

ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ФГАОУ ВО НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

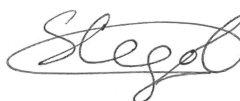
Факультет компьютерных наук
Образовательная программа «Прикладная математика и информатика»

УДК XXXXX

Отчет об исследовательском проекте на тему:
Устойчивость и обобщающая способность состязательных алгоритмов,
базирующихся на анализе эмпирической функции риска

Выполнили:

студент группы БПМИ202
Седов Сергей Алексеевич



(подпись)

18.05.2023

(дата)

Принял руководитель проекта:

Двинских Дарина Михайловна
Доцент
Факультета компьютерных наук НИУ ВШЭ



(подпись)

18.05.2023

(дата)

Содержание

Аннотация	3
1 Введение	4
1.1 Описание предметной области	4
1.1.1 Стохастическая оптимизация выпуклой функции	4
1.1.2 Формальное описание стохастического зеркального спуска	5
1.1.3 Формальное описание метода Монте Карло	6
1.2 Постановка задачи	7
1.3 Обзор литературы	8
1.3.1 Выпуклая оптимизация	8
1.3.2 Выпукло-вогнутая оптимизация	8
2 Основные результаты	9
2.1 Оценка точности оптимизации седловой функции	9
2.1.1 Оптимальность полученной оценки	13
2.2 Зеркальный спуск на симплексах	14
2.2.1 Вывод явной формулы шага	15
2.3 Стохастический зеркальный спуск	16
2.3.1 Рандомизация при проектировании в зеркальном спуске на симплексах	16
2.3.2 Зеркальный спуск в задаче билинейной формы	17
2.4 Оптимизация алгоритма в офлайн-постановке	17
2.4.1 Анализ собственных значений	17
2.4.2 Билинейные матричные игры	18
2.5 Сравнительный анализ описанных методов	19
2.5.1 Сравнение на сэмплированных матрицах	20
2.5.2 Описание способа нахождения векторов PageRank	21
2.5.3 Сравнение на циклических веб-графах	22
3 Выводы	23
Список литературы	24

Аннотация

Задача седловой оптимизации тщательно изучалась последние десятилетия, особенно в случае евклидовой постановки. В то же время для некоторых задач энтропийная постановка подходит лучше - в этой работе проводится сравнительный анализ методов стохастической оптимизации билинейной формы на симплексах в ней. Помимо этого, в работе выводятся теоретические оценки точности решения эмпирической седловой задачи.

Ключевые слова

Выпукло-вогнутая стохастическая оптимизация, зеркальный спуск, метод Монте-Карло.

1 Введение

1.1 Описание предметной области

1.1.1 Стохастическая оптимизация выпуклой функции

Рассмотрим задачу стохастической оптимизации

$$x^* = \arg \min_{x \in \mathcal{X}} F(x) := \mathbb{E}_{\xi} f(x, \xi), \quad (1)$$

где ξ - случайная величина, имеющая неизвестное распределение, а f выпукла по x . Оптимизация будет основываться на наборе независимых ξ_i из этого распределения.

Такая модель может описывать задачу регрессии, только в ней принято обозначать оптимизируемые параметры как w , а в качестве ξ берутся объекты в выборке из неизвестного нам распределения: $\xi = (x, y)$. Пусть $\phi(x)$ - вектор признаков объекта из выборки, тогда $f(w, \xi) = \text{Loss}(\langle w, \phi(x) \rangle, y)$ - некоторая выпуклая функция потерь. В таком случае, наша задача - подобрать наилучшие веса w , минимизируя её.

Существуют два классических подхода к решению этой задачи: онлайн подход, или стохастическая аппроксимация (Stochastic Approximation) и оффлайн подход, или аппроксимация выборочного среднего (Sample Average Approximation). Последний также известен как метод Монте-Карло:

- В стохастической оптимизации x_i оптимизируется итеративно, простейшим примером является градиентный спуск. Будем рассматривать его обобщение - зеркальный спуск. Зеркальный спуск основывается на prox-setup'ах - наборе предположений о постановке задачи. Как правило, имеются в виду выбор нормы и 1-сильно выпуклой функции, задающей расстояние на нашем множестве. Помимо этого, на каждом шаге вызывается оракул первого рода - градиент функции в последней точке. Засчёт выбора prox-setup'а мы можем обобщить решение задач в привычных евклидовых пространствах или же эффективнее находить решение, например, рассматривая оптимизации на симплексах или даже на пространствах матриц. Основным отличием от второго подхода является отсутствие ограничений на изменение оракула - может изменяться сама оптимизируемая функция, например, в выборку могут добавляться новые наблюдения.
- Метод Монте-Карло прост в постановке - вместо оптимизации самой задачи будем ре-

шать эмпирическую задачу оптимизации по выборке ξ_1, \dots, ξ_n :

$$\hat{x} = \arg \min_{x \in \mathcal{X}} \hat{F}(x) = \arg \min_{x \in \mathcal{X}} \frac{1}{n} \sum_{i=1}^n f(x, \xi_i) \quad (2)$$

В отличие от предыдущего случая, эмпирическая функция фиксированна. В данной задаче оценивание точности оптимизации складывается из удалённости эмпирического решения от истинного и из точности решения эмпирической задачи. В зависимости от ограничений, накладываемых на оптимизируемую функцию, и размера выборки доказываются оценки на обе части. Как правило, речь идёт о (строгой) выпуклости, Липшицевости и Липшицевости градиента. Однако в случае оптимизации седловых задач играет роль мера качества (точности) решения, используют не только лишь евклидово расстояние. Тем не менее, основные результаты по оценке точности решения удаётся перенести на седловой случай, служащий обобщением выпуклого.

1.1.2 Формальное описание стохастического зеркального спуска

Будем считать, что $f(x)$ выпукла и L-Липшицева на множестве \mathcal{X} . Определим Prox-setup:

- 1 Множество \mathcal{X} - замкнутое выпуклое подмножество Евклидова пространства E .
- 2 Норма $\|\cdot\|$ на $\mathbb{R}^n \ni \mathcal{X}$ и сопряжённая ей $\|\cdot\|_*$.
- 3 Функция, задающая расстояние $\omega(x) : \mathcal{X} \rightarrow \mathbb{R}$, непрерывна дифференцируемая и сильно выпуклая функция с константой 1 относительно выбранной нормы.
- 4 Дивергенция Брегмана $V_{x_0}(x) = \omega(x) - \langle \omega'(x_0), x - x_0 \rangle - \omega(x_0)$. $V_{x_0}(x) \geq \frac{1}{2} \|x - x_0\|^2$, то есть интуитивно эта величина оценивает, насколько близко значение функции в точке от приближенного первой степенью разложения в ряд Тейлора.
- 5 Диаметр множества: $\Omega = \sqrt{2\Theta}$, $\Theta = \sup_{x, x_0 \in \mathcal{X}} V_{x_0}(x)$. Засчёт ограничений на $V_{x_0}(x)$, ω -диаметр оценивает сверху $\max_{x, x_0 \in \mathcal{X}} \|x - x_0\|$.
- 6 Прокс-отображение $Prox_{x_0}(g) : \mathbb{R}^n \rightarrow \mathcal{X}$,
 $Prox_{x_0}(g) = \arg \min_{x \in \mathcal{X}} \{V_{x_0}(x) + \langle g, x \rangle\} = \arg \min_{x \in \mathcal{X}} \{\omega(x) + \langle g - \omega'(x_0), x \rangle\}$.

Приведём в пример две самых популярных постановки: евклидову и энтропийную.

1. Евклидовая постановка: на подмножестве \mathcal{X} Евклидова пространства $E = \mathbb{R}^n$ возьмём

вторую норму, $\omega(x) = \frac{1}{2} \langle x, x \rangle$, тогда:

$$Prox_{x_0}(g) = \Pi_{\mathcal{X}}(x_0 - g) = \arg \min_{x \in \mathcal{X}} \|(x_0 - g) - x\|_2$$

2. Энтропийная постановка: $E = \mathbb{R}^n$, \mathcal{X} - замкнутое выпуклое подмножество симплекса Δ_n^+ .

Берём регуляризованную энтропию в качестве $\omega(x)$:

$$\omega(x) = (1 + \delta) \sum_{i=1}^n (x_i + \delta/n) \ln(x_i + \delta/n) : \Delta_n^+ \rightarrow \mathbb{R}$$

Теперь можно сформулировать сам алгоритм зеркального спуска. Выбираем prox-setup, learning-rates (γ) и x_0 , после чего итеративно обновляем x_i :

$$x_{t+1} = Prox_{x_t}(\gamma_t f'(x_t))$$

Нетрудно заметить, что в случае евклидовой постановки зеркальный спуск - обычный градиентный. Однако введённая теоретическая база позволяет оптимизировать и другие задачи, где обычный евклидовый подход справляется плохо. В частности, нас будет интересовать поведение рандомизированного зеркального спуска при оптимизации билинейных седел на энтропийной постановке.

1.1.3 Формальное описание метода Монте Карло

В данном случае мы будем оптимизировать эмпирическое среднее для выборки ξ_1, \dots, ξ_n .

$$\hat{x} = \arg \min_{x \in \mathcal{X}} \hat{F}(x) = \arg \min_{x \in \mathcal{X}} \frac{1}{n} \sum_{i=1}^n f(x, \xi_i) \quad (3)$$

Как правило сходимость рассматривают в следующих предположениях:

- $f(x, \xi)$ выпукла / сильно выпукла с константой λ
- $f(x, \xi)$ обладает Липшицевостью с константой M .
- Градиент $f(x, \xi)$ обладает Липшицевостью с константой L .

В качестве обобщения задачи выпуклой оптимизации рассмотрим выпукло-вогнутую, также называемую седловой:

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \Phi(x, y) := \mathbb{E}_{\xi} [\Phi(x, y, \xi)] \quad (4)$$

Её оптимальным решением будет $z^* = (x^*, y^*)$. Эмпирической задачей назовём:

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \hat{\Phi}_n(x, y) := \frac{1}{n} \sum_{i=1}^n \Phi(x, y, \xi_i) \quad (5)$$

Аналогично, оптимальным решением эмпирической задачи является $\hat{z} = (\hat{x}, \hat{y})$. Для оценки точности сходимости эмпирического оптимума к истинному как правило используют часть из предположений:

- [SC-SC] Сильная выпуклость - сильная вогнутость Φ_ξ , то есть $\Phi_\xi(x, y)$ μ_x сильно выпукла относительно $\|\cdot\|_x$ и μ_y сильно вогнута относительно $\|\cdot\|_y$ (п.н. по ξ).
- [Lipschitz Continuity] Липшицевость Φ по обоим аргументам с константами M_x и M_y .
- [Gradient Lipschitz Continuity] Липшицевость градиентов $\nabla_x \Phi$ и $\nabla_y \Phi$ по каждой из переменных с соответствующими константами L_x , L_y и L_{xy} .

В отличие от выпуклой задачи, в выпукло-вогнутой оптимизации мы можем по-разному вводить метрики оценки качества. Первая из них - квадрат евклидового расстояния до точки оптимума.

$$d^2(\hat{x}, \hat{y}) = \|\hat{x} - x^*\|_x^2 + \|\hat{y} - y^*\|_y^2 \quad (6)$$

Помимо этого вводят так называемые WGM - weak generality measure и SGM - strong generality measure (aka duality gap - зазор двойственности):

$$\Delta^w(\hat{x}, \hat{y}) = \max_{y \in \mathcal{Y}} \mathbb{E}_\xi [\Phi(\hat{x}, y)] - \min_{x \in \mathcal{X}} \mathbb{E}_\xi [\Phi(x, \hat{y})], \quad \Delta^s(\hat{x}, \hat{y}) = \mathbb{E}_\xi \left[\max_{y \in \mathcal{Y}} \Phi(\hat{x}, y) - \min_{x \in \mathcal{X}} \Phi(x, \hat{y}) \right] \quad (7)$$

Легко показать, что для сильно выпуклой - сильно вогнутой Φ_ξ :

$$\mathbb{E}_\xi d^2(\hat{x}, \hat{y}) \leq \Delta^w(\hat{x}, \hat{y}) \leq \Delta^s(\hat{x}, \hat{y}) \quad (8)$$

То есть ограничения сходимости по SGM - самые строгие. В самой работе мы будем оценивать точность решения эмпирической задачи именно по этой мере.

1.2 Постановка задачи

Первая часть курсовой работы заключается в изучении предметной области и получении теоретического результата о необходимой точности решения эмпирической оффлайн-задачи

по заданной точности решения истинной. Фактически это является переносом результатов выпуклой оптимизации на седловой случай, показывая схожесть между ними.

Во второй части курсовой работы были реализованы несколько подходов оптимизации методом зеркального спуска в энтропийной постановке для билинейных седел на произведении симплексов. Было рассмотрено 2 подхода онлайн-оптимизации и 1 офлайн-подход, не допускающий изменения матрицы в итеративном процессе оптимизации. также эксперименты проводились на комбинациях двух онлайн-подходов.

- 1 Стохастический зеркальный спуск с рандомизацией Немировского-Юдицкого. Здесь предполагается эффективная реализация рандомизированного алгоритма на симплексах, с последующим анализом поведения метода.
- 2 Зеркальный спуск с рандомизацией на этапе проектирования на единичный симплекс. В этом алгоритме на каждом шаге оптимизации выбирается конкретная вершина симплекса, наподобие хода в билинейной матричной игре. То есть, по своей природе алгоритм может быть применим к немного другим задачам, но основная задача - сравнение скорости сходимости для общих билинейных матричных игр на симплексах.
- 3 Метод из статьи "Accelerating Smooth Games by Manipulating Spectral Shapes" предложенный в качестве оптимального для решения билинейных задач. Этот метод был представлен авторами статьи после изучения собственных значений Гессианов оптимизируемых функций и сравнении с их помощью известных методов оптимизации. В частности, для билинейных матричных игр авторы предложили улучшение метода градиентного спуска с моментом Бориса Поляка. Предложенный метод использует сингулярные значения матрицы билинейной игры, то есть является офлайн-методом, и не предполагает изменения матрицы в процессе игры.

1.3 Обзор литературы

1.3.1 Выпуклая оптимизация

Основными используемыми материалами по выпуклой оптимизации являются книга Аркадия Немировского [4], статья Shalev-Shartz'a [1] и диссертация Д.М.Двинских [7].

1.3.2 Выпукло-вогнутая оптимизация

Отправной точкой послужит статья Zhang'a [3], где исследуется сходимость решения эмпирической седловой задачи. Её результаты в каком-то смысле обобщают до выпукло-вогнутого

случая результаты Shalev-Shwartz'a. На ограничения Zhang'a мы опираемся в выводе теоретического результата для выпукло-вогнутого случая, делая это по аналогии с выпуклым случаем в Теореме 2.1.3 ранее упомянутой диссертации Д.М.Двинских [7].

В планах реализация алгоритма из статьи Accelerating Smooth Games by Manipulating Spectral Shapes [2], упомянутой в разделе постановки задачи.

При погружении в онлайн-методы оптимизации, в основном использовались книги А.С.Немировского [4], J.C.Duchi [6] и S.Bubeck [5]. Для дальнейшего изучения / реализации подходов используется статья А.В.Гасникова [8] и отрывок из книги Л.Г.Хачияна [10], описывающий подход с рандомизацией при проектировании на симплекс. Применение этого подхода при решении задачи PageRank так же описано в статье А.В.Гасникова [9].

2 Основные результаты

2.1 Оценка точности оптимизации седловой функции

Первой задачей был вывод в выпукло-вогнутом случае точности ε' решения эмпирической задачи, необходимой для ε точности истинной.

В статье Shalev-Shwartz'a [1] показана важность сильной выпуклости оптимизируемой функции - без неё равномерной сходимости эмпирического минимума к истинному нет. Контр-примеры приведены в этой же статье, они строятся на независимости размерности $x \in \mathcal{X}$ от размера выборки n . Это позволяет интерпретировать регуляризацию как необходимое для равномерной сходимости сильно-выпуклого слагаемое. Основным результатом статьи даёт гарантии на точность решения эмпирической задачи при λ -сильно выпуклой и M -Липшицевой функции f . Напомню, $F(x) = \mathbb{E}_\xi f(x, \xi)$. С вероятностью не меньше $1 - \delta$ выполнено:

$$F(\hat{x}) - F(x^*) \leq \frac{4M^2}{\delta\lambda n} \quad (9)$$

В Теореме 2.1.3 диссертации Д.М.Двинских [7] показана оценка на точность ε' решения эмпирической задачи, необходимой для гарантирования точности ε решения истинной, для этого используются результаты ранее упомянутой статьи. Пусть f λ -сильно выпуклая M -Липшицевая и эмпирическая задача решена с точностью ε' :

$$\hat{F}(\hat{x}_{\varepsilon'}) - \hat{F}(\hat{x}^*) \leq \varepsilon'$$

Тогда с вероятностью не меньше $1 - \delta$ выполняется:

$$F(\hat{x}_{\varepsilon'}) - F(x^*) \leq \sqrt{\frac{2M^2}{\lambda} \varepsilon'} + \frac{4M^2}{\delta \lambda n}$$

Рассмотрим офлайн-решение стохастической седловой задачи:

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \Phi(x, y) := \mathbb{E}_{\xi}[\Phi(x, y, \xi)], \quad \min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \hat{\Phi}_n(x, y) := \frac{1}{n} \sum_{i=1}^n \Phi(x, y, \xi_i) \quad (10)$$

Решением эмпирической задачи будем считать точку (\hat{x}, \hat{y}) . Оценивать точность решения стохастической задачи мы будем по зазору двойственности:

$$\Delta^s(\hat{x}, \hat{y}) = \mathbb{E}_{\xi} \left[\max_{y \in \mathcal{Y}} \Phi(\hat{x}, y, \xi) - \min_{x \in \mathcal{X}} \Phi(x, \hat{y}, \xi) \right] \quad (11)$$

$$\Delta_n^s(\hat{x}, \hat{y}) = \mathbb{E}_{\xi} \left[\max_{y \in \mathcal{Y}} \hat{\Phi}_n(\hat{x}, y, \xi) - \min_{x \in \mathcal{X}} \hat{\Phi}_n(x, \hat{y}, \xi) \right] \quad (12)$$

Допустим решение эмпирической задачи сошлось в точку $(\hat{x}_{\varepsilon'}, \hat{y}_{\varepsilon'})$. В последующей теореме мы оценим $\Delta^s(\hat{x}_{\varepsilon'}, \hat{y}_{\varepsilon'})$ сверху, показав, с какой точностью нужно решать эмпирическую задачу, чтобы гарантировать ε -точность решения истинной седловой задачи.

Теорема

Пусть функция $\Phi(x, y, \xi)$ сильно выпукла - сильно вогнута (1.1.3), и она сама, и её градиент Липшицевы (1.1.3). Тогда если решение $(\hat{x}_{\varepsilon'}, \hat{y}_{\varepsilon'})$ обеспечивает точность $\varepsilon' = \Delta_n^s(\hat{x}_{\varepsilon'}, \hat{y}_{\varepsilon'})$, то:

$$\varepsilon = \Delta^s(\hat{x}_{\varepsilon'}, \hat{y}_{\varepsilon'}) \leq C \sqrt{\varepsilon'} + B, \quad (13)$$

$$C = \max \left\{ \frac{2M_x^s}{\sqrt{\mu_x}}, \frac{2M_y^s}{\sqrt{\mu_y}} \right\}, \quad \Delta^s(\hat{x}, \hat{y}) \leq \frac{2\sqrt{2}}{n} \cdot \sqrt{\frac{L_{xy}^2}{\mu_x \mu_y} + 1} \cdot \left(\frac{(M_x^s)^2}{\mu_x} + \frac{(M_y^s)^2}{\mu_y} \right) = B \quad (14)$$

Таким образом, мы должны решать эмпирическую задачу с точностью $\varepsilon' = \mathcal{O}(\varepsilon^2)$, чтобы решить исходную задачу с точностью ε . Теорема является обобщением для седлового случая аналогичных результатов Теоремы 2.1.3 Д.Двинских [7], оценивающих точность приближенного решения эмпирической функции в выпуклой оптимизационной задаче. Ограничение на $\Delta^s(\hat{x}, \hat{y})$ - результат Теоремы 3 из Zhang21 [3].

Доказательство

Заметим, что по неравенству Йенсена для вогнутой функции:

$$\begin{cases} M_x^s = \sqrt{\sup_{y \in \mathcal{Y}} \mathbb{E}_\xi [M_x^2(\xi, y)]} = \sup_{y \in \mathcal{Y}} \sqrt{\mathbb{E}_\xi [M_x^2(\xi, y)]} \geq \sup_{y \in \mathcal{Y}} \mathbb{E}_\xi [M_x(\xi, y)] \\ M_y^s = \sqrt{\sup_{x \in \mathcal{X}} \mathbb{E}_\xi [M_y^2(\xi, x)]} = \sup_{x \in \mathcal{X}} \sqrt{\mathbb{E}_\xi [M_y^2(\xi, x)]} \geq \sup_{x \in \mathcal{X}} \mathbb{E}_\xi [M_y(\xi, x)] \end{cases} \quad (15)$$

Таким образом,

$$\begin{cases} M_x^s \geq |\Phi(x', y, \xi) - \Phi(x, y, \xi)| / \|x' - x\|_{p_x} \quad \forall x', x \in \mathcal{X} \\ M_y^s \geq |\Phi(x, y', \xi) - \Phi(x, y, \xi)| / \|y' - y\|_{p_y} \quad \forall y', y \in \mathcal{Y} \end{cases} \quad (16)$$

Распишем зазор двойственности $(\hat{x}_{\varepsilon'}, \hat{y}_{\varepsilon'})$:

$$\Delta^s(\hat{x}_{\varepsilon'}, \hat{y}_{\varepsilon'}) = \mathbb{E}_\xi \left[\max_{y \in \mathcal{Y}} \Phi(\hat{x}_{\varepsilon'}, y) - \min_{x \in \mathcal{X}} \Phi(x, \hat{y}_{\varepsilon'}) \right] - \Delta^s(\hat{x}, \hat{y}) + \Delta^s(\hat{x}, \hat{y}) = \quad (17)$$

$$= \mathbb{E}_\xi \left[\max_{y \in \mathcal{Y}} [\Phi(\hat{x}_{\varepsilon'}, y) - \Phi(\hat{x}, y)] + \min_{x \in \mathcal{X}} [\Phi(x, \hat{y}) - \Phi(x, \hat{y}_{\varepsilon'})] \right] + \Delta^s(\hat{x}, \hat{y}) \leq \quad (18)$$

$$\leq \mathbb{E}_\xi [M_x^s \cdot \|\hat{x}_{\varepsilon'} - \hat{x}\|_{p_x} + M_y^s \cdot \|\hat{y}_{\varepsilon'} - \hat{y}\|_{p_y}] + \Delta^s(\hat{x}, \hat{y}) \quad (19)$$

Переход к третьей строчке использует Липшицевость функции и показанное выше неравенство для M_x^s . Теперь ограничим $\|\hat{x}_{\varepsilon'} - \hat{x}\|_{p_x}$ и $\|\hat{y}_{\varepsilon'} - \hat{y}\|_{p_y}$, пользуясь сильной выпуклостью - сильной вогнутостью. Заметим, что эмпирическая функция тоже обладает этим свойством:

$$\|\hat{x}_{\varepsilon'} - \hat{x}\|_{p_x}^2 \leq \sqrt{\frac{2}{\mu_x} |\Phi^n(\hat{x}_{\varepsilon'}, y) - \Phi^n(\hat{x}, y)|} \leq \sqrt{\frac{2}{\mu_x} |\Phi^n(\hat{x}_{\varepsilon'}, \hat{y}) - \Phi^n(\hat{x}, \hat{y})|} \quad (20)$$

$$\|\hat{y}_{\varepsilon'} - \hat{y}\|_{p_y}^2 \leq \sqrt{\frac{2}{\mu_y} |\Phi^n(x, \hat{y}_{\varepsilon'}) - \Phi^n(x, \hat{y})|} \leq \sqrt{\frac{2}{\mu_y} |\Phi^n(\hat{x}, \hat{y}_{\varepsilon'}) - \Phi^n(\hat{x}, \hat{y})|} \quad (21)$$

Подставим в неравенство на зазор двойственности:

$$\Delta^s(\hat{x}_{\varepsilon'}, \hat{y}_{\varepsilon'}) \leq M_x^s \cdot \mathbb{E}_\xi \sqrt{\frac{2}{\mu_x} |\Phi^n(\hat{x}_{\varepsilon'}, \hat{y}) - \Phi^n(\hat{x}, \hat{y})|} + M_y^s \cdot \mathbb{E}_\xi \sqrt{\frac{2}{\mu_y} |\Phi^n(\hat{x}, \hat{y}_{\varepsilon'}) - \Phi^n(\hat{x}, \hat{y})|} + \Delta^s(\hat{x}, \hat{y}) \quad (22)$$

Введём $C' = \max \left\{ \sqrt{2(M_x^s)^2/\mu_x}, \sqrt{2(M_y^s)^2/\mu_y} \right\}$, и воспользуемся неравенством Йенсена:

$$\Delta^s(\hat{x}_{\varepsilon'}, \hat{y}_{\varepsilon'}) \leq C' \cdot \left(\sqrt{\mathbb{E}_\xi |\Phi^n(\hat{x}_{\varepsilon'}, \hat{y}) - \Phi^n(\hat{x}, \hat{y})|} + \sqrt{\mathbb{E}_\xi |\Phi^n(\hat{x}, \hat{y}_{\varepsilon'}) - \Phi^n(\hat{x}, \hat{y})|} \right) + \Delta^s(\hat{x}, \hat{y}) \quad (23)$$

$\varepsilon = \Delta^s(\hat{x}_{\varepsilon'}, \hat{y}_{\varepsilon'})$, точность нашего решения истинной задачи. Наша цель - свести правую часть неравенства к $\varepsilon' = \Delta_n^s(\hat{x}_{\varepsilon'}, \hat{y}_{\varepsilon'})$, точности решения эмпирической задачи. Обозначим:

$$\varepsilon'_x = \mathbb{E}_\xi |\Phi^n(\hat{x}_{\varepsilon'}, \hat{y}) - \Phi^n(\hat{x}, \hat{y})|, \quad \varepsilon'_y = \mathbb{E}_\xi |\Phi^n(\hat{x}, \hat{y}_{\varepsilon'}) - \Phi^n(\hat{x}, \hat{y})| \quad (24)$$

Заметим, что так как $\hat{\Phi}_n(\hat{x}_{\varepsilon'}, \hat{y}) \geq \hat{\Phi}_n(\hat{x}, \hat{y})$ и $\hat{\Phi}_n(\hat{x}, \hat{y}) \geq \hat{\Phi}_n(\hat{x}, \hat{y}_{\varepsilon'})$, то:

$$\varepsilon'_x + \varepsilon'_y = \mathbb{E}_\xi \left[\hat{\Phi}_n(\hat{x}_{\varepsilon'}, \hat{y}) - \hat{\Phi}_n(\hat{x}, \hat{y}) + \hat{\Phi}_n(\hat{x}, \hat{y}) - \hat{\Phi}_n(\hat{x}, \hat{y}_{\varepsilon'}) \right] = \quad (25)$$

$$= \mathbb{E}_\xi \left[\hat{\Phi}_n(\hat{x}_{\varepsilon'}, \hat{y}) - \hat{\Phi}_n(\hat{x}, \hat{y}_{\varepsilon'}) \right] \leq \mathbb{E}_\xi \left[\max_{y \in \mathcal{Y}} \hat{\Phi}_n(\hat{x}_{\varepsilon'}, \hat{y}) - \max_{x \in \mathcal{X}} \hat{\Phi}_n(\hat{x}, \hat{y}_{\varepsilon'}) \right] = \Delta_n^s(\hat{x}_{\varepsilon'}, \hat{y}_{\varepsilon'}) = \varepsilon' \quad (26)$$

Чтобы ограничить последнее слагаемое - зазор двойственности точного эмпирического решения - воспользуемся результатами Теоремы 3 из Zhang21 [3]:

$$\Delta^s(\hat{x}, \hat{y}) \leq \frac{2\sqrt{2}}{n} \cdot \sqrt{\frac{L_{xy}^2}{\mu_x \mu_y} + 1} \cdot \left(\frac{(M_x^s)^2}{\mu_x} + \frac{(M_y^s)^2}{\mu_y} \right) =: B \quad (27)$$

Нам остаётся записать (23) в терминах ε , ε'_x и ε'_y , подставив в него ограничение на $\Delta^s(\hat{x}, \hat{y})$:

$$\varepsilon \leq C' \cdot \left(\sqrt{\varepsilon'_x} + \sqrt{\varepsilon'_y} \right) + B \quad (28)$$

$$\left(\frac{\varepsilon - B}{C'} \right)^2 \leq \varepsilon'_x + \varepsilon'_y + 2\sqrt{\varepsilon'_x \varepsilon'_y} \leq 2 \cdot (\varepsilon'_x + \varepsilon'_y) \leq 2 \cdot \varepsilon' \quad (29)$$

$$\varepsilon \leq C' \sqrt{2 \cdot \varepsilon'} + B = C \sqrt{\varepsilon'} + B \quad (30)$$

Таким образом, мы вывели ограничение на точность решения седловой задачи через точность решения эмпирической. Важно отметить, что для ε точности решения истинной задачи нам необходима $\varepsilon' = \mathcal{O}(\varepsilon^2)$ точность решения эмпирической. Опять-таки, результат сошёлся с аналогичной теоремой для выпуклой оптимизации из диссертации [7]. Это ещё раз указывает на то, что седловая оптимизация - обобщение выпуклой, и зачастую результаты для второй переносятся на первую. В реальных задачах зачастую μ_x и μ_y выбираются порядка ε .

Если ввести $\mu_{min} = \min \{\mu_x, \mu_y\}$, то в силу $\varepsilon' \geq (\varepsilon - C)^2/B^2$, точность можно записать как:

$$\varepsilon' = \mathcal{O}(\mu_{\min} \cdot \varepsilon^2 + n^{-2} \cdot \mu_{\min}^{-3})$$

То есть при оценивании точности метрикой SGM на практике будет выходить третья степень зависимости от ε . Для уменьшения второго слагаемого нам необходимо больший размер выборки - в выпуклом случае было $n^{-2}\mu^{-1}$. На самом деле различие кроется в выборе метрики SGM , при выборе WGM потенциально можно получить асимптотически такой же результат.

2.1.1 Оптимальность полученной оценки

Возникает вопрос об оптимальности полученной оценки на точность эмпирического решения. На самом деле, пример, доказывающий её оптимальность, строится полностью аналогично выпуклому случаю в выражении 19 из Shalev-Shwartz et al. [1]. Рассмотрим функцию $\Phi(x, y) = \|x * \theta\| - \|y * \theta\| + \frac{C}{2}\|x\|^2 - \frac{C}{2}\|y\|^2$, где $\theta \in \Theta = \{0, 1\}^d$, $d = 2^n$. Если при этом θ равномерно сэмплируется из Θ n раз, то с вероятностью $1 - e^{-1}$ существует координата $j \in \{1, \dots, 2^n\} : \forall i \Rightarrow \theta_i[j] = 0$. В таком случае эмпирическая функция в точке (te_j, te_j) :

$$\Phi_n(te_j, te_j) = \frac{1}{n} \sum_{i=1}^n \left(\|te_j * \theta_i\| - \|te_j * \theta_i\| + \frac{1}{2C^2}\|te_j\|^2 - \frac{1}{2C^2}\|te_j\|^2 \right) \quad (31)$$

$$\varepsilon' = \Delta_n^s(te_j, te_j) = \frac{1}{n} \sum_{i=1}^n \left[\max_{y \in \mathcal{Y}} \Phi(te_j, y) - \min_{x \in \mathcal{X}} \Phi(x, te_j) \right] = \frac{1}{2C^2} (\|te_j\|^2 + \|te_j\|^2) = t^2/C^2 \quad (32)$$

Но при этом зазор двойственности истинной задачи:

$$\varepsilon = \Delta^s(te_j, te_j) = \mathbb{E}_\theta \left[\max_{y \in \mathcal{Y}} \Phi(te_j, y) - \min_{x \in \mathcal{X}} \Phi(x, te_j) \right] = t + t^2/C^2 \quad (33)$$

При $t = C\sqrt{\varepsilon'}$, подставив его в ограничение из теоремы, мы получим $C\sqrt{\varepsilon'} + \varepsilon' \leq C\sqrt{\varepsilon'} + B$, то есть ограничение на C в первом слагаемом полученной оценки оптимальное. Такого результата и следовало ожидать, ведь выпуклая оптимизация - частный случай выпукло-вогнутой. Так как в выпуклом случае аналогичная оценка является оптимальной, то и в более широком классе функций её некуда улучшать. В построенном примере ключевая разница между эмпирической и истинной функциями заключается в нерегуляризирующих (первых) слагаемых. Засчёт того, что размерность векторов 2^n (при выборке из n элементов), удаётся сконструировать пример, в котором точность эмпирического и истинного решения отличаются друг от друга на t . Как уже было сказано, эта идея является одним из основных результатов статьи Shalev-Shwartz et al [1], а здесь приведена лишь её адаптация к седловому случаю.

2.2 Зеркальный спуск на симплексах

Вернёмся к алгоритму зеркального спуска и опишем его применение для оптимизации седловых задач на симплексах. Для начала опишем, как выглядит наша задача:

$$\min_{x \in \Delta_n(1)} \max_{y \in \Delta_m(1)} \phi(x, y) \quad (34)$$

где $x \in \Delta_n(1)$ - точка на единичном симплексе размерности n , то есть x - вектор размерности n : $\|x\|_1 = 1, \forall i \Rightarrow x_i \geq 0$. Вектор x можно интерпретировать как распределение вероятностей по вершинам симплекса. Функция ϕ предполагается липшицевой и выпукло-вогнутой.

Для того, чтобы оптимизироваться на симплексах, мы будем использовать алгоритм Зеркального Спуска в евклидовой постановке. По сути это является обобщением всем хорошо известного градиентного спуска для симплексной постановки задачи.

Сначала опишем Prox-setup.

- 1 В качестве нормы мы будем использовать: $\|x, y\| = \sqrt{\alpha \|x\|_1^2 + \beta \|y\|_1^2}$, сопряженной к ней будет $\|\xi, \eta\|_* = \sqrt{\alpha^{-1} \|\xi\|_\infty^2 + \beta^{-1} \|\eta\|_\infty^2}$.
- 2 Функция, задающая расстояние - комбинация регуляризованных энтропий:

$$\omega(x, y) = \alpha \cdot (1 + \delta) \sum_{i=1}^n (x_i + \delta/n) \cdot \ln(x_i + \delta/n) + \beta \cdot (1 + \delta) \sum_{j=1}^m (y_j + \delta/n) \cdot \ln(y_j + \delta/n)$$
- 3 Из соображений минимизации ω -диаметра множества коэффициенты α, β выбирают:

$$\alpha = \left(\sqrt{\ln n} \cdot \left(\sqrt{\ln n} + \sqrt{\ln m} \right) \right)^{-1}, \beta = \left(\sqrt{\ln m} \cdot \left(\sqrt{\ln n} + \sqrt{\ln m} \right) \right)^{-1}$$

Напомним, что Прокс-отображение задаётся как:

$$Prox_x(\xi) = \arg \min_{y \in \mathcal{X}} \{V_x(y) + \langle \xi, y \rangle\} = \arg \min_{y \in \mathcal{X}} \{\omega(y) + \langle \xi - \omega'(x), y \rangle\}, \quad (35)$$

$$V_x(y) = \omega(y) - \langle \omega'(x), y - x \rangle - \omega(x) \geq \frac{1}{2} \|y - x\|^2 \quad (36)$$

Теперь, определив оракул первого рода для нашей функции как $F(x, y) = [F_x(x, y); F_y(x, y)] \in \partial_x(\phi(x, y)) \times \partial_y(-\phi(x, y))$, запишем алгоритм зеркального спуска для такой постановки:

$$z_0 = (x_0, y_0) \in \Delta_n \times \Delta_m = Z \quad (37)$$

$$z_{t+1} = Prox_{z_t}(\gamma_t F(z_t)) = \arg \min_{z \in Z} \{\omega(z) + \langle \gamma_t F(z_t) - \omega'(z_t), z \rangle\} \quad (38)$$

$$z^{(k)} = \left[\sum_{t=1}^k \gamma_t \right]^{-1} \sum_{t=1}^k \gamma_t z_t \quad (39)$$

2.2.1 Вывод явной формулы шага

Принципиальных изменений относительно ранее описанного алгоритма нет - мы так же шагаем по прокс-отображениям от точки к точке. Остаётся вывести явные формулы обновления z_t для нашей постановки задачи. Для удобства обозначим $g = \gamma_t F(z_t)$, выпишем прокс-отображение:

$\arg \min_z \{ \langle g, z \rangle + \omega(z) + \langle \omega'(z_t), z - z_t \rangle \}$ Нам необходимо ограничить $z^1 = x, z^2 = y$ на симплексы соответствующих размерностей. Делать это мы будем, введя множитель Лагранжа:

$$\arg \min_z \left\{ f(z) := \langle g, z \rangle + \omega(z) + \langle \omega'(z_t), z - z_t \rangle + \lambda_1 \cdot \left(\sum_{i=1}^n z_i^1 - 1 \right) + \lambda_2 \cdot \sum_{i=1}^m z_i^2 - 1 \right\} \quad (40)$$

Легко заметить, что процесс минимизации можно выписывать отдельно для z^1 и z^2 , так что проведём вывод только для z^1 :

$$\frac{\partial f}{\partial z^1} = g^1 + \omega'(z^1) - \omega'(z_t^1) + \lambda_1 \cdot \mathbf{1} = (1 + \delta)(1 + \ln(z^1 + \delta/n)) + g^1 - \omega'(z_t^1) + \lambda_1 \cdot \mathbf{1} = 0 \quad (41)$$

$$z^1 = \exp \left\{ \frac{\omega'(z_t^1) - g^1 - \lambda_1 \cdot \mathbf{1}}{1 + \delta} - 1 \right\} - \frac{\delta}{n} \quad (42)$$

Из этого выражения необходимо убрать λ_1 , для этого запишем изначальные ограничения на z^1 : $\|z^1\|_1 = 1, (z^1)_i \geq 0$. Второе условие заведомо выполняется из-за \exp , подставим в первое.

$$\sum_i (z^1)_i = \sum_i \exp \left\{ \frac{\omega'((z_t^1)_i) - g_i^1 - \lambda_1}{1 + \delta} - 1 \right\} - \delta = \quad (43)$$

$$= \sum_i \exp \left\{ \frac{\omega'((z_t^1)_i) - g_i^1}{1 + \delta} - 1 \right\} \cdot \exp \left\{ -\frac{\lambda_1}{1 + \delta} \right\} - \delta = 1 \quad (44)$$

$$\lambda_1 = -(1 + \delta) \cdot \ln \left(\frac{1 + \delta}{\sum \exp \{ \dots \}} \right) \Rightarrow \quad (45)$$

$$v^1 := \exp \left\{ \frac{\omega'(z_t^1) - g^1}{1 + \delta} - 1 \right\}, \quad z^1 = (1 + \delta) \cdot \frac{v^1}{\sum_i (v^1)_i} - \frac{\delta}{n} \quad (46)$$

Если отбросить δ , которые мы изначально используем ради численной стабильности операций, то формула становится значительно проще и читабельнее:

$$z^1 = \frac{\exp \{ \omega'(z_t^1) - g^1 - 1 \}}{\| \exp \{ \omega'(z_t^1) - g^1 - 1 \} \|_1} = \frac{\exp \{ \ln z_t^1 - g^1 \}}{\| \exp \{ \ln z_t^1 - g^1 \} \|_1} \quad (47)$$

$$z_{t+1} = \left(\frac{\exp \{ \ln x_t - g_x \}}{\| \exp \{ \ln x_t - g_x \} \|_1}; \frac{\exp \{ \ln y_t - g_y \}}{\| \exp \{ \ln y_t - g_y \} \|_1} \right) \quad (48)$$

То есть сложность шага оптимизации на симплексах не особо отличается от сложности вычисления оракула $F(z_t)$, так как все операции проводятся только над векторами, в отличие

от (как правило) вычисления $F(z_t)$. Таким образом, логично упростить вычисление детерминированного оракула, перейдя к стохастическому - как в SGD.

2.3 Стохастический зеркальный спуск

Введём стохастический оракул $G = G(z, \xi_t)$, где $\xi_1, \xi_2 \dots$ - независимые одинаково распределённые случайные величины, на каждом шаге добавляющие шум к вызову оракула; необходимо потребовать несмещённость G , а также ограниченность второго момента:

$$\mathbf{E}_\xi G(z, \xi) := F(z) = F(x, y) = [F_x(x, y); F_y(x, y)] \in \partial_x (\phi(x, y)) \times \partial_y (-\phi(x, y)) \quad (49)$$

$$\sup_z \mathbf{E}_\xi \|G(z, \xi)\|_*^2 < \infty \quad (50)$$

Сам алгоритм стохастического зеркального спуска не отличается от детерминированной версии, просто на каждом шаге зовётся стохастический оракул: $z_{t+1} = \text{Prox}_{z_t}(\gamma_t G(z_t, \xi_t))$. Возникает логичный вопрос - не потеряем ли мы в скорости сходимости алгоритма, введя такие на первый взгляд слабые ограничения на G ? Оказывается, что нет. Точность сходимости мы будем измерять по зазору двойственности, то есть $\varepsilon(\hat{x}, \hat{y}) = \max_y \phi(\hat{x}, y) - \min_x \phi(x, \hat{y})$. В теоремах 5.3.4 и 5.3.6 из книги Немировского [4] доказана оценка и для $F(z)$, и для $G(z, \xi)$: $\varepsilon(x_T, y_T) \leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$. В этой работе мы сверим практические результаты с этой оценкой.

2.3.1 Рандомизация при проектировании в зеркальном спуске на симплексах

Однако упрощение подсчёта градиента - не единственная причина использования стохастичности, в прикладных задачах зачастую рандомизация помогает находить лучшие локальные оптимумы. Некоторые онлайн-задачи по своей сути на каждом шаге оптимизации требуют выбор конкретной вершины симплекса, а не распределения на всех. Например, задача об одноруких бандитах или антагонистические матричные игры [8]. Но как оптимизироваться на пространствах из n или m векторов, соответствующих вершинам симплексов? Возникает идея, что не нужно отказываться от оптимизации распределения, на каждом шаге будем выбирать вершину согласно имеющемуся распределению, обновляя его соответственно.

$$p_{t+1} = \left(\frac{\exp\{\ln x_{(t)} - g_x\}}{\|\exp\{\ln x_{(t)} - g_x\}\|_1}, \frac{\exp\{\ln y_{(t)} - g_y\}}{\|\exp\{\ln y_{(t)} - g_y\}\|_1} \right) \quad (51)$$

$$\text{index}(t+1) \sim p_{t+1} \rightarrow z_{t+1} = e_{\text{index}(t+1)}, z_{(k)} = (x_{(k)}, y_{(k)}) = \left[\sum_{t=1}^k \gamma_t \right]^{-1} \sum_{t=1}^k \gamma_t z_t \quad (52)$$

2.3.2 Зеркальный спуск в задаче билинейной формы

Мы будем рассматривать задачи оптимизации билинейной формы, в общем виде представимые как $\phi(x, y) = y^T A x$. В таком случае детерминированный оракул принимает вид $F(x, y) = [A^T y, -Ax]$. Определим стохастический оракул как $G(x, y) = [A^T[j], -A[i]]$, $i \sim x, j \sim y$. То есть согласно распределению вектора y выбирается строка матрицы A (столбец транспонированной), а согласно распределению x - столбец.

Заметим, что вышеописанные алгоритмы решают задачу онлайн, то есть никак не ограничивая изменение оракула (матрицы A в случае билинейной формы) между разными итерациями процесса. В следующем разделе мы обсудим улучшение алгоритма оптимизации с использованием априорных знаний о матрице A - как в офлайн подходе.

2.4 Оптимизация алгоритма в офлайн-постановке

2.4.1 Анализ собственных значений

Рассматриваемый метод предложен в статье Azizian et al. [2] и опирается на анализ собственных значений якобиана оракула. Авторы начинают с ограничения спектра якобиана оракула $\mathbf{J}_F(z) = \nabla^2 \phi(z)$ через константы L -гладкости и μ -сильной выпуклости:

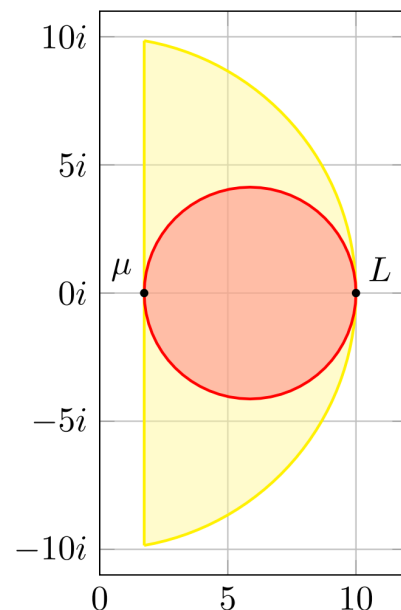
$$K = \{\lambda \in \mathbb{C} : 0 < \mu \leq \text{Re}(\lambda), |\lambda| < L\} \quad (53)$$

Здесь обе константы взяты для $\mathbf{J}_F(z)$, а не $\phi(z)$. На иллюстрации 2.1 из статьи, приведённой справа, обозначенное множество K - это жёлтый полукруг, образованный пересечением круга радиуса L с полуплоскостью. Основным теоретическим результатом авторов является ограничение скорости сходимости алгоритмов через оценивание их спектральных множеств снизу и сверху с помощью эллипсов - на картинке приведена оценка снизу кругом с центром в $\frac{1}{2}(\mu + L)$.

Теорема

Рассмотрим произвольный эллипс $E(a, b, c) \in \mathcal{C}$, симметричный относительно действительной оси и включающий в себя (включающийся в) множество K . Тогда для всех задач оптимизации: $\text{Span } \mathbf{J}_F(z) \in K$, метод градиентного спуска спуска с моментом Бориса Поляка, запущенный на параметрах эллип-

Рис. 2.1: Иллюстрация ограничений на спектр $\mathbf{J}_F(z)$. [2]



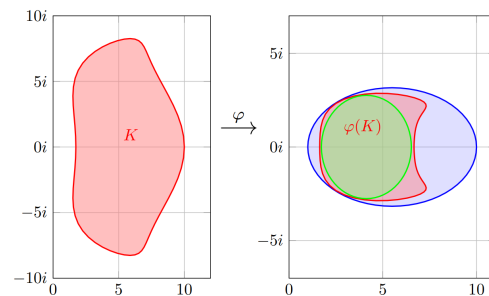
са, даёт верхнюю (нижнюю) оценку на скорость сходимости:

$$V(z, z_{prev}) = z - \alpha(a, b, c) \cdot F(z) + \beta(a, b, c) \cdot (z - z_{prev})$$

Это утверждение опирается на тот факт, что градиентный спуск с моментом является оптимальным методом для спектров эллипсоидной формы.

Ключевым переходом от утверждения этой теоремы к оптимизации алгоритмов является идея изменение самого оракула в нашей задаче. То есть множество K будет преобразовываться некоторым отображением $\varphi(K)$, причём нашей мотивацией будет ограничение новой формы спектра эллипсами, параллельными действительной оси. Воспользовавшись оптимальностью градиентного спуска с моментом, нам вероятно удастся ускорить имеющиеся алгоритмы. На Рис. 2.2 изображено изменение спектра отображением $\lambda \rightarrow \lambda(1 - \eta\lambda)$.

Рис. 2.2: Изменение спектра оракула для оптимальной оценки на эллипсах. Azizian et al. [2]



2.4.2 Билинейные матричные игры

В случае оптимизации $\min_{x \in X} \max_{y \in Y} \phi(x, y) := y^T A x$, якобиан $\mathbf{J}_F(x, y) = \nabla^2 \phi(x, y)$ - константный и кососимметричный. Выписав характеристический многочлен и воспользовавшись связью между собственными и сингулярными значениями матрицы, несложно показать, что в таком случае $Span \mathbf{J}_F(z) \subset [i\sigma_{min}(A), i\sigma_{max}(A)] \cup [-i\sigma_{min}(A), -i\sigma_{max}(A)]$. Введём

$$F^{real}(z) := \frac{1}{\eta} (F(z - \eta F(z)) - F(z)) \approx \nabla \left(\frac{1}{2} \|F(z)\|_2^2 \right) \quad (54)$$

Так как якобиан правой части - это $\mathbf{J}_{F^{real}}(z) = -\mathbf{J}_F^2(z)$, то все его собственные значения - действительные числа. Таким образом, $\varphi(K)$ легко ограничить сверху и снизу "оптимальными" эллипсами, гарантируя скорость сходимости соответственно градиентным спуском с моментом. Следуя такой интуиции, авторы предлагают улучшенный алгоритм градиентного спуска с моментом для решения задач билинейных матричных игр.

Утверждение

При выборе $\sqrt{\alpha} = \frac{2}{\sigma_{min} + \sigma_{max}}$, $\sqrt{\beta} = \frac{\sigma_{max} - \sigma_{min}}{\sigma_{min} + \sigma_{max}}$ и замене оракула $F \rightarrow F^{real}$, метод градиентного спуска с моментом:

$$z_{t+1} = z_t - \alpha F^{real}(z_t) + \beta(z_t - z_{t-1})$$

сходится с линейной скоростью $\mathcal{O} \left(\left(1 - \frac{2\sigma_{min}}{\sigma_{min} + \sigma_{max}} \right)^t \right)$, в то время как предыдущие наилучшие

результаты имели скорость порядка $\mathcal{O}\left(\left(1 - C \cdot \frac{\sigma_{min}^2}{\sigma_{max}^2}\right)^t\right)$.

В нашем случае момент предыдущего шага добавляется к изменённому оракулу и подаётся в прокс-отображение вместо изначального. Заметим, что при вычислении оракула $F^{real}(z)$ зовётся $F(z - \eta F(Z))$, так что в случае оптимизации на симплексах необходимо проецировать обе координаты $z - \eta F(z)$ обратно на симплексы. При достаточно малых η можно просто делить на первую норму векторов, но в строгом общем случае подходит проекция на симплекс из предыдущих разделов - по сути оператор софтмакс. В паре экспериментов со способами проекции не было замечено ощутимой разницы в сходимости методов, ведь η как правило очень малы. Итого, предложенный метод использует априорные знания о матрице билинейной игры - её минимальное и максимальное сингулярные значения. Их можно эффективно оценить, оптимизируя отношение Релэ или же запустив Truncated Randomized SVD. Так как этот офлайн подход не допускает изменения матрицы игры в итерационном процессе оптимизации, можно использовать приближительные знания о сингулярных значениях изменяющихся матриц, однако это зависит от природы конкретной задачи.

2.5 Сравнительный анализ описанных методов

Методы, описанные выше, сравнивались на двух задачах билинейных матричных игр. Обе постановки выбирались из соображений заранее известного правильного ответа, чтобы можно было оценивать сходимость методов не только относительно друг друга.

Поведение методов рассматривалось на нескольких запусках из случайных достаточно плохих точек с точки зрения трёх метрик:

- 1 Усреднённое отклонение от оптимального значения функции $\phi(x_{opt}, y_{opt})$:

$$\frac{1}{K} \sum_{k=1}^K |\phi_k(x_t, y_t) - \phi(x_{opt}, y_{opt})|$$

- 2 Усреднённый зазор двойственности:

$$\frac{1}{K} \sum_{k=1}^K (\max_{y \in \Delta_m} \phi_k(x_t, y) - \min_{x \in \Delta_n} \phi_k(x, y_t))$$

- 3 Усреднённое расстояние от найденного решения до теор. оптимума:

$$\frac{1}{K} \sum_{k=1}^K (\|x_t - x_{opt}\|_1 + \|y_t - y_{opt}\|_1)$$

В общем случае между собой сравниваем:

- 1 Детерминированный (deterministic) и стохастический (stochastic) оракулы
- 2 Рандомизацию на этапе проектирования на симплекс: с обычным (stochastic proj) и стохастическим (stochastic all) оракулами.

3 Адаптированную оптимизацию зеркального спуска с моментом (polyak optimized).

В скобках приведены названия методов на последующих графиках.

Помимо общих экспериментов проводились минорные, например: по влиянию разных констант α, β в прокс-сетапе, по влиянию разных стратегий выбора длины шага на скорость сходимости методов, по оптимальному выбору проекций в последнем алгоритме.

2.5.1 Сравнение на сэмплированных матрицах

Первый простой подход - сэмплирование элементов матрицы A из стандартного нормального распределения. В таком случае усреднённым оптимумом по нескольким матрицам A и нескольким стартовым позициям (x_0, y_0) очевидно будут вектора равномерного распределения, то есть $(1/n, \dots, 1/n)$ и $(1/m, \dots, 1/m)$. Усреднённым оптимальным значением функции является 0. Ниже приведены результаты сходимости методов в общем эксперименте на 10 сгенерированных матрицах A , для каждой из которых алгоритм запускался по 10 раз из точек (e_i, e_j) , $i \sim U[1, \dots, n]$, $j \sim U[1, \dots, m]$. Ниже приведены результаты соответствующих экспериментов для рассмотренных методов. Графики построены в логарифмической шкале по обеим осям, для удобства сравнения скорости сходимости алгоритмов с $1/\sqrt{T}$ и $1/T$. Теоретически доказанной оценкой является $\mathcal{O}(1/\sqrt{T})$, однако на практике методы могут сходиться быстрее - для этого рассматривается $1/T$.

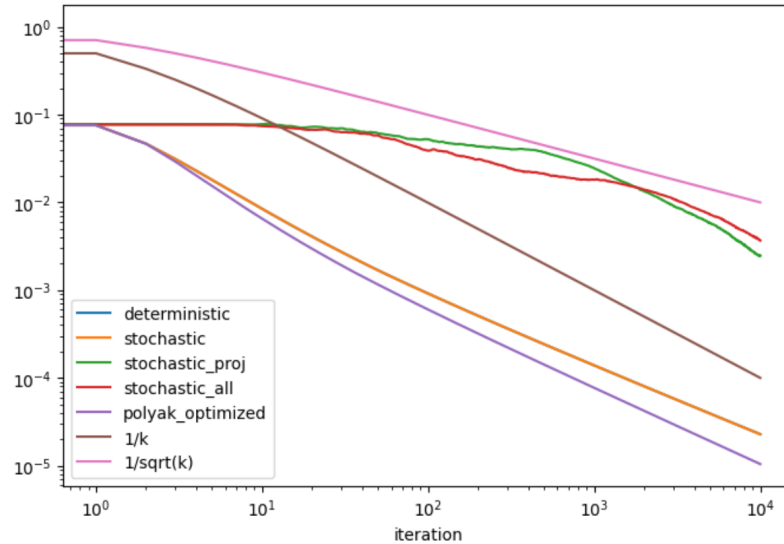


Рис. 2.3: Mean function value

На графике 2.3 отклонения значения функций от оптимального детерминированный и стохастический оракулы ведут себя идентично, вместе с оптимизированной версией градиентного спуска с моментом Бориса Поляка показывая линейную скорость сходимости. Последний из

них показывает лучшую скорость сходимости, при этом используя офлайн-информацию о сингулярных значениях матрицы A . Методы с рандомизацией при проектировании на симплекс показывают себя значительно медленнее, достигая скорости сходимости порядка корня лишь на достаточно большом количестве итераций. Причиной такого медленного старта является выбор начальных точек, поначалу метод практически не двигается по градиенту из-за того, что постоянно выбирает одни и те же координаты соответствующие единицам в векторах x и y . Эти методы позволяют выбирать ход на каждом шаге итерации в матричной игре, в отличие от оценивания распределения предыдущими, однако потеря в скорости сходимости из-за конкретного выбора координаты, как видно, существенна. Тем не менее, методы сходятся как $\mathcal{O}(1/\sqrt{T})$. Далее приведены графики 2.4, 2.5 отклонений по другим метрикам, более-менее соответствующие описанному выше.

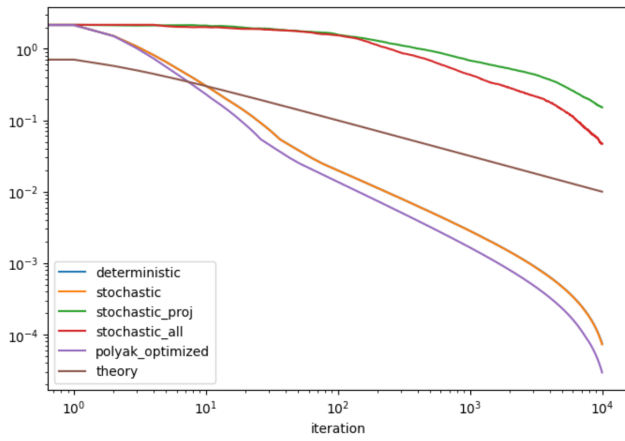


Рис. 2.4: Mean duality gap

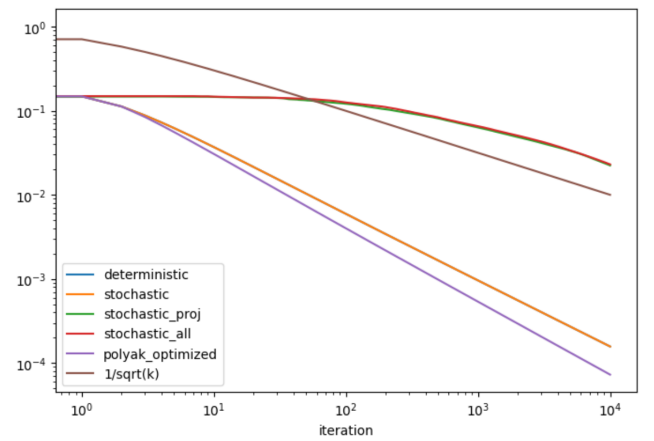


Рис. 2.5: Mean distance

2.5.2 Описание способа нахождения векторов PageRank

В качестве второй постановки с заведомо известным теоретическим решением была выбрана задача PageRank. Для начала опишем саму задачу, а потом уже перейдём к оптимизации соответствующей билинейной формы.

Мы хотели бы оценить распределение вероятностей перехода на n страницах. Пусть задана матрица вероятностей перехода в веб-графе P , тогда в каждый момент времени t вектор вероятностей перехода по страницам задаётся как $p^T(t+1) = p^T(t)P$, а вектором PageRank называется $v = \lim_{t \rightarrow \infty} p(t)$, его можно искать как $v^T = v^T P \Leftrightarrow (P^T - I)v = 0$.

Запишем задачу в виде билинейной формы:

$$\min_{v \in \Delta_n(1)} \max_{u \in \Delta_n(1)} \phi(v, u) := u^T (P^T - I)v$$

Тогда седловой точкой как раз будет $\phi(v_{opt}, u_{opt}) = 0$, и именно v_{opt} будет вектором PageRank. Эту задачу можно привести к случаю симметричной матрицы билинейной формы:

$$\min_{x \in \Delta_{2n+1}(1)} \max_{y \in \Delta_{2n+1}(1)} y^T A x, \quad \text{где } A := \begin{pmatrix} 0 & P^T - I & -e \\ I - P & 0 & e \\ e^T & -e^T & 0 \end{pmatrix}$$

Как раз в этой постановке мы сравниваем работу методов, ожидая скорость сходимости отклонения от оптимального значения функции порядка или выше теоретической границы $\mathcal{O}(1/\sqrt{T})$.

2.5.3 Сравнение на циклических веб-графах

Помимо известного значения функции в оптимуме, предпочтительно также считать и расстояние до оптимального решения на пространстве векторов PageRank. Очевидным примером известного вектора являются циклические веб-графы. Пусть вероятности перехода от i -й вершины к соседним равны 0.25, а вероятность остаться в этой вершине в следующий момент времени 0.5, то есть это буквально циклический граф. Тогда вектор PageRank - это равномерное распределение вероятностей по страницам. Первым графиком 2.6 по результатам экспериментов опять рассмотрим среднее отклонение значения функции:

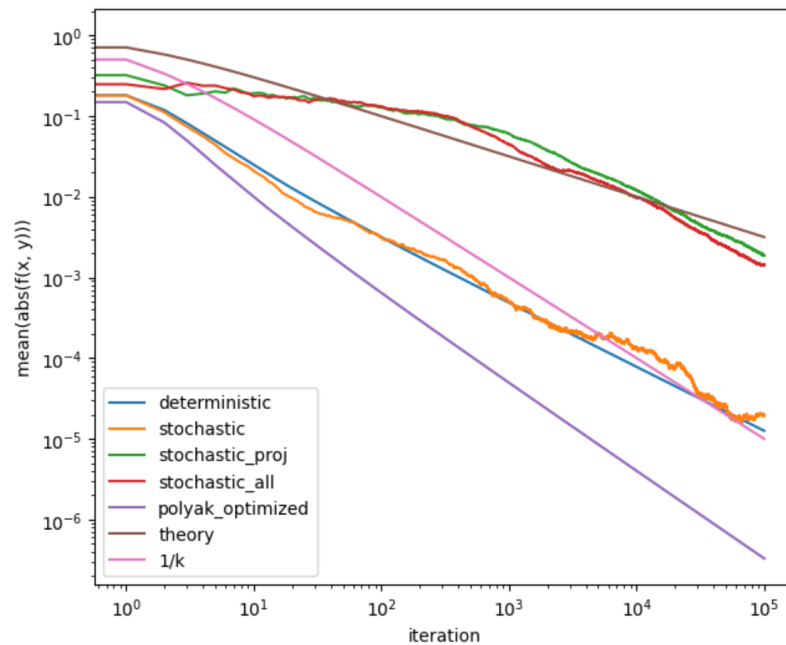


Рис. 2.6: Mean function value

На этом графике ещё ярче подтверждаются рассуждения предыдущей части, разница между оптимизированным методом Поляка и стохастическим / детерминированным оракулом вид-

на намного сильнее. Засчёт отсутствия усреднений по сэмплированным матрицам заметно поведение при обычной стохастичности по колонкам матрицы. Методы рандомизации при проекции на симплекс, как стохастический по колонкам, так и нет, постепенно нагоняют других, но при этом на 10^5 итераций, проделанных для этой задачи, всё ещё остаются значительно позади. Тем не менее, важно отметить, что они начинают сходиться быстрее $O(1/\sqrt{T})$.

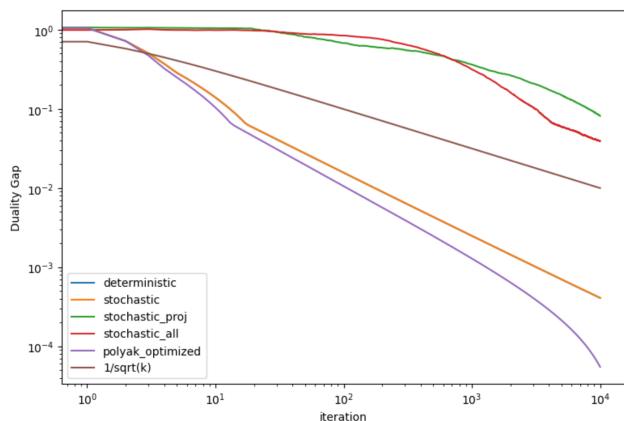


Рис. 2.7: Mean duality gap

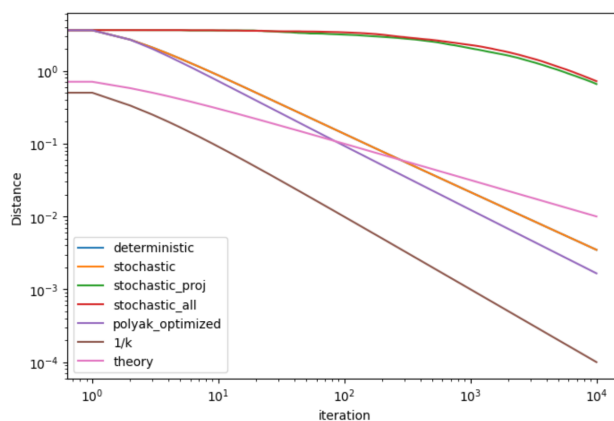


Рис. 2.8: Mean distance

Картина по скорости сходимости для зазора двойственности 2.7 и расстоянию до решения на пространстве оптимизируемых векторов 2.8 остаётся примерно такой же. На последнем графике заметно, что даже спустя 10^4 итераций алгоритма, решения методов рандомизации при проецировании существенно хуже, чем решения обычных стохастических и оптимизированного.

3 Выводы

В процессе работы над курсовым проектом было изучено несколько статей из предметной области теоретических оценок сходимости метода Монте-Карло. Основной задачей этой части проекта был перенос доказанного теоретического результата о необходимой точности решения эмпирической задачи с задач выпуклой оптимизации на задачи выпукло-вогнутой. Этот результат стал частью статьи для журнала Компьютерные Исследования и Моделирование. Во второй части курсового проекта была изучена литература по методам оптимизации седловых задач на симплексах, был проведён сравнительный анализ имплементаций этих методов на двух постановках задач оптимизации билинейной формы. В частности, сравнение показало значительно более медленную скорость сходимости зеркального спуска с рандомизацией при проекции относительно других методов, а также ощутимое улучшение оптимизированного метода градиентного спуска Бориса Поляка относительно обычного зеркального спуска.

Список литературы

- [1] Shalev-Shwartz et al. “Stochastic Convex Optimization”. В: (2009).
- [2] W.Azizian et al. “Accelerating Smooth Games by Manipulating Spectral Shapes”. В: (2020).
- [3] Zhang et al. “Generalization Bounds for Stochastic Saddle Point Problems”. В: (2020).
- [4] Aharon Bet-Tal и Arkadi Nemirovski. *Lectures on Modern Convex Optimization*. 2022, с. 400—441.
- [5] S´ebastien Bubeck. *Theory of Convex Optimization for Machine Learning*. 2015.
- [6] John C. Duchi. *Introductory Lectures on Stochastic Optimization*. 2010.
- [7] Darina Dvinskikh. *Decentralized Algorithms for Wasserstein Barycenters, Dissertation*. 2021.
- [8] В.Г. Спокойный А.В.Гасников Ю.Е. Нестеров. “Об эффективности одного метода рандомизации зеркального спуска в задачах онлайн оптимизации”. В: (2015).
- [9] Д.Ю.Дмитриев А.В.Гасников. “Об эффективных рандомизированных алгоритмах поиска вектора PageRank”. В: (2015).
- [10] Л.Г.Хачиян. *Избранные труды*. 2014, с. 38—49.