сравнению с оригинальными.



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

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, (в) – оптимизация со стохастической редукцией дисперсии градиентного спуска.

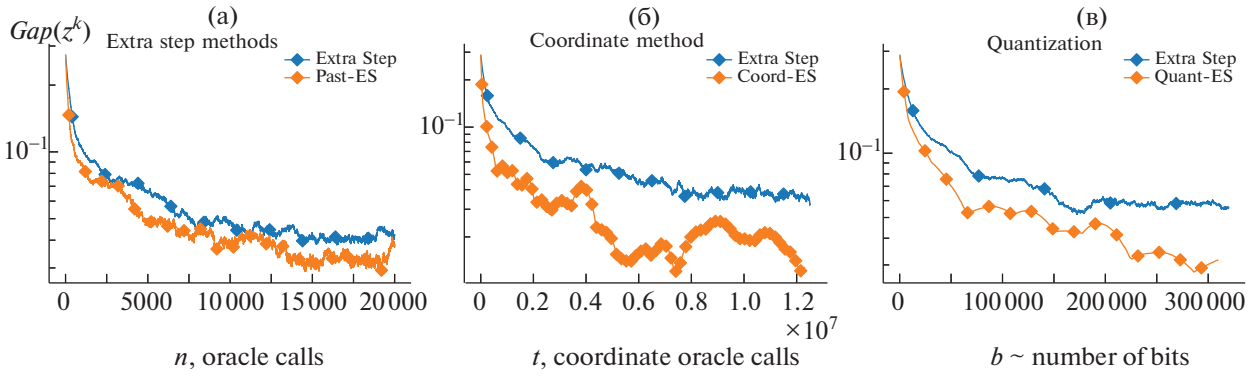
• Метод редукции дисперсии позволял оптимизировать функционал точнее, чем оригинальный способ, на протяжении всей эпохи оптимизации, однако, приближаясь к конечным итерациям каждой эпохи обучения, полный градиент достаточно сильно отличался от направления градиента, получаемого на последних итерациях, 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 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 генеративных моделей и могут быть применены для оптимизации GAN, но квантизация позволяет быстро получить наилучшие результаты при сохранении вычислительных ресурсов.

B.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



Фиг. 3. Сходимость методов с дополнительным шагом для задачи Полицейский vs. Грабитель (20): (а) – методы Extra Step и Past-ES, (б) – Extra Step и Coord-ES, (в) – Extra Step и Quant-ES.

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 $\exp(-\theta d(i,j))$ для функции расстояния d, которая вводится выше. Мы можем сформулировать описанную постановку как задачу о поиске билинейной седловой точки:

$$\min_{x\in\Delta(n^2)}\max_{y\in\Delta(n^2)} f(x,y) := \frac{1}{n}\sum_{k=1}^{n} y^\top A^{(k)} x, \qquad (\text{П}.22)$$

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

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

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

$$w_i = 1 - \frac{2}{n}\min\{\|\lfloor i/n\rfloor - n/2|, |i \bmod n - n/2|\},$$
$$w_i^{(k)} = w_i(1+\xi^{(k)}) \quad \text{для} \quad \xi^{(k)} \sim \mathcal{U}(0,\sigma),$$
$$d(i,j) = \sqrt{(\lfloor i/n\rfloor - \lfloor j/n\rfloor)^2 + (i \bmod n - j \bmod n)^2}.$$

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

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=lX%5C_7Rce3%5C_Q0C.
2. *Nesterov Y.* Smooth minimization of non-smooth functions // Math. Program. 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 J. Optimizat. 2004. V. 15. № 1. P. 229–251.

4. *Chambolle A., Pock T.* A first-order primal-dual algorithm for convex problems with applications to imaging // J. Math. Imag. and Vis. 2011. V. 40. № 1. P. 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 J. Imag. Sci. 2010. V. 3. № 4. P. 1015–1046.
6. Generative Adversarial Networks / I.J. Goodfellow [и др.]. 2014. arXiv: 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 // Ekonomika i Matematicheskie Metody. 1976. V. 12. № 4. P. 747–756.
10. *Juditsky A., Nemirovski A., Tauvel C.* Solving variational inequalities with stochastic mirror-prox algorithm // Stochastic System. 2011. V. 1. № 1. P. 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 // Ann. Math. Statistic. 1951. V. 22. № 3. P. 400–407. URL: https://doi.org/10.1214/aoms/1177729586
13. *Johnson R., Zhang T.* Accelerating stochastic gradient descent using predictive variance reduction // Adv. Neural Informat. Processing System. 2013. V. 26. P. 315–323.
14. QSGD: Communication-efficient SGD via gradient quantization and encoding / D. Alistarh [и др.] // Adv. Neural Informat. Processing System. 2017. P. 1709–1720.
15. *Hanzely F., Mishchenko K., Richtarik P.* SEGA: Variance reduction via gradient sketching // arXiv preprint arXiv:1809.03054. 2018.
16. *Gorbunov E., Hanzely F., Richtarik P.* A unified theory of SGD: Variance reduction, sampling, quantization and coordinate descent // Internat. Conf. Artific. Intelligence and Statistic. PMLR. 2020. P. 680–690.
17. On the convergence of single-call stochastic extra-gradient methods / Y.-G. Hsieh [и др.]. 2019. arXiv: 1908.08465 [math.OC].
18. Revisiting stochastic extragradient / K. Mishchenko [и др.] // Internat. Conf. Artific. Intelligence and Statistic. PMLR. 2020. P. 4573–4582.
19. *Tseng P.* A Modified Forward-Backward Splitting Method for Maximal Monotone Mappings // SIAM J. Control and Optimizat. 2000. V. 38. № 2. P. 431–446. https://doi.org/10.1137/S0363012998338806
20. *Nesterov Y.* Dual extrapolation and its applications to solving variational inequalities and related problems // Math. Program. 2007. V. 109. № 2. P. 319–344.
21. *Palaniappan B., Bach F.* Stochastic Variance Reduction Methods for Saddle- Point Problems // Adv. Neural Informat. Processing System. T. 29 / Под ред. D. Lee [и др.]. Curran Associates, Inc., 2016. URL: https://proceedings.neurips.cc/paper/2016/file/1aa48fc4880bb0c9b8aPaper.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 Ann. Symp. Foundat. of Comput. Sci. (FOCS). IEEE. 2018. P. 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 // Internat. Conf. Artific. Intelligence and Statistic. PMLR. 2021. P. 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. *Stich S.U.* Unified optimal analysis of the (stochastic) gradient method // arXiv preprint arXiv:1907.04232. 2019.
29. *Wright S.J.* Coordinate descent algorithms // Math. Program. 2015. V. 151. № 1. C. 3–34.
30. *Nesterov Y.* Efficiency of coordinate descent methods on huge-scale optimization problems // SIAM J. Optimizat. 2012. V. 22. № 2. P. 341–362.
31. On Biased Compression for Distributed Learning / A. Beznosikov [и др.] arXiv preprint arXiv:2002.12410. 2020.
32. *Barratt S., Sharma R.* A note on the inception score // arXiv preprint arXiv:1801.01973. 2018.
33. *Radford A., Metz L., Chintala S.* Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. 2016. arXiv: 1511.06434 [cs.LG].
34. *Mirza M., Osindero S.* Conditional generative adversarial nets // arXiv preprint arXiv:1411.1784. 2014.