УДК 519.688

А. В. Гасников[1,2], *П. Е. Двуреченский*[2,4], *В. Г. Спокойный*[1,2,3,4], *П. И. Стецюк*[5], *А. Л. Суворикова*[1,2,6]

[1]Московский физико-технический институт (государственный университет)
[2]Институт проблем передачи информации РАН
[3]Национальный исследовательский университет Высшая школа экономики
[4]Weierstrass Institute for Applied Analysis and Stochastics, Berlin
[5]Институт кибернетики им. В. М. Глушкова НАН Украины
[6]The International Research Training Group 1792, Berlin

Суперпозиция метода балансировки и универсального градиентного метода для поиска энтропийно-регуляризованного барицентра Вассерштейна и равновесий в многостадийных моделях транспортных потоков

Представлен обзор современных численных методов поиска барицентра Вассерштейна конечного семейства вероятностных мер с одинаковым конечным носителем. Такие задачи в последнее время стали очень популярны в связи с всевозможными приложениями к сравнительному анализу изображений, в частности, к обнаружению разладок в ряде изображений. Например, подобные задачи возникают при изучении деятельности головного мозга.

В основном мы исходили из цикла работ M. Cuturi с соавторами. Общая идея этих работ – найти барицентр вероятностных мер согласно энтропийно-регуляризованному расстоянию Вассерштейна. Такое (регуляризованное) расстояние можно заметно эффективнее посчитать, чем исходное расстояние Вассерштейна. В одной из работ отмеченного цикла [8] содержалась идея сочетания метода Синхорна (балансировки) для решения внутренней задачи (расчета соответствующих регуляризованных расстояний и их субградиентов) и быстрого градиентного метода для решения внешней задачи (поиск барицентра). К сожалению, в описанном авторами виде метод оказался не пригодным для использования на практике (не было также никаких теоретических гарантий его сходимости). В [1] показано, как можно доработать данный метод (в частности, доказана сходимость предложенной модификации). Однако мы были сконцентрированы на другом приложении (к поиску равновесий в многостадийных транспортных моделях).

В данной работе мы рассматриваем оба отмеченных приложения. Главным результатом работы является разработка в общем случае (не только для этих двух приложений) концепции суперпозиции методов, когда мы можем выделить в исходной задаче часть переменных, по которым задача эффективно решается внутреннем методом (но допускается, что лишь приближенно) при замороженных остальных переменных. А по оставшейся группе переменных запускается внешний метод, на каждой итерации которого требуется запускать внутренний метод. В статье получен частичный ответ на довольно общий вопрос: как оптимально сочетать эти методы (внутренней и внешний), т.е. насколько точно надо решать на каждой итерации внешнего метода внутреннюю задачу, чтобы минимизировать общее время работы метода при заданной точности решения, которую хотим получить?

Ключевые слова: седловая задача, энтропия, метод балансировки, метод Синхорна, универсальный метод, неточный оракул, задача Монжа–Канторовича, расстояние Вассерштейна, барицентр Вассерштейна.

A. V. Gasnikov[1,2], *P. E. Dvurechensky*[2,4], *V. G. Spokoiny*[1,2,3,4], *P. I. Stetsyuk*[5], *A. L. Suvorikova*[1,2,6]

[1]Moscow Institute of Physics and Technology (State University)
[2]Institute for Information Transmission Problems RAS
[3]Higher School of Economics National Research University
[4]Weierstrass Institute for Applied Analysis and Stochastics, Berlin
[5]V. M. Glushkov Institute of Cybernetics of NAS of Ukraine
[6]The International Research Training Group 1792, Berlin

Superposition of the balancing algorithm and the universal gradient method for search of the regularized Wasserstein brycenter and equilibria in multistage transport models

The paper presents a survey of modern computational algorithms the mean (in the sense of 2–Wasserstein distance) of a finite set of empirical probability measures. Each measure is assumed to have the same support. The problem is quite popular due to numerous applications to comparative image analysis, for example the detection of structural changes in a set of fMRI images.

The work is inspired by a number of papers written by M. Cuturi and his colleagues. They propose to use regularized 2–Wasserstein distance instead of the original one. The new approach significantly improves computational efficiency. The paper [8] proposes to combine Sinkhorn algorithm with the fast gradient method. The first algorithm computes the regularized distance and its subgradient, while the second one is used for barycenter search. However, this approach uses some predefined parameters that cannot be selected in practice. Moreover, there is no theoretical guarantee that it converges. The paper [1] proposes a modification of the algorithm, which solves the above problems.

The work [1] is mainly focused on searching equilibria in the multistage traffic model, whereas this paper investigates its application to the search for Wasserstein barycenters. The key result is a general concept that allows constructing the superposition of algorithms.

It is used if the assumption holds. The set of variables can be divided into two groups (A) and (B), s.t. by fixing (A) one can obtain (an inexact) solution by (B). Thus, the initial problem can be reduced to successive optimization by (A) and (B). The present work aims to answer the following questions: how to combine the steps and how accurate one should be when optimizing by (B) so that to achieve the predefined accuracy?

Key words: saddle point problem, entropy, balancing algorithm, Sinkhorn's algorithm, universal gradient method, inexact oracle, Monge–Kantorovich problem, Wasserstein distance, Wasserstein barycenter.

1. Введение

Данная статья представляет собой продолжение недавно опубликованной работы [1]. Для удобства мы повторим во введении постановку задачи и основные обозначения.

Поиск (стохастических) равновесий в многостадийных моделях транспортных потоков приводит к решению следующей седловой задачи с правильной (выпукло-вогнутой) структурой [2] – [5]:

$$\min_{\substack{\sum_{j=1}^n x_{ij} = L_i,\, \sum_{i=1}^n x_{ij} = W_j \\ x_{ij} \geqslant 0,\, i,j = 1,\ldots,n}} \max_{y \in Q} \Big\{ \sum_{i,j=1}^n x_{ij} \ln x_{ij} + \sum_{i,j=1}^n c_{ij}(y) x_{ij} + g(y) \Big\} =$$

$$= \max_{y \in Q} \max_{\lambda,\mu \in \mathbb{R}^n} \Big\{ \langle \lambda, L \rangle + \langle \mu, W \rangle - \sum_{i,j=1}^n \exp\big(-c_{ij}(y) - 1 + \lambda_i + \mu_j\big) + g(y) \Big\}, \qquad (1)$$

где $g(y)$ и $c_{ij}(y) \geqslant 0$ – вогнутые гладкие функции (если ищутся не стохастические равновесия, то $c_{ij}(y)$ могут быть негладкими), Q – множество простой структуры, например,

$$Q = \{y \in \mathbb{R}^n : y \geqslant \bar{y}\}.$$

Легко понять, что система балансовых ограничений в (1) либо несовместна $\sum\limits_{i=1}^{n} L_i \neq \sum\limits_{j=1}^{n} W_j$, либо вырождена (имеет не полный ранг). В последнем случае это приводит к тому, что двойственные переменные (λ, μ) определены с точностью до произвольной постоянной C:

$$(\lambda + Ce, \mu - Ce), \quad e = \underbrace{(1, ..., 1)}_{n}.$$

Задачу (1) также можно переписать следующим образом (не ограничивая общности, считаем $\sum\limits_{i=1}^{n} L_i = \sum\limits_{j=1}^{n} W_j = 1$):

$$\min_{\substack{\sum\limits_{j=1}^{n} x_{ij} = L_i,\ \sum\limits_{i=1}^{n} x_{ij} = W_j \\ x_{ij} \geqslant 0,\ \sum\limits_{i,j=1}^{n} x_{ij} = 1,\, i,j = 1, ..., n}} \max_{y \in Q} \Big\{ \sum_{i,j=1}^{n} x_{ij} \ln x_{ij} + \sum_{i,j=1}^{n} c_{ij}(y) x_{ij} + g(y) \Big\} =$$

$$= \max_{y \in Q} \max_{\lambda, \mu \in \mathbb{R}^n} \Big\{ \langle \lambda, L \rangle + \langle \mu, W \rangle - \ln \Big[\sum_{i,j=1}^{n} \exp\big(-c_{ij}(y) + \lambda_i + \mu_j\big) \Big] + g(y) \Big\} = \tag{2}$$

$$= - \min_{y \in Q} f(y),$$

где выпуклая функция $f(y)$ определяется как

$$f(y) = \max_{\substack{\sum\limits_{j=1}^{n} x_{ij} = L_i,\ \sum\limits_{i=1}^{n} x_{ij} = W_j \\ x_{ij} \geqslant 0,\ \sum\limits_{i,j=1}^{n,n} x_{ij} = 1,\, i,j = 1, ..., n}} \Big\{ - \sum_{i,j=1}^{n} x_{ij} \ln x_{ij} - \sum_{i,j=1}^{n} c_{ij}(y) x_{ij} - g(y) \Big\} =$$

$$= \min_{\lambda, \mu \in \mathbb{R}^n} \Big\{ \ln \Big[\sum_{i,j=1}^{n} \exp(-c_{ij}(y) + \lambda_i + \mu_j) \Big] - \langle \lambda, L \rangle - \langle \mu, W \rangle - g(y) \Big\}. \tag{3}$$

Поскольку мы добавили в ограничения условие $\sum\limits_{i,j=1}^{n} x_{ij} = 1$, являющееся следствием балансовых уравнений, то это привело к тому, что двойственные переменные (λ, μ) определены с точностью до двух произвольных постоянных C_λ, C_μ: $(\lambda + C_\lambda e, \mu + C_\mu e)$.

Заметим, что расчет градиента $\nabla f(y)$ (в ряде транспортных приложений вогнутые функции $c_{ij}(y)$ – негладкие, тогда вместо градиентов стоит понимать суперградиенты $c_{ij}(y)$ и субградиент $f(y)$) осуществляется по следующей формуле (Демьянова–Данскина–Рубинова, см., например, [2, 6]):

$$\nabla f(y) = - \frac{\sum\limits_{i,j=1}^{n} \exp\big(-c_{ij}(y) + \lambda_i^* + \mu_j^*\big) \nabla c_{ij}(y)}{\sum\limits_{i,j=1}^{n} \exp\big(-c_{ij}(y) + \lambda_i^* + \mu_j^*\big)} - \nabla g(y) =$$

$$= - \sum_{i,j=1}^{n} x_{ij}(\lambda^*, \mu^*) \nabla c_{ij}(y) - \nabla g(y), \tag{4}$$

где (λ^*, μ^*) – решение задачи (3), не важно какое именно, градиент $\nabla f(y)$ от выбора C_λ, C_μ (см. выше) не зависит. В данной статье мы (так же, как и в [1]) ограничимся изучением только полноградиентных методов для задачи (2), т.е. не будем рассматривать, например, рандомизацию при вычислении градиента по формуле (4). Планируется отдельно исследовать вопрос о возможности ускорения вычислений за счет введения рандомизации для

внешней задачи. На данный момент нам представляется (см. формулу (13) в п. 3), что это может принести дивиденды только в случае, когда вспомогательная задача расчета $\nabla c_{ij}(y)$ достаточно сложная. Тут требуется много оговорок, в частности, в большинстве приложений умение рассчитывать $\nabla c_{ij}(y)$ для конкретной пары (i,j) без дополнительных затрат позволяет заодно рассчитать и все $\nabla c_{ij}(y)$, $j = 1, ..., n$. Также отдельно планируется исследовать вопрос о том, какие подходы и насколько хорошо допускают распараллеливание. Вопрос о целесообразности рандомизации оказывается завязанным и на вопрос о возможности распараллеливания.

Структура статьи следующая. В п. 2 мы рассматриваем популярную в последнее время (в связи с большим числом приложений) задачу вычисления барицентра Вассерштейна различных вероятностных мер [7]. Эта задача оказывается тесно связанной с задачей (1). Мы разбираем в статье этот пример, потому что он хорошо проясняет возможные альтернативы предлагаемому нами основному подходу решения задач (1), (2), изложенному в п. 3. В основе подхода п. 3 (см. также [1, 8]) лежит сочетание метода балансировки для решения внутренней задачи оптимизации по двойственным множителям и универсального градиентного метода с неточным оракулом для внешней задачи (2). В работе [1] мы сконцентрировались на обобщении универсального метода на случай неточного оракула. В данной работе акцентируем внимание на получении условий на шум оракула для внешней задачи, исходя из оценок точности решения внутренней задачи. Таким образом, данная статья дополняет работу [1] в теоретическом плане. Практическим экспериментам планируется посвятить отдельную работу.

2. Поиск барицентра Вассерштейна

К похожей на (1) задаче приводит поиск барицентра Монжа–Канторовича (в западной литературе чаще говорят барицентра Вассерштейна [7, 8])[1]. Изложим вкратце постановку задачи [8] – [10]. Вводится энтропийно регуляризованное транспортное расстояние (см. рис. 1 в [11]) с матрицей $\|c_{ij}\|_{i,j}^{n,n}$, сформированной из квадратов попарных расстояний $c_{ij} = l_{ij}^2$ от носителя i меры L до носителя j меры W ($L, W \in S_n(1)$):

$$\Delta(L,W) = \min_{\substack{\sum\limits_{j=1}^{n} x_{ij} = L_i,\ \sum\limits_{i=1}^{n} x_{ij} = W_j \\ x_{ij} \geqslant 0,\, i,j = 1,\ldots,n}} \Big\{\gamma \sum_{i,j=1}^{n} x_{ij} \ln x_{ij} + \sum_{i,j=1}^{n} c_{ij} x_{ij}\Big\} =$$

$$= \max_{\lambda,\mu \in \mathbb{R}^n} \left\{ \langle \lambda, L\rangle + \langle \mu, W\rangle - \gamma \sum_{i,j=1}^{n} \exp\Big(\frac{-c_{ij} + \lambda_i + \mu_j}{\gamma} - 1\Big) \right\} =$$

$$= \max_{\lambda \in \mathbb{R}^n} \Bigg\{ \langle \lambda, L\rangle - \underbrace{\gamma \sum_{j=1}^{n} W_j \ln\Big[\frac{1}{W_j} \sum_{i=1}^{n} \exp\Big(\frac{-c_{ij} + \lambda_i}{\gamma}\Big)\Big]}_{H_W^*(\lambda)} \Bigg\}. \tag{5}$$

Определим при $L \in S_n(1)$ функцию $H_W(L) = \Delta(L,W)$. Эта гладкая на $L \in S_n(1)$ функция с градиентом (см. утверждение 3 [10]):

$$\nabla H_W(L) = \lambda^*,$$

[1]Строго говоря, мы будем искать барицентр вероятностных мер не согласно настоящему (негладкому) расстоянию Вассерштейна (на наш взгляд, исторически более правильно это расстояние называть расстоянием Монжа–Канторовича–Добрушина), как это можно было подумать из названия, а согласно энтропийно-регуляризованному расстоянию Вассерштейна [8].

где λ^* – единственное решение (5), удовлетворяющее условию[2] $\langle \lambda^*, e\rangle = 0$. Отсюда следует, что

$$H^*_W(\lambda) = \max_{L\in S_n(1)}\Big\{\langle \lambda, L\rangle - H_W(L)\Big\} = \\ = \gamma\sum_{j=1}^n W_j \ln\Big[\frac{1}{W_j}\sum_{i=1}^n \exp\Big(\frac{-c_{ij}+\lambda_i}{\gamma}\Big)\Big].$$

Теперь можно перейти к изложению основной конструкции. Задача поиска барицентра Вассерштейна[3] записывается следующим образом:

$$\sum_{k=1}^m H_{W_k}(L) \to \min_{L\in S_n(1)}. \tag{6}$$

К сожалению, в такой формулировке мы не можем оценить константу Липшица градиента функционала (6), явно входящую в большинство современных быстрых (ускоренных) численных методов. Однако оказывается (см. п. 3), что существуют быстрые методы, которым для работы не требуется такая информация (константа Липшица градиента).

Перепишем задачу (6), следуя п. 3 работы [10], следующим образом:

$$-\sum_{k=1}^m H_{W_k}(L_k) \to \max_{\substack{L_1 = L_m|\ \lambda^1,\, L_1\in S_n(1)\\ \dots \\ L_{m-1} = L_m|\ \lambda^{m-1},\, L_m\in S_n(1)}},$$

$$\sum_{k=1}^{m-1}\max_{L_k\in S_n(1)}\Big\{\langle\lambda^k, L_k\rangle - H_{W_k}(L_k)\Big\} + \max_{L_m\in S_n(1)}\Big\{\langle -\sum_{k=1}^{m-1}\lambda^k, L_m\rangle - H_{W_m}(L_m)\Big\} \to \min_{\lambda^1,\dots,\lambda^{m-1}},$$

$$\sum_{k=1}^{m-1} H^*_{W_k}(\lambda^k) + H^*_{W_m}\Big(-\sum_{k=1}^{m-1}\lambda^k\Big) \to \min_{\lambda^1,\dots,\lambda^{m-1}}, \tag{7}$$

$L_* = \nabla H^*_{W_k}(\lambda^k_*)$ для любого $k = 1, \dots, m-1$, где L_* – единственное решение задачи (6), $\{\lambda^k_*\}_{k=1}^{m-1}$ – единственное решение задачи (7). Важное свойство функционала задачи (7) – равномерная ограниченность константы Липшица градиента (следует из [15]). Задача безусловной минимизации (7) может быть эффективно решена различными способами (в зависимости от того, насколько велики n и m). В частности, для больших n и m неплохо с задачей справляются различные модификации метода сопряженных градиентов и быстрых градиентных методов [10]. Структура задачи (7) позволяет эффективно использовать покомпонентные методы (см., например, [16,17]), которые к тому же хорошо параллелятся для данной задачи. Задача (7) хорошо также решается с помощью распределенных вычислений [18].

В приложениях к поиску разладки требуется много раз перерешивать задачу (7), которую для симметричности перепишем следующим образом:

$$\sum_{k=1}^m H^*_{W_k}(\lambda^k) \to \min_{\sum\limits_{k=1}^m \lambda^k = 0},$$

[2]Решая задачу (6) каким-нибудь прокс-методом с KL прокс-структурой [12], легко понять, что от того, как именно выбирать λ^*, задаваемое с точностью до сдвига всех компонент на одно и то же произвольное число, метод зависеть не будет. Единственное, для чего имеет смысл стремиться к выполнению этого нормирующего условия, так это для лучшей практической обработки (меньшее накопление ошибок округления из-за конечной длины мантиссы) экспоненциального взвешивания компонент градиента, возникающего на каждом шаге итерационного процесса при выборе KL прокс-структуры.

[3]К сожалению, пока не так много известно о статистической обоснованности использования расстояния Вассерштейна. Другими словами, хотелось бы иметь связь барицентра Вассерштейна с оценками максимального правдоподобия, ну или хотя бы с состоятельными оценками для соответствующих схем экспериментов. Пока установлена только связь с состоятельными оценками [13,14].

немного смещая окно, т.е. заменяя каждый раз несколько первых слагаемых в сумме

$$H^*_{W_1}(\lambda^1), .., H^*_{W_r}(\lambda^r)$$

на столько же новых (которые, как ожидается, близки к $H^*_{W_m}(\lambda^m)$). В таком случае предлагается в итерационном процессе стартовать при сдвиге окошка с того, на чем остановились на прошлом положении окошка, экстраполируя λ^m_* на вновь пришедшие слагаемые. Ясно, что для новой задачи набор

$$(\lambda^{r+1}_*, ..., \lambda^{m-1}_*, \lambda^m_*, \underbrace{\lambda^m_*, ..., \lambda^m_*}_{r}.),$$

с которого стартуем, уже не будет оптимальным, однако мы вправе надеяться на его близость к оптимальному набору, что существенно сокращает число последующих итераций. Интересной, особенно в данном контексте, представляется возможность использования (и интерпретации) распределенных вычислений [18].

Может показаться, что подход, сводящий поиск решения задачи (6) к задаче (7), не доминируем, поскольку, в отличие от задачи (6), в задаче (7) мы можем явно выписать функционал и по простым формулам, рассчитать градиент, который к тому же имеет равномерно ограниченную константу Липшица. С одной стороны, это, действительно, преимущество, но получено оно дорогой ценой – ценой раздутия пространства, в котором происходит оптимизация почти в m раз. И это раздутие скажется не только на сложности одной итерации. Для задачи (6) осуществление одной итерации будет еще более дорогим ввиду необходимости на каждой итерации решать m отдельных подзадач расчета $\nabla H_{W_k}(L)$. Скажется это, прежде всего, на числе необходимых итераций. В следующем разделе будет отмечено, что расчет $\nabla H_{W_k}(L)$ с помощью метода балансировки не намного сложнее расчета $\nabla H^*_{W_k}(\lambda^k)$. При этом задача (6) решается в пространстве намного меньшей размерности, и мы вправе ожидать, что необходимое число итераций может быть намного меньше, чем для задачи (7). Кроме того, задача (6) решается на компакте, т.е. размер решения (если быть точным, то расстояние от точки старта до решения), входящий в оценку необходимого числа итераций, заведомо ограничен размером симплекса. Задача (7) – задача безусловной оптимизации, причем без свойства сильной выпуклости функционала. Размер ее решения может быть большим, и входит он в оценки необходимого числа итераций так же, как и для задачи (6), к сожалению, степенным образом (для быстрых (ускоренных) методов можно ожидать линейной зависимости необходимого числа итераций от этого размера). Наконец, для постановок задач об обнаружении разладки (см. выше) также ожидается, что использовать близость решений прямых задач (6) при смещении окошка удастся намного лучше, чем близость в решении двойственных задач (7). В итоге выгода от подхода, связанного с переходом к задаче (7), уже не столь очевидна и требует отдельного и более аккуратного исследования с решающей ролью численных экспериментов.

В ряде задач требуется искать параметрический барицентр Вассерштейна. В таком случае в одном из вариантов постановки предполагают наличие параметрической зависимости $L(\theta) \in S_n(1)$, $\theta \in \Theta$, где размерность вектора параметров $\dim\theta \ll n$. К сожалению, в этом случае нельзя гарантировать с помощью стандартных приемов [19, с. 86] выпуклости задачи

$$\sum_{k=1}^{m} H_{W_k}\big(L(\theta)\big) \to \min_{\theta\in\Theta}, \tag{8}$$

за исключением случая, когда $L(\theta) = A\theta + b$, а Θ – выпуклое множество. В этом случае конструкция (7) видоизменяется следующим образом:

$$-\sum_{k=1}^{m} H_{W_k}\big(L_k\big) \to \max_{\substack{L_1 = L_m|\ \lambda^1,\ L_1 \in S_n(1)\\ \cdots \\ L_{m-1} = L_m|\ \lambda^{m-1},\ L_{m-1} \in S_n(1)\\ L_m = A\theta + b|\ \tilde\lambda,\ L_m \in S_n(1), \theta \in \Theta}},$$

$$\sum_{k=1}^{m-1} H^*_{W_k}(\lambda^k) + H^*_{W_m}\Big(-\sum_{k=1}^{m-1}\lambda^k - \tilde{\lambda}\Big) + \langle \tilde{\lambda}, A\theta + b\rangle \to \min_{\substack{\lambda^1,\,...,\,\lambda^{m-1},\,\tilde{\lambda}\\ \theta\,\in\,\Theta}}, \tag{9}$$

$L_* = \nabla H^*_{W_k}(\lambda^k_*)$ для любого $k = 1, ..., m-1$. Как следствие, нет никаких гарантий, что изложенная выше конструкция, связанная с переходом к двойственной задаче (7) и восстановлением решения прямой задачи (6) по явным формулам через двойственные множители, в общем случае будет работать хотя бы для поиска локальных решений.[4] Другими словами, необходимо искать глобальный оптимум задачи (8) исходя из работы с прямой задачей (8). Один из вариантов того, как это можно делать, будет описан в следующем пункте.[5]

Однако при другом варианте постановки (более предпочтительном) можно задавать зависимость $L(\theta)$ с помощью аффинных равенств и выпуклых неравенств:

$$\sum_{k=1}^{m} H_{W_k}(L) \to \min_{\substack{A\theta + BL = c\\ g(\theta,\ L)\leqslant 0\\ L\,\in\,S_n(1);\ \theta\,\in\,\Theta}}.$$

Многие параметрические зависимости представимы в таком виде [21]. В частности, отметим полиэдральные представления Фурье–Моцкина [21], возникающие, например, в робастной оптимизации:

$$\sum_{k=1}^{m} H_{W_k}(L) \to \min_{L\in\left\{L\in S_n(1):\,\exists\,\theta:\,A\theta+BL\leqslant c\right\}}.$$

Можно переписать задачи (7), (9) и на эти случаи, причем сделать это корректно в том смысле, что правомочность подхода полностью сохранится. При этом принципиально ничего из сказанного выше не поменяется. Подробнее об этом планируется написать в отдельной работе.

В действительности, в приложениях наиболее интересен случай, когда ищется барицентр именно расстояний Вассерштейна,[6] а не энтропийно-регуляризованных расстояний [8] – [11]. Другими словами, интересно изучать предельное поведение $\gamma \to 0+$ (см. п. 3.1 [9], утверждение 1 [10], п. 3 и конец п. 4 [11]). К сожалению, методы из пп. 2, 3 оказываются весьма чувствительными к этому предельному переходу. Для метода из этого раздела константа Липшица градиента в задаче (7) будет расти как γ^{-1}, соответственно, число итераций будет увеличиваться (при использовании быстрых (ускоренных) методов) как $\gamma^{-1/2}$. Еще более плохое поведение (см. [6]) можно ожидать от метода балансировки, использующегося в подходе из п. 3. Планируется в отдельной публикации исследовать вопрос о том, как следует действовать при малых $\gamma > 0$. В частности, в вырожденном случае $\gamma = 0$. По-видимому, в этом случае поможет философия искусственного сглаживания [7] [15], в которой искусственно введенная энтропийная регуляризация уже задается с четко заданным

[4] Впрочем, есть результаты (см. формулу (8) п. 3 § 2 главы 8 [20]) о локальной сходимости обычного градиентного спуска для задачи (9) при некоторых дополнительных предположениях.

[5] При этом правомочность подхода п. 3 для постановки задачи (8) имеет место при дополнительном предположении, что метод стартует из выпуклой окрестности точки минимума с небольшим запасом, допускающим возможность по ходу итерационного процесса оказаться дальше от решения, чем в начальный момент.

[6] Численные методы поиска «честного» барицентра Вассерштейна вероятностных мер в основном строятся на том, что когда меры заданы на прямой, существуют эффективные способы решения задачи поиска барицентра [7]. Далее проектируют (считают преобразования Радона) меры на случайные прямые и решают одномерные задачи. По их решениям восстанавливают решение исходной задачи [22, 23]. В отличие от других подходов, здесь существенно используется структура матрицы затрат $c_{ij} = l^2_{ij}$ (в дискретном случае). Интересно было бы исследовать вопрос о применимости этого подхода к постановкам задач о разладках, в которых требуется много раз пересчитывать барицентр (см. выше). Также интересно было бы сравнить описанные подходы с остальными. Этому планируется посвятить отдельную публикацию.

[7] Это сглаживание правильно называть двойственным сглаживанием, поскольку для того чтобы добиться гладкости в прямой негладкой задаче, которая имеет лежандровское (седловое) представление [15, 21], в это представление, которое также можно понимать как двойственное, вводят аддитивным образом с небольшим коэффициентом сильно выпуклый (вогнутый) функционал. Этот функционал и обеспечивает

коэффициентом регуляризации $\gamma > 0$, зависящим от итоговой точности по функции, с которой требуется решить задачу. Другой способ – использовать менее чувствительные (чем метод балансировки) способы решения двойственной задачи, например, при небольших значениях n ожидается, что лучше сработает r-алгоритм Н. З. Шора и некоторые его обобщения [24, 25]. В данной работе мы фиксируем $\gamma > 0$ и далее уже не будем возвращаться к подобного рода вопросам.

В заключение отметим, что поиск барицентра Вассерштейна в случае $m = 1$ может быть осуществлен явно: $L = W$. Обоснование этого частного результата представляется довольно полезным для понимания основной конструкции этого раздела.

3. Универсальный метод с неточным оракулом

Из п. 2 следует, что внутренняя задача максимизации по (λ, μ) может быть явно решена по μ при фиксированном λ, и наоборот (это верно для задач (1) и (2) и приводит к одним и тем же формулам). Собственно, таким образом, получается метод балансировки расчета матрицы корреспонденций по энтропийной модели, см., например, [6] (тесно связанный с методом Синхорна [9, 11]), как метод простой итерации для явно выписываемых условий экстремума (принципа Ферма): $\lambda = \Lambda(\mu)$, $\mu = \mathrm{M}(\lambda)$. Метод балансировки имеет вид ($[\lambda]_0 = [\mu]_0 = 0$):

$$[\lambda_i]_{k+1} = -\ln\Big[\frac{1}{L_i}\sum_{j=1}^{n}\exp\big(-c_{ij}(y) - 1 + [\mu_j]_k\big)\Big],$$

$$[\mu_j]_{k+1} = -\ln\Big[\frac{1}{W_j}\sum_{i=1}^{n}\exp\big(-c_{ij}(y) - 1 + [\lambda_i]_k\big)\Big]$$

или

$$[\mu_j]_{k+1} = -\ln\Big[\frac{1}{W_j}\sum_{i=1}^{n}\exp\big(-c_{ij} - 1 + [\lambda_i]_{k+1}\big)\Big].$$

В этих формулах «-1» в экспоненте для метода (2) (в отличие от метода (1)) можно не писать, поскольку двойственные множители определяются неоднозначным образом с большим произволом для задачи (2) (см. выше), достаточным для справедливости этого замечания.

Оператор $(\lambda, \mu) \to \big(\Lambda(\mu), \mathrm{M}(\lambda)\big)$ является сжимающим в метрике Биркгофа–Гильберта ρ [26]. Это означает, что после $N \sim \ln(\sigma^{-1})$ итераций метода балансировки можно получить такие (λ_N, μ_N), что $\big\{\big(\lambda_*(y), \mu_*(y)\big)\big\}$ – двумерное аффинное множество решений (см. п. 1):

$$\rho\Big(\big(\lambda_N, \mu_N\big); \big\{\big(\lambda_*(y), \mu_*(y)\big)\big\}\Big) \leqslant \sigma. \tag{10}$$

Причем на практике наблюдается очень быстрая сходимость, т.е. коэффициент пропорциональности не большой [6]. Таким образом, мы можем приближенно решить внутреннюю задачу.[8]

Далее предлагается воспользоваться прямодвойственным (эта важно, поскольку нужно восстанавливать двойственные переменные) универсальным методом [27] для решения

гладкость (а еще точнее липшицевость градиента) в прямой задаче. В нашем случае мы исходим из задачи о перемещении масс (Монж–Канторович), являющейся задачей ЛП. Однако для удобства вычисления расстояний Вассерштейна мы перешли к двойственной задаче. Мы хотим сделать гладкой двойственную задачу, потому что именно с ней в дальнейшем и идет работа. Для этого двойственное сглаживание (в нашем случае энтропийное) применяется к двойственной задаче для двойственной задачи к транспортной задаче, т.е. применяется просто к транспортной задаче.

[8]Заметим, что в пп. 3.1, 3.2 работы [9] предлагается за счет раздутия прямого пространства с помощью обобщения описанного метода балансировки Брэгмана (метода проекций Брэгмана) решать задачу поиска барицентра напрямую, т.е. отпадает необходимость в решении внешней задачи. Плата за это достаточно большая – увеличение размера прямого пространства в m раз, но метод при этом хорошо параллелится.

внешней задачи оптимизации по y (имеется видеопрезентация с описанием этого метода [28]). К сожалению, в формулировке (1) (в отличие от формулировки (2)) кроме того, что внешняя задача гладкая (при условии гладкости $c_{ij}(y)$ [29, 30]), больше ничего о ней сказать нельзя (константа Липшица градиента не ограничена). Также не понятна гладкость задачи (6). Поэтому и по ряду других причин, о которых будет сказано далее, было отдано предпочтение универсальному методу, оптимально адаптивно настраивающемуся на гладкость функционала $f(y)$ на текущем участке пребывания итерационного процесса.[9] Однако нам потребуется использовать этот метод в варианте с неточным оракулом, выдающим градиент [31]. Напомним (см. п. 1), что мы решаем задачу 2, представимую в виде (здесь в max представлении $x = x$,

$$\bar{Q} = \Big\{x_{ij} \geqslant 0, i, j = 1, ..., n : \sum_{j=1}^{n} x_{ij} = L_i, \sum_{i=1}^{n} x_{ij} = W_j\Big\},$$

а в min представлении $x = (\lambda, \mu)$, $\bar{Q} = \mathbb{R}^{2n}$):

$$f(y) = \max_{x \in \bar{Q}} \Psi(x, y) = \min_{x \in \bar{Q}} \Phi(x, y) \to \min y \in Q.$$

Далее везде будем считать, что $y \in Q$.

Определение 1 (см. главу 4 [32])**.** *(δ, L)-оракул выдает (на запрос, в котором указывается только одна точка y) такие $\big(F(y), G(y)\big)$, что и для любых $y, y' \in Q$*

$$0 \leqslant f(y') - F(y) - \big\langle G(y), y' - y\big\rangle \leqslant \frac{L}{2}\|y' - y\|^2 + \delta.$$

Из определения 1 сразу следует, что для любого $x \in Q$

$$F(x) \leqslant f(x) \leqslant F(x) + \delta$$

и для любых $x, y \in Q$

$$f(y) \geqslant f(y) - \langle G(x), y - x\rangle - \delta.$$

Из последнего свойства получаем, что определение (δ, L)-оракула можно понимать как обобщение на гладкие задачи классического понятия негладкой выпуклой оптимизации: δ-субградиента (см. п. 5 § 1 главы 5 [20]). В приводимом далее утверждении в первой его части следует сохранить обозначения для задачи (2), (3) и следует обозначить $x = \lambda$, $y = L$ для задачи (5), (6); а во второй части утверждения следует обозначить $x = (\lambda, \mu)$, $y = y$ для задачи (2), (3). Таким образом, на задачу (2), (3) можно посмотреть с двух разных ракурсов, однако второй ракурс менее привлекателен ввиду необходимости рассмотрения ограниченных множеств $\bar{Q}$, что в интересующих нас приложениях место не имеет.

[9] Бытует мнение, с которым столкнулись и авторы данной статьи, что любой универсальный метод должен чем-то платить за свою универсальность, и в этой связи возникает много вопросов, в частности: насколько дорога эта плата? В принципе, в статье [27] довольно подробно проясняется этот момент. Тем не менее мы повторим здесь соображения из [27]. Действительно, плата за универсальность есть. Универсальный метод из работы [27] может сделать где-то в 4 раза больше обращений к оракулу для задачи с более менее одинаковой константой Липшица градиента во всей области (где довелось пройти итерационному процессу), по сравнению с обычным быстрым градиентным методом [15]. Тем не менее замечательная особенность универсального метода не только в том, что он настраивается на гладкость задачи и применим к любым задачам, но и в том, что (в отличие от подавляющего большинства методов) этот метод локально настраивается на гладкость функционала. И для сильно неоднородных функционалов типично, что универсальный метод делает заметно меньше итераций, чем, скажем, быстрый градиентный метод (плата за универсальность уже учтена в отмеченном выше потенциально возможном увеличении числа итераций в 4 раза в худшем случае). Примеры, поясняющие сказанное, имеются в работе [27].

Утверждение 1. Если $\psi(y) = \max\limits_{x\in\bar{Q}} \Psi(x,y)$, где $\Psi(x,y)$ – выпуклая по y и вогнутая по x функция, и найден такой $\tilde{x} \in \bar{Q}$, что

$$\psi(y) - \Psi(\tilde{x},y) \leqslant \delta,$$

то субградиент $\partial_y\Psi(\tilde{x},y)$ есть δ-субградиент функции $\psi(y)$ в точке y. Если $\varphi(y) = \min\limits_{x\in\bar{Q}} \Phi(x,y)$, где $\Phi(x,y)$ – выпуклая по совокупности переменных функция, и найден такой $\tilde{x} \in \bar{Q}$, что

$$\max_{z\in\bar{Q}} \langle \Phi_x(\tilde{x},y), \tilde{x} - z\rangle \leqslant \delta,$$

то

$$\Phi(\tilde{x},y) - \varphi(y) \leqslant \delta$$

и субградиент $\Phi_y(\tilde{x},y) = \partial_y\Phi(\tilde{x},y)$ есть δ-субградиент функции $\varphi(y)$ в точке y.

Доказательство. Ограничимся доказательством только второй части этого утверждения. Доказательство первой части см. на с. 124 (лемма 13) книги [20]. Из выпуклости $\Phi(x,y)$ по совокупности переменных имеем

$$\Phi(x',y') \geqslant \Phi(x,y) + \big\langle \Phi_x(x,y), x'-x\big\rangle + \big\langle \Phi_y(x,y), y'-y\big\rangle. \tag{11}$$

Определим зависимость $x(y)$ из соотношения

$$\varphi(y) = \min_{x\in\bar{Q}} \Phi(x,y) = \Phi\big(x(y),y\big).$$

Заметим, что $\Phi(\tilde{x},y) \geqslant \varphi(y)$. Положим в (11) $x' = x(y')$, $x = \tilde{x}$. Тогда

$$\begin{aligned}\varphi(y') = \Phi(x',y') &\geqslant \Phi(\tilde{x},y) + \big\langle \Phi_x(\tilde{x},y), x'-\tilde{x}\big\rangle + \big\langle \Phi_y(\tilde{x},y), y'-y\big\rangle \geqslant\\ &\geqslant \varphi(y) + \big\langle \Phi_x(\tilde{x},y), x(y')-\tilde{x}\big\rangle + \big\langle \Phi_y(\tilde{x},y), y'-y\big\rangle \geqslant\\ &\geqslant \varphi(y) + \big\langle \Phi_y(\tilde{x},y), y'-y\big\rangle - \delta.\end{aligned}$$

В последней формуле мы использовали, что

$$\big\langle \Phi_x(\tilde{x},y), \tilde{x} - x(y')\big\rangle \leqslant \delta.$$

В свою очередь из выпуклости $\Phi(x,y)$ по x (для всех допустимых y) имеем

$$\Phi(\tilde{x},y) - \Phi\big(x(y'),y\big) \leqslant \big\langle \Phi_x(\tilde{x},y), \tilde{x} - x(y')\big\rangle.$$

Беря в этой формуле $y' = y$ и воспользовавшись определением $x(y)$, получаем, что

$$\Phi(\tilde{x},y) - \varphi(y) \leqslant \big\langle \Phi_x(\tilde{x},y), \tilde{x} - x(y)\big\rangle.$$

Однако не хочется довольствоваться возможностью находить только δ-субградиент (из утверждения 1 эта возможность очевидна), поскольку в определенных ситуациях явно можно рассчитывать на некоторую гладкость итоговой (внешней) задачи (2). Понятие (δ, L)-оракула в некотором смысле налагает наиболее слабые условия на возможные неточности в вычислении функции и градиента, при которых можно рассчитывать, что скорость сходимости метода, учитывающего гладкость (липшицевость градиента функционала) задачи, сильно не пострадает (см. теорему 1 ниже).

На первый взгляд может показаться, что применимость описанной концепции (δ, L)-оракула к задаче (1) следует из следующего результата (см. п. 4.2.2 [32]).

Утверждение 2. Пусть подзадача энтропийно-линейного программирования (ЭЛП) в (2) решена (по функции) с точностью δ, т.е. найден такой $\tilde{x}(c)$, удовлетворяющей балансовым ограничениям, что

$$\sum_{i,j=1}^{n} \tilde{x}_{ij}(c) \ln \tilde{x}_{ij}(c) + \sum_{i,j=1}^{n} c_{ij}\tilde{x}_{ij}(c) -$$

$$- \min_{\substack{\sum\limits_{j=1}^{n} x_{ij} = L_i,\ \sum\limits_{i=1}^{n} x_{ij} = W_j \\ x_{ij} \geqslant 0,\, i,\, j = 1, \ldots, n}} \left\{ \sum_{i,j=1}^{n} x_{ij} \ln x_{ij} + \sum_{i,j=1}^{n} c_{ij} x_{ij} \right\} \leqslant \delta.$$

Тогда для функции

$$\bar{f}(c) = - \min_{\substack{\sum\limits_{j=1}^{n} x_{ij} = L_i,\ \sum\limits_{i=1}^{n} x_{ij} = W_j \\ x_{ij} \geqslant 0,\, i,\, j = 1, \ldots, n}} \left\{ \sum_{i,j=1}^{n} x_{ij} \ln x_{ij} + \sum_{i,j=1}^{n} c_{ij} x_{ij} \right\}$$

набор

$$-\left(\sum_{i,j=1}^{n} \tilde{x}_{ij}(c) \ln \tilde{x}_{ij}(c) + \sum_{i,j=1}^{n} c_{ij}\tilde{x}_{ij}(c), \left\{ \tilde{x}_{ij}(c) \right\}_{i,j=1}^{n,n} \right)$$

является $(\delta, \max\limits_{i,j=1,\ldots,n} c_{ij})$-оракулом.

К сожалению, большинство методов (в том числе метод балансировки) не удовлетворяют одному пункту утверждения 2, а именно, они выдают вектор $\tilde{x}$, который лишь приближенно удовлетворяет балансовым ограничениям (в утверждении требование точного удовлетворения балансовых ограничений является существенным и не может быть как-то равнозначно релаксировано). Связанно это с тем, что для задачи ЭЛП, когда ограничений намного меньше числа прямых переменных, обычно решается двойственная задача, по которой восстанавливается решение прямой задачи [6, 33]. Как следствие, приобретается невязка и в ограничениях. Собственно, в представлении градиента функционала по формуле (4) имеются два способа. Первый через двойственные множители (λ, μ), второй через решение прямой задачи x. Функционал прямой задачи сильно выпуклый по x, поскольку энтропия 1-сильно выпуклая функция в 1-норме [15]. Поэтому сходимость в решении прямой задачи по функции обеспечивает сходимость и по аргументу, что и означает возможность определения с хорошей точностью градиента по формуле (4) через x. Другая ситуация возникает, если смотреть на двойственную задачу к задаче ЭЛП (в приводимом далее утверждении следует обозначить $x = (\lambda, \mu)$, $y = y$ для задачи (2), (3)).

Утверждение 3. Пусть $\varphi(y) = \min\limits_{x \in \bar{Q}} \Phi(x, y)$, где $\Phi(x, y)$ – такая достаточно гладкая, выпуклая по совокупности переменных функция, что[10]

$$\left\| \nabla\Phi(x', y') - \nabla\Phi(x, y) \right\|_2 \leqslant L \|(x', y') - (x, y)\|_2.$$

Пусть для произвольного $y \in Q$ (считаем, что множество Q содержит внутри себя шар радиуса более $\sqrt{2\delta/L}$) можно найти такой $\tilde{x} = \tilde{x}(y) \in \bar{Q}$, что

$$\max_{z \in \bar{Q}} \left\langle \Phi_x(\tilde{x}, y), \tilde{x} - z \right\rangle \leqslant \delta.$$

Тогда

$$\Phi(\tilde{x}, y) - \varphi(y) \leqslant \delta,$$

[10]Это утверждение имеет достаточно простую геометрическую интерпретацию. Проекция надграфика выпуклой функции будет выпуклым множеством, то есть, в свою очередь, надграфиком некоторой выпуклой функции. Кривизна границы у полученного при проектировании множества будет не больше, чем была у исходного множества. Это следует из того, что проектирование — сжимающий оператор.

$$\big\|\nabla\varphi(y') - \nabla\varphi(y)\big\|_2 \leqslant L\|y' - y\|_2,$$

и $\Big(\Phi(\tilde{x}, y) - 2\delta, \Phi_y(\tilde{x}, y)\Big)$ будет $(6\delta, 2L)$-оракулом для $\varphi(y)$ на выпуклом множестве, полученном из множества Q отступанием от границы ∂Q во внутрь Q на расстояние $\sqrt{2\delta/L}$ (по условию это множество не пусто).

Доказательство. По условию задачи имеем при всех допустимых значениях аргументов Φ:

$$\lambda_{\max}\left(\left\| \begin{array}{cc} \Phi_{xx} & \Phi_{xy} \\ \Phi_{yx} & \Phi_{yy} \end{array} \right\|\right) = \sup_{\|h\|_2 \leqslant 1} \left\langle h, \left\| \begin{array}{cc} \Phi_{xx} & \Phi_{xy} \\ \Phi_{yx} & \Phi_{yy} \end{array} \right\| h \right\rangle \leqslant L. \tag{12}$$

Заметим, что также по условию при всех допустимых значениях аргументов Φ:

$$\left\| \begin{array}{cc} \Phi_{xx} & \Phi_{xy} \\ \Phi_{yx} & \Phi_{yy} \end{array} \right\| \succ 0, \quad \Phi_{xx} \succ 0, \quad \Phi_{yy} \succ 0, \quad \Phi_{yx} = \Phi_{xy}^T, \quad \Phi_{xx} = \Phi_{xx}^T, \quad \Phi_{yy} = \Phi_{yy}^T.$$

Для упрощения последующих рассуждений (в частности, чтобы не работать с псевдообратными матрицами) будем, считать, что матрица $\Phi_{xx} \succ 0$ положительно определена (исходя из условий, гарантировать можно лишь неотрицательную определенность). Также будем считать (в интересующих нас приложениях к задачам (2), (6) это имеет место), что зависимость $x(y)$, определяемая из соотношения

$$\varphi(y) = \min_{x \in \bar{Q}} \Phi(x, y) = \Phi\big(x(y), y\big)$$

однозначным образом, и удовлетворяет соотношению

$$\Phi_x\big(x(y), y\big) \underset{y}{\equiv} 0,$$

из которого имеем

$$\Phi_{xx}\big(x(y), y\big)\left\| \frac{\partial x(y)}{\partial y} \right\| + \Phi_{xy}\big(x(y), y\big) \underset{y}{\equiv} 0,$$

т.е.

$$\big\|\partial x/\partial y\big\| = \big\|\partial x_i/\partial y_j\big\| = -\Phi_{xx}^{-1}\Phi_{xy}.$$

Поскольку $\varphi(y) = \Phi\big(x(y), y\big)$, то

$$\begin{aligned} \varphi_{yy} = {} & \big\|\partial x/\partial y\big\|^T \Phi_{xx}\big\|\partial x/\partial y\big\| + \big\|\partial x/\partial y\big\|^T \Phi_{xy} + \Phi_{yx}\big\|\partial x/\partial y\big\| + \Phi_{yy} = \\ & = \Phi_{yy} - \Phi_{yx}\Phi_{xx}^{-1}\Phi_{xy}. \end{aligned}$$

С учетом этой формулы и из формулы дополнения по Шуру [34], получаем

$$\left\| \begin{array}{cc} \Phi_{xx} & \Phi_{xy} \\ \Phi_{yx} & \Phi_{yy} \end{array} \right\| = \left\| \begin{array}{cc} E_x & 0 \\ \Phi_{yx}\Phi_{xx}^{-1} & E_y \end{array} \right\| \left\| \begin{array}{cc} \Phi_{xx} & 0 \\ 0 & \Phi_{yy} \end{array} \right\| \left\| \begin{array}{cc} E_x & \Phi_{xx}^{-1}\Phi_{xy} \\ 0 & E_y \end{array} \right\|,$$

где E_x, E_y – единичные матрицы соответствующих размеров. Поскольку

$$\left\| \begin{array}{cc} E_x & 0 \\ \Phi_{yx}\Phi_{xx}^{-1} & E_y \end{array} \right\| = \left\| \begin{array}{cc} E_x & \Phi_{xx}^{-1}\Phi_{xy} \\ 0 & E_y \end{array} \right\|^T$$

и эти матрицы полного ранга, то из (12) имеем, что

$$\begin{aligned} & \sup_{\|h\|_2 \leqslant 1} \big\langle h, \varphi_{yy}\, h \big\rangle = \lambda_{\max}(\varphi_{yy}) \leqslant \\ & \leqslant \max\big\{\lambda_{\max}(\Phi_{xx}), \lambda_{\max}(\varphi_{yy})\big\} = \lambda_{\max}\left(\left\| \begin{array}{cc} \Phi_{xx} & \Phi_{xy} \\ \Phi_{yx} & \Phi_{yy} \end{array} \right\|\right) \leqslant L. \end{aligned}$$

Таким образом, установлено, что

$$\varphi(y) \leqslant \varphi(x) + \langle \nabla\varphi(x), y - x\rangle + \frac{L}{2}\|y - x\|_2^2.$$

Согласно утверждению 1

$$\varphi(y) \geqslant \varphi(x) + \langle \Phi_y(\tilde{x}, y), y - x\rangle - \delta.$$

Далее проведем рассуждения аналогично рассуждениям на с. 107 (и немного отлично от с. 115) диссертации [32]. Вычитая из первого неравенства второе, получим

$$\langle \Phi_y(\tilde{x}, y) - \nabla\varphi(x), y - x\rangle \leqslant \frac{L}{2}\|y - x\|_2^2 + \delta.$$

Положим $t > 0$:

$$y - x = \frac{\Phi_y(\tilde{x}, y) - \nabla\varphi(x)}{\|\Phi_y(\tilde{x}, y) - \nabla\varphi(x)\|_2}t,$$

получим

$$\|\Phi_y(\tilde{x}, y) - \nabla\varphi(x)\|_2 \leqslant \frac{Lt}{2} + \frac{\delta}{t}.$$

Минимизируя правую часть неравенства по $t > 0$, при $t = \sqrt{2\delta/L}$ получаем

$$\|\Phi_y(\tilde{x}, y) - \nabla\varphi(x)\|_2 \leqslant \sqrt{2\delta L}.$$

Отсюда и из утверждения 1 находим

$$\begin{aligned}
\varphi(y) &\leqslant \varphi(x) + \langle \nabla\varphi(x), y - x\rangle + \frac{L}{2}\|y - x\|_2^2 \leqslant \\
&\leqslant \varphi(x) + \langle \Phi_y(\tilde{x}, y), y - x\rangle + \sqrt{2\delta L}\|y - x\|_2 + \frac{L}{2}\|y - x\|_2^2 \leqslant \\
&\leqslant \Phi(\tilde{x}, y) - 2\delta + \langle \Phi_y(\tilde{x}, y), y - x\rangle + \sqrt{2\delta L}\|y - x\|_2 + \frac{L}{2}\|y - x\|_2^2 + 2\delta \leqslant \\
&\leqslant \Phi(\tilde{x}, y) - 2\delta + \langle \Phi_y(\tilde{x}, y), y - x\rangle + L\|y - x\|_2^2 + 6\delta.
\end{aligned}$$

С учетом того, что (см. утверждение 1)

$$\begin{aligned}
\varphi(y) &\geqslant \varphi(x) + \langle \Phi_y(\tilde{x}, y), y - x\rangle - \delta \\
&\geqslant \Phi_y(\tilde{x}, y) - 2\delta + \langle \Phi_y(\tilde{x}, y), y - x\rangle,
\end{aligned}$$

из определения 1 получаем доказываемое утверждение.

Это утверждение позволяет установить гладкость задачи (2), (3) (но не (5), (6)). Таким образом, для (5), (6) необходимость использования универсального метода для внешней задачи является отражением надежды сходиться быстрее, чем в негладком случае, в то время как для (2), (3) использование универсального метода для внешней задачи является скорее отражением желания настраиваться на правильную константу Липшица градиента. Можно, конечно, пытаться использовать приведенные выше формулы, однако из способа рассуждений (см., например, доказательство утверждения 3) видно, что полученная таким образом константа Липшица может оказаться завышенной.

К сожалению, практическое применение утверждения 3 натыкается на следующие сложности:

1) необходимости отступать от границы множества Q во внутрь на $\sqrt{2\delta/L}$,

2) необходимости рассмотрения ситуации (см. доказательство утверждения 3):

$$\|\partial x_i/\partial y_j\| = -\Phi_{xx}^{-1}\Phi_{xy},$$

3) необходимости предположения о компактности множества $\bar{Q}$, иначе невозможно будет добиться выполнения условия

$$\max_{z\in\bar{Q}}\langle \Phi_x(\tilde{x},y),\tilde{x}-z\rangle \leqslant \delta.$$

Первая сложность на практике разрешима за счет возможности доопределения функционала задачи с сохранением всех свойств на $\sqrt{2\delta/L}$-окрестность множества Q (заметим, что доопределение часто не требуется, поскольку функционал и так задан «с запасом»). Например, для рассматриваемых нами транспортных приложений с $Q=\{y:\ \ y\geqslant\bar{y}\}$ это возможно [2] – [5]. Сложность 2 часто вообще не возникает (разве что оговорка о существовании Φ_{xx}^{-1}, впрочем, приведенные выше рассуждения можно провести, сохранив все результаты в идентичном виде, так, что эта оговорка будет не нужна), поскольку $\bar{Q}$ совпадает со всем (двойственным) пространством. А вот сложность 3, действительно, портит дело. К сожалению, простых теоретически обоснованных способов борьбы с этой сложностью мы пока не знаем. Тем не менее полезно заметить, что в действительности нужно гарантировать выполнение (см. доказательство утверждения 1)

$$\big\langle \Phi_x\big(\tilde{x}(y),y\big),\tilde{x}(y)-x(y')\big\rangle \leqslant \delta,$$

где точки y и y' близки, поскольку возникают на соседних итерациях внешнего метода. С учетом ожидаемой «близости» $\tilde{x}=\tilde{x}(y)$ и $x(y)$, мы можем заменить в этом критерии настоящее множество $\bar{Q}$, которое, как правило, совпадает со всем пространством, на шар конечного радиуса. Более детальные исследования (для задачи (2), (3)) и практические эксперименты показывают, что для выполнения приведенного выше условия достаточно обеспечить для внутреннего итерационного процесса $\{x_k\}\to x(y)$ условия

$$\|\Phi_x(x_k,y)\|_2\ \|x_k\|_2\leqslant\delta/2,\quad \|\Phi_x(x_k,y)\|_2\leqslant\delta.$$

Соответствующее $x_k=(\lambda_k,\mu_k)$ порождает нужное $\tilde{x}(y)=x_k$. С учетом специфики рассматриваемой нами задачи (2), (3), имеем следующий критерий (возвращаемся к обозначениям (2), (3)):

$$\big\|Ax(\lambda_k,\mu_k)-b\big\|_2\ \big\|(\lambda_k,\mu_k)\big\|_2\leqslant\delta/2,\quad \big\|Ax(\lambda_k,\mu_k)-b\big\|_2\leqslant\delta,$$

где $x(\lambda_k,\mu_k)$ определяется в формуле (4), а введённая линейная система балансовых уравнений $Ax=b$ есть общая запись аффинных (транспортных) ограничений:

$$\sum_{j=1}^{n}x_{ij}=L_i,\ \sum_{i=1}^{n}x_{ij}=W_j,\quad i,j=1,...,n.$$

В связи со сказанным выше заметим, что (это следует из оценки (10)) метод балансировки обеспечивает сходимость и по аргументу, что для других методов (без введения регуляризации) решения двойственной задачи, вообще говоря, нельзя гарантировать. Это свойство наряду с линейной скоростью сходимости метода (со скоростью геометрической прогрессии) позволяет надеяться, что выбранный критерий является достаточно точным (точнее, не слишком грубым).

Принципиально важно для гладкого случая ($c_{ij}(y)$ – функции с липшицевым градиентом), как это будет следовать из дальнейших оценок (см. теорему 1), не просто уметь решать двойственную задачу, т.е. находить (λ,μ) так, чтобы была сходимость по аргументу, а делать это так, чтобы сложность решения задачи зависела от точности ее решения логарифмическим образом. Выше мы отмечали, что это имеет место для метода балансировки. Также это имеет место и для быстрых методов, примененных к регуляризованной двойственной задачи. При фиксации параметра регуляризации, исходя из итоговой желаемой точности, быстрые градиентные методы (для сильно выпуклых функций) решают

регуляризованную двойственную задачу так, что зависимость сложности от точности ее решения – логарифмическая.

Хочется, чтобы при решении внешней задачи в (2), т.е. задачи

$$\min_{y\in Q} f(y),$$

можно было не задумываться ни о какой гладкости. Если она есть, то метод бы это хорошо учитывал, не требуя знания констант Липшица градиента (это намного более существенно для возможности применять описанный подход к поиску барицентра Вассерштейна вероятностных мер, см. п. 2), если ее нет, то метод также работал бы оптимальным (для негладкого случая) образом. Именно таким свойством и обладает универсальный метод [27], работающий и в концепции неточного оракула [31] (см. определение 1).

Заметим [27], что можно погрузить задачу с гёльдеровым градиентом ($\nu \in [0,1]$)

$$\|\nabla f(y') - \nabla f(y)\|_* \leqslant L_\nu \|y' - y\|^\nu$$

(в том числе и негладкую задачу с ограниченной нормой разности субградиентов при $\nu = 0$) в класс гладких задач с оракулом, характеризующимся точностью δ и

$$L = L_\nu \left[\frac{L_\nu(1-\nu)}{2\delta(1+\nu)}\right]^{\frac{1-\nu}{1+\nu}}.$$

Это позволяет даже в случае, когда можно рассчитывать только на δ-субградиент[11] (с ограниченной нормой субградиента (разности субградиентов), причем, какой именно константой ограниченной, методу знать не обязательно), все равно работать в концепции (δ, L)-оракула.

Итак, у нас есть внешняя задача (2)

$$\min_{y\in Q} f(y),$$

для которой обращение к (δ, L)-оракулу за значением функции и градиента стоит $\sim \ln(\delta^{-1})$. Насколько быстро мы можем решить такую задачу, т.е. при каком $N(\varepsilon)$ можно гарантировать, что

$$f(y_{N(\varepsilon)}) - \min_{y\in Q} f(y) \leqslant \varepsilon?$$

Ответ можно получить из следующего результата.

Теорема 1 (см. [1, 21, 27, 31])**.** *Существует однопараметрическое семейство универсальных градиентных методов (параметр $p \in [0,1]$), не получающих на вход, кроме p, больше никаких параметров (в частности, не использующих значения L_ν и R – «расстояние» от точки старта до решения, – априорно не известное), которое приводит к следующей оценке на требуемое число итераций:*

$$N_p(\varepsilon) = \mathrm{O}\left(\inf_{\nu\in[0,1]}\left[\frac{L_\nu R^{1+\nu}}{\varepsilon}\right]^{\frac{2}{1+2p\nu+\nu}}\right),$$

если $\delta \leq \mathrm{O}\big(\varepsilon N_p(\varepsilon)^{-p}\big)$.

[11] На δ-субградиент всегда можно рассчитывать согласно утверждению 1. Причем, как уже отмечалось раньше, для получения δ-субградиента не нужна сходимость по аргументу для вспомогательной задачи.

Из теоремы 1 можно заключить, что если мы рассчитываем на некоторую гладкость $f(y)$, то стоит выбирать значение параметра $p = 1$, при этом общие трудозатраты машинного времени будут

$$\mathrm{O}\Big(N_1(\varepsilon)\big(T\ln(\varepsilon^{-1}) + \tilde{T}\big)\Big), \tag{13}$$

где $\tilde{T}$ – время вычисления (суб-)градиента функционала (в основном это вычисления $\{\nabla c_{ij}(y)\}_{i,j=1}^{n,n}$ [29, 30]), T – время решения вспомогательной задачи методом балансировки с относительной точностью 1%. Численные эксперименты показывают, что на одном современном ноутбуке при $n \sim 10^2$ время $T \approx 1$ с. [31], что сопоставимо с временем $\tilde{T}$ для таких n [29].

Выгода от описанной выше конструкции по сравнению с обычным способом решения исходной задачи минимизации (2), (3) сразу по совокупности всех переменных (см., например, [5]) заключается в гарантированном не увеличении константы Липшица градиента в оценке необходимого числа итераций (см. утверждение 3) и ожидаемое уменьшение в этой же оценке расстояния от точки старта до (неизвестного априори) решения. Выгода здесь вполне может достигать одного порядка и более. При этом можно ожидать лишь незначительного увеличения стоимости одной итерации. Причем стоит иметь в виду, что при оптимизации сразу по всем переменным требуется рассчитывать градиент функционала по большему числу переменных, чем в описанном выше подходе, что также играет нам на пользу. В конечном итоге, сокращение числа итераций заметно превалирует над небольшим увеличением стоимости одной итерации.

Что касается задач (5), (6), то описанный выше подход представляется естественным и не имеющим альтернатив в рассматриваемом классе. Альтернативные методы, с которыми можно сравнивать, мы упоминали по ходу статьи, но все они были предложены из принципиально других подходов. Сравнению (практическому) всех этих методов планируется посвятить отдельную публикацию.

Резюмируем ключевой результат этого раздела (и всей статьи) следующим образом.

Для решения задач типа (2), (6) *или* (8) *предлагается использовать универсальный метод из работы* [27] *(а точнее его модификацию из* [31]*). Если рассчитываем на гладкость*[12] $f(y)$*, то полагаем в методе* $p = 1$*. Если на гладкость рассчитывать не приходится*[13]*, то полагаем* $p = 0$*. В обоих случаях, кроме априорной подсказки относительно параметра* p*, методу больше ничего от нас знать не надо!*

Авторы выражают глубокую признательность Ю. Е. Нестерову за серию продуктивных обсуждений.

Исследование А. В. Гасникова, П. Е. Двуреченского, В. Г. Спокойного и А. Л. Суворикoвой в части 2 выполнено в ИППИ РАН за счет гранта Российского научного фонда (проект № 14-50-00150); исследование П. Е. Двуреченского в части 3 выполнено при поддержке гранта РФФИ 15-31-20571-мол_а_вед; исследование А. В. Гасникова в части 3 выполнено при поддержке гранта РФФИ 15-31-70001 мол_а_мос.

Литература

1. *Гасников А.В., Двуреченский П.Е., Камзолов Д.И., Нестеров Ю.Е., Спокойный В.Г., Стецюк П.И., Суворикова А.Л., Чернов А.В.* Поиск равновесий в многостадийных транспортных моделях // Труды МФТИ. 2015. Т. 7, № 4. С. 143–155.

[12] В этом случае как раз существенна логарифмическая сложность приближенного вычисления градиента и значения функции $f(y)$ от точности, обеспеченная методом балансировки.

[13] В этом случае точность решения вспомогательной задачи расчета δ-субградиента можно завязать на желаемую точность решения задачи (2) ε (или (6) или (8)) по формуле $\delta = \mathrm{O}(\varepsilon)$ (см. теорему 1 при $\nu = 0$), с константой порядка 1.

2. *Гасников А.В., Дорн Ю.В., Нестеров Ю.Е, Шпирко С.В.* О трехстадийной версии модели стационарной динамики транспортных потоков // Математическое моделирование. 2014. Т. 26:6. С. 34–70. arXiv:1405.7630

3. *Гасников А.В.* Об эффективной вычислимости конкурентных равновесий в транспортно-экономических моделях // Математическое моделирование. 2015. Т. 27, № 12. С. 121–136. arXiv:1410.3123

4. *Бабичева Т.С., Гасников А.В., Лагуновская А.А., Мендель М.А.* Двухстадийная модель равновесного распределения транспортных потоков // Труды МФТИ 2015. Т. 7, № 3. С. 31–41. https://mipt.ru/upload/medialibrary/971/31-41.pdf

5. *Гасников А.В., Гасникова Е.В., Мациевский С.В., Усик И.В.* О связи моделей дискретного выбора с разномасштабными по времени популяционными играми загрузок // Труды МФТИ. 2015. Т. 7, № 4. С. 129–142. arXiv:1511.02390

6. *Гасников А.В., Гасникова Е.В., Нестеров Ю.Е., Чернов А.В.* Об эффективных численных методах решения задач энтропийно-линейного программирования // ЖВМ и МФ. 2016. Т. 56, № 4. С. 523–534. arXiv:1410.7719

7. *Agueh M., Carlier G.* Barycenters in the Wasserstein space // SIAM J. Math. Anal. 2011. V. 43, N 2. P. 904–924.

8. *Cuturi M., Doucet A.* Fast Computation of Wasserstein Barycenters // ICML. 2014. http://www.iip.ist.i.kyoto-u.ac.jp/member/cuturi/

9. *Benamou J.D., Carlier G., Cuturi M., Nenna L., Peyré G.* Iterative Bregman Projections for Regularized Transportation Problems // e-print, 2015. (to appear in SISC) arXiv:1412.5154

10. *Cuturi M., Peyré G., Rolet A.* A Smoothed Dual Formulation for Variational Wasserstein Problems // e-print, 2015. arXiv:1503.02533

11. *Cuturi M.* Sinkhorn Distances: Lightspeed Computation of Optimal Transport // NIPS. 2013.

12. *Немировский А.С., Юдин Д.Б.* Сложность задач и эффективность методов оптимизации. М.: Наука, 1979. http://www2.isye.gatech.edu/simnemirovs/Lect_EMCO.pdf

13. *Boissard E., Le Gouic T., Loubes J.-M.* Distribution's Template Estimate with Wasserstein Metrics // e-print, 2013. (to be published in Bernoulli) arXiv:1111.5927

14. *Bigot J., Klein T.* Consistent estimation of a population barycenter in the Wasserstein space // e-print, 2015. arXiv:1212.2562

15. *Nesterov Y.* Smooth minimization of nonsmooth function // Math. Program. Ser. A. 2005. V. 103, N 1. P. 127–152.

16. *Fercoq O., Richtarik P.* Accelerated, Parallel and Proximal Coordinate Descent // e-print, 2013. arXiv:1312.5799

17. *Qu Z., Richtarik P.* Coordinate Descent with Arbitrary Sampling I: Algorithms and Complexity // e-print, 2014. arXiv:1412.8060

18. *Boyd S., Parikh N., Chu E., Peleato B., Eckstein J.* Distributed optimization and statistical learning via the alternating direction method of multipliers // Foundations and Trends in Machine Learning. 2011. V. 3(1). P. 1–122. http://stanford.edu/boyd/papers.html

19. *Boyd S., Vandenberghe L.* Convex optimization. Cambridge University Press, 2004. http://stanford.edu/boyd/cvxbook/

20. *Поляк Б.Т.* Введение в оптимизацию. М.: Наука, 1983.

21. *Nemirovski A.* Lectures on modern convex optimization analysis, algorithms, and engineering applications. Philadelphia: SIAM, 2013.
http://www2.isye.gatech.edu/simnemirovs/Lect_ModConvOpt.pdf

22. *Rabin J., Peyér G., Delon J. Bernot M.* Wasserstein barycenter and its applications to texture mixing // LNCS. 2011. Proc. SSVM'11. Springer. V. 6667. P. 435–446. https://hal.archives-ouvertes.fr/hal-00476064/document

23. *Bonnel N., Pfister H.* Sliced Wasserstein barycenter of multiple densities // Harvard Technical Report. 2013. TR-02-13.
ftp://ftp.deas.harvard.edu/techreports/tr-02-13.pdf

24. *Стецюк П.И.* Методы эллипсоидов и r-алгоритмы. Кишинев: Эврика, 2014.

25. *Стецюк П.И., Гасников А.В.* NLP-программы и r-алгоритм в задаче энтропийно-линейного программирования // Теория оптимальных решений. Киев: Институт кибернетики им. В.М.Глушкова НАН Украины, 2015. С. 73–78.

26. *Franklin J., Lorenz J.* On the scaling of multidimensional matrices // Linear Algebra and its applications. 1989. V. 114. P. 717–735.

27. *Nesterov Yu.* Universal gradient methods for convex optimization problems // CORE Discussion Paper 2013/63. 2013; Math. Program. Ser. A. 2015. V. 152. P. 381–404. https://www.uclouvain.be/cps/ucl/doc/core/documents/coredp2013_26web.pdf

28. *Nesterov Yu.* http://www.youtube.com/watch?v=Fm9h92pcbvg

29. *Гасников А.В., Двуреченский П.Е., Дорн Ю.В., Максимов Ю.В.* Численные методы поиска равновесного распределения потоков в моделях Бэкмана и стабильной динамики // Математическое моделирование. 2016. Т. 28, № 10. С. 40–64. arXiv:1506.00293

30. *Гасников А.В., Гасникова Е.В., Ершов Е.И., Двуреченский П.Е., Лагуновская А.А.* Поиск стохастических равновесий в моделях равновесного распределения потоков // Труды МФТИ. 2015. Т. 7, № 4. С. 114–128. arXiv:1505.07492

31. *Гасников А.В, Двуреченский П.Е., Камзолов Д.И.* Градиентные и прямые методы с неточным оракулом для задач стохастической оптимизации // Динамика систем и процессы управления. Труды Международной конференции, посвященной 90-летию со дня рождения академика Н.Н. Красовского. Екатеринбург, 15–20 сентября 2014. Издательство: Институт математики и механики УрО РАН им. Н.Н. Красовского (Екатеринбург). 2014. С. 111–117. arXiv:1502.06259

32. *Devolder O.* Exactness, inexactness and stochasticity in first-order methods for large-scale convex optimization. CORE UCL, PhD thesis, March 2013.

33. *Anikin A., Dvurechensky P., Gasnikov A., Golov A., Gornov A., Maximov Yu., Mendel M., Spokoiny V.* Modern efficient numerical approaches to regularized regression problems in application to traffic demands matrix calculation from link loads // Proceedings of International Conference ITAS-2015. Russia, Sochi, September, 2015. arXiv:1508.00858

34. *Zhang F.* The Schur Complement and Its Applications. Springer, 2005.

References

1. *Gasnikov A., Dvurechensky P., Kamzolov D., Nesterov Y., Spokoiny V., Stetsyuk P., Suvorikova A., Chernov A.* Search for the stochastic equilibria in the transport models of equilibrium flow distribution. Proceedings of MIPT. 2015. V. 7, N 4. P. 143–155. (in Russian)

2. *Gasnikov A., Dorn Yu., Nesterov Yu., Shpirko S.* On the three-stage version of stable dynamic model. Matem. Mod. 2014. V. 26:6. P. 34–70. arXiv:1405.7630 (in Russian)

3. *Gasnikov A.* About reduction of searching competetive equillibrium to the minimax problem in application to different network problems. Mathematical modelling. 2015. V. 27, N 12. P. 121–136. (in Russian) arXiv:1410.3123

4. *Babicheva T., Gasnikov A., Lagunovskaya A., Mendel M.* Two-stage model of equilibrium distributions of traffic flows. 2015. V. 7, N 3. P. 31–41. (in Russian) https://mipt.ru/upload/medialibrary/971/31-41.pdf

5. *Gasnikov A., Gasnikova E., Matsievsky S., Usik I.* Searching of equilibriums in hierarchical congestion population games. Proceedings of MIPT. 2015. V. 7, N 4. P. 129–142. (in Russian) arXiv:1511.02390

6. *Gasnikov A., Gasnikova E., Nesterov Y., Chernov A.* Entropy linear programming. Computational Mathematics and Mathematical Physics. 2016. V. 56, N 4. P. 523–534. (In Russian). arXiv:1410.7719

7. *Agueh M., Carlier G.* Barycenters in the Wasserstein space. SIAM J. Math. Anal. 2011. V. 43. N 2. P. 904–924.

8. *Cuturi M., Doucet A.* Fast Computation of Wasserstein Barycenters. ICML, 2014. http://www.iip.ist.i.kyoto-u.ac.jp/member/cuturi/

9. *Benamou J.D., Carlier G., Cuturi M., Nenna L., Peyré G.* Iterative Bregman Projections for Regularized Transportation Problems. e-print, 2015. (to appear in SISC). arXiv:1412.5154

10. *Cuturi M., Peyré G., Rolet A.* A Smoothed Dual Formulation for Variational Wasserstein Problems. e-print, 2015. arXiv:1503.02533

11. *Cuturi M.* Sinkhorn Distances: Lightspeed Computation of Optimal Transport. NIPS. 2013.

12. *Nemirovsky A., Yudin D.* Problem Complexity and Method Efficiency in Optimization. SIAM Review 27.2 (1985): 264.

13. *Boissard E., Le Gouic T., Loubes J.-M.* Distribution's template estimate with Wasserstein metrics. e-print, 2013. (to be published in Bernoulli). arXiv:1111.5927

14. *Bigot J., Klein T.* Consistent estimation of a population barycenter in the Wasserstein space. e-print, 2015. arXiv:1212.2562

15. *Nesterov Y.* Smooth minimization of nonsmooth function. Math. Program. Ser. A. 2005. V. 103. N 1. P. 127–152.

16. *Fercoq O., Richtarik P.* Accelerated, Parallel and Proximal Coordinate Descent. e-print, 2013. arXiv:1312.5799

17. *Qu Z., Richtarik P.* Coordinate Descent with Arbitrary Sampling I: Algorithms and Complexity. e-print, 2014. arXiv:1412.8060

18. *Boyd S., Parikh N., Chu E., Peleato B., Eckstein J.* Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends in Machine Learning. 2011. V. 3(1). P. 1–122. http://stanford.edu/boyd/papers.html

19. *Boyd S., Vandenberghe L.* Convex optimization. Cambridge University Press, 2004. http://stanford.edu/boyd/cvxbook/

20. *Polyak B.* Introduction to optimization. New York: Optimization Software, 1987.

21. *Nemirovski A.* Lectures on modern convex optimization analysis, algorithms, and engineering applications. Philadelphia: SIAM, 2013.
http://www2.isye.gatech.edu/simnemirovs/Lect_ModConvOpt.pdf

22. *Rabin J., Peyér G., Delon J. Bernot M.* Wasserstein barycenter and its applications to texture mixing. LNCS. 2011. Proc. SSVM'11. Springer. V. 6667. P. 435–446. https://hal.archives-ouvertes.fr/hal-00476064/document

23. *Bonnel N., Pfister H.* Sliced Wasserstein barycenter of multiple densities. Harvard Technical Report. 2013. TR-02-13.
ftp://ftp.deas.harvard.edu/techreports/tr-02-13.pdf

24. *Stetsyuk P.* Ellipsoid methods and r-algorithms. Kishinev: Evrika. 2014. (in Russian)

25. *Stetsyuk P., Gasnikov A.* NLP-programms and r-algorithms in linear-entropic programming. Optimal Decision Theory. Kiev: Glushkov Institute of Cybernetics NASU, 2015. P. 73–78.

26. *Franklin J., Lorenz J.* On the scaling of multidimensional matrices. Linear Algebra and its applications. 1989. V. 114. P. 717–735.

27. *Nesterov Yu.* Universal gradient methods for convex optimization problems. CORE Discussion Paper 2013/63. 2013; Math. Program. Ser. A. 2015. V. 152 P. 381–404. https://www.uclouvain.be/cps/ucl/doc/core/documents/coredp2013_26web.pdf

28. *Nesterov Yu.* http://www.youtube.com/watch?v=Fm9h92pcbvg

29. *Gasnikov A., Dvurechensky P., Dorn Y., Maksimov Y.* Searching equillibriums in Beckmann's and Nesterov–de Palma's models. Mathematical modelling. 2016. V. 28, N 10. P. 40–60. (in Russian). arXiv:1506.00293

30. *Gasnikov A., Gasnikova E., Ershov E., Dvurechensky P., Lagunovskaya A.* Search for the stochastic equilibria in the transport models of equilibrium flow distribution. Proceedings of MIPT. 2015. V. 7, N 4. C. 114–128. arXiv:1505.07492

31. *Gasnikov A., Dvurechensky P., Kamzolov D.* Gradient and direct methods with inexact oracle in stochastic optimization. Proceedings of the International Conference «Systems Dynamics and Control Processes» (SDCP'2014) dedicated to the 90th Anniversary of Academician Nikolay Nikolayevich Krasovskii, P. 111–117. Ekaterinburg, 2014. (in Russian) arXiv:1502.06259

32. *Devolder O.* Exactness, inexactness and stochasticity in first-order methods for large-scale convex optimization. CORE UCL, PhD thesis, March 2013.

33. *Anikin A., Dvurechensky P., Gasnikov A., Golov A., Gornov A., Maximov Yu., Mendel M., Spokoiny V.* Modern efficient numerical approaches to regularized regression problems in application to traffic demands matrix calculation from link loads. Proceedings of International Conference ITAS-2015. Russia, Sochi, September, 2015. arXiv:1508.00858

34. *Zhang F.* The Schur Complement and Its Applications. Springer, 2005.

Поступила в редакцию 09.03.2016