© 2019 г. А.Б. ЮДИЦКИЙ, канд. тех. наук
(anatoli.juditsky@univ-grenoble-alpes.fr)
(LJK, Университет Гренобль-Альпы, Гренобль, Франция),
А.С. НЕМИРОВСКИЙ, д-р физ.-мат. наук (nemirovs@isye.gatech.edu)
(ISyE, Технологический институт Джорджии, Атланта, США)

ВОССТАНОВЛЕНИЕ СИГНАЛОВ С ПОМОЩЬЮ СТОХАСТИЧЕСКОЙ ОПТИМИЗАЦИИ[1]

Рассматривается подход к восстановлению сигналов в обобщенных линейных моделях, при котором задача оценивания сигналов сводится к решению стохастических монотонных вариационных неравенств (ВН). Решения таких ВН могут быть получены с помощью эффективно вычислительных процедур, а в случае сильно монотонных ВН допускают верхнюю границу на конечном времени для ожидаемой ошибки $\|\cdot\|_2^2$, сходящуюся к нулю со скоростью $O(1/K)$ с ростом числа K наблюдений. Принятые структурные предположения существенно слабее тех, которые необходимы для обеспечения выпуклости оптимизационной задачи, возникающей при применении метода максимального правдоподобия. Прослеживается связь предлагаемого подхода с идеями, лежащими в основе алгоритма персептрона Розенблата.

Ключевые слова: обобщенные линейные модели, задача статистического оценивания, стохастическая оптимизация, вариационные неравенства.

DOI: 10.1134/S0005231019100088

1. Введение

Задачи статистического оценивания составляют одну из важнейших прикладных областей стохастической оптимизации. Типичная постановка задачи выглядит следующим образом (например, см. [1] и обширную библиографию, приведенную в этой книге). Допустим, что имеются независимые одинаково распределенные наблюдения $\omega^K = (\omega_1, \ldots, \omega_K)$, $\omega_k = (\eta_k, y_k)$, где $\eta_k \in \mathbf{R}^{n\times m}$, $y_k \in \mathbf{R}^m$ — реализации регрессоров (независимые переменные) и отклики (метки) соответственно. Предполагаем, что наблюдения описываются обобщенной линейной моделью (ОЛМ) [2, 3], т.е. условное (по η) математическое ожидание y равно $f(\eta^{\mathrm{T}}x)$, где верхний индекс $^{\mathrm{T}}$ — знак транспонирования, $f(\cdot) : \mathbf{R}^m \to \mathbf{R}^m$ — известная *функция связи*, а $x \in \mathbf{R}^m$ — неизвестный "сигнал", т.е. вектор параметров модели. Цель состоит в "подгонке модели", т.е. в восстановлении вектора x по наблюдениям ω^K. При стандартном подходе к решению задачи оценивания выбирают функцию потерь $\ell(y, \theta) : \mathbf{R}^m \times \mathbf{R}^m \to \to \mathbf{R}$ и в качестве оценки x принимают оптимальное решение оптимизацион-

[1] Работа первого автора поддержана грантом 2016-2032H, PGMO. Работа первого и второго авторов финансировалась грантом CCF-1523768, NSF.

ной задачи

$$\min_{u\in\mathcal{X}} \mathbf{E}_{\omega\sim P_x}\left\{\ell(y, f(\eta^{\mathrm{T}}u))\right\}, \tag{1}$$

где P_x — функция распределения наблюдений $\omega = (\eta, y)$, отвечающая "истинному сигналу x", а $\mathcal{X}$ — априори известное множество сигналов. Иными словами, задача статистического оценивания сводится к задаче стохастической оптимизации (1), которая должна решаться приближенно, исходя из имеющихся наблюдений ω^K. Это может быть сделано или напрямую ("пакетно") путем минимизации по $u \in \mathcal{X}$ аппроксимации выборочного среднего (ВСА)

$$\frac{1}{K}\sum_{k=1}^{K} \ell(y_k, f(\eta_k^{\mathrm{T}}u)) \tag{2}$$

для математического ожидания в (1) (например, см. [4]) или путем применения итеративных алгоритмов стохастической оптимизации типа стохастической аппроксимации (СА) [5, 6].

Пусть функция условного по η распределения $P^x_{|\eta}$ величины y, индуцированная распределением P_x, принадлежит известному параметрическому семейству $\mathcal{P} = \{P^\theta : \theta \in \Theta \subset \mathbf{R}^m\}$, а именно, $P^x_{|\eta} = P^{f(\eta^{\mathrm{T}}x)}$; тогда стандартный выбор функции потерь дается методом максимального правдоподобия (МП). Иными словами, считая, что распределения P^θ имеют плотности p_θ относительно некоторой меры П, полагают

$$\ell(y, \theta) = -\ln(p_\theta(y)).$$

Например, в классической *логистической регрессии* имеем $m = 1$, $f(s) = (1 + \mathrm{e}^{-s})^{-1}$, $\Theta = (0, 1)$, а P^θ, $\theta \in \Theta$, — распределение Бернулли, т.е. y принимает значение единица с вероятностью $q = (1 + \exp\{-\eta^{\mathrm{T}}x\})^{-1}$ и нуль с вероятностью $1 - q$, что приводит к

$$\ell(y, f(\eta^{\mathrm{T}}u)) = \ln\left(1 + \exp\{\eta^{\mathrm{T}}u\}\right) - y\eta^{\mathrm{T}}u. \tag{3}$$

В этом случае задача (1) становится оптимизационной задачей вида

$$\min_{u\in\mathcal{X}} \mathbf{E}_{(\eta,y)\sim P_x}\left\{\ln(1 + \exp\{\eta^{\mathrm{T}}u\}) - y\,\eta^{\mathrm{T}}u\right\}, \tag{4}$$

и соответствующая задача ВСА записывается как

$$\min_{u\in\mathcal{X}} \frac{1}{K}\sum_{k=1}^{K}\left[\ln(1 + \exp\{\eta_k^{\mathrm{T}}u\}) - y_k\eta_k^{\mathrm{T}}u\right], \tag{5}$$

а ее оптимальное решение $\widehat{x}_{\mathrm{ML}}(\omega^K)$ и есть МП-оценка сигнала x. При предположении о выпуклости множества $\mathcal{X}$ сигналов обе эти задачи выпуклы и, следовательно, глобальный оптимум в задаче ВСА может быть эффективно вычислен, а для алгоритмов СА для (4) могут быть получены гарантированные скорости сходимости.

В более общем случае, когда распределения наблюдений образуют *семейство условных экспоненциальных распределений* [7, 8], функция правдоподобия имеет вид

$$\ell(y, \eta^{\mathrm{T}} u) = F(\eta^{\mathrm{T}} u) - y\eta^{\mathrm{T}} u$$

с выпуклой *кумулянтной функцией* F, а соответствующая задача (1) минимизации риска записывается как

$$\min_{u \in \mathcal{X}} \mathbf{E}_{(\eta,y)\sim P_x} \left\{ F(\eta^{\mathrm{T}} u) - y\eta^{\mathrm{T}} u \right\}. \tag{6}$$

В этом случае, так же как и при логистической регрессии, для получения МП-оценок параметров модели можно применять ВСА или СА.

Заметим, однако, что предположение о принадлежности распределений экспоненциальному семейству довольно ограничительно. С другой стороны, для распределений, которые не удовлетворяют этому предположению, выпуклость оптимизационной задачи, получающейся при выборе функции потерь $\ell(\cdot)$ по методу максимального правдоподобия, является скорее исключением, чем правилом. Например, рассмотрим нелинейную схему наименьших квадратов, в которой величина y получается из $f(\eta^{\mathrm{T}} x)$ добавлением центрированного гауссовского шума, не зависящего от регрессоров:

$$y = f(\eta^{\mathrm{T}} x) + \xi, \quad \xi \sim \mathcal{N}(0, \sigma^2 I_m).$$

В этом случае задача (1) и ее ВС-аппроксимация при выборе $\ell(\cdot)$ по методу максимального правдоподобия приобретают вид

$$\begin{gathered} \min_{u \in \mathcal{X}} \mathbf{E}_{\eta\sim Q} \Big\{ \|f(\eta^{\mathrm{T}} x) - f(\eta^{\mathrm{T}} u)\|_2^2 \Big\}, \\ \min_{u \in \mathcal{X}} \left\{ \frac{1}{K} \sum_{k=1}^{k} \|y_k - f(\eta_k^{\mathrm{T}} u)\|_2^2 \right\}, \end{gathered} \tag{7}$$

где Q — распределение регрессоров (они предполагаются независимыми от сигнала). Если функция f нелинейна, то обе эти задачи обычно невыпуклы, и их численное решение затруднено. Аналогично, в задаче "неэкспоненциальной логистической регрессии", в которой сигмоидная функция $f(s) = (1 + \exp\{-s\})^{-1}$ заменена на общего вида неубывающую функцию связи $f(s) : \mathbf{R} \to (0, 1)$ (например, так называемая *пробит-функция* или *дополняющая* $\log - \log$ *функция*), МП-выбор функции потерь, как правило, влечет невыпуклость задачи (1) и ее ВС-аппроксимации.

Цель последующего изложения — предложить альтернативу решению задачи настройки модели, основанному на решении оптимизационной задачи (1) с МП-выбором функции потерь при оценивании сигнала по наблюдениям в ОЛМ. Идеи, лежащие в основе подхода, предлагаемого в настоящей статье, восходят к итеративному алгоритму персептрона Розенблата [9, 10] и его "пакетной" версии [11]. В этом случае структурные требования, предъявляемые к модели, оказываются существенно слабее тех, которые приводят к вы-

пуклости задач (1) с МП-выбором функции потерь и их ВС-аппроксимаций[2]. При этих предположениях вместо того, чтобы использовать классический подход с формированием задачи стохастической оптимизации (1) и функции потерь [12–16], задача оценивания сводится к другой задаче с выпуклой структурой — *сильно монотонному вариационному неравенству* (ВН), *представимому в виде стохастического оракула.* Полученное вариационное неравенство может быть эквивалентно задаче выпуклой минимизации (например, при $m = 1$ такое ВН эквивалентно задаче выпуклой оптимизации, аналогичной (6)); но даже и в этом случае получающаяся задача, как правило, отличается от МП-версии задачи (1). Решение такого ВН может быть получено в результате эффективной численной процедуры, и оно оказывается "хорошей" оценкой "истинного" сигнала по имеющимся наблюдениям. Для этой оценки здесь доказываются верхние границы на ожидаемую квадратичную ошибку в норме $\|\cdot\|_2^2$, сходящуюся к нулю со скоростью $O(1/K)$ при $K \to \infty$[3].

2. Постановка задачи

Рассматривается обобщенная линейная модель наблюдений, характеризующаяся следующими предположениями.

Наблюдения зависят от неизвестного сигнала x, принадлежащего заданному выпуклому компактному множеству $\mathcal{X} \subset \mathbf{R}^n$, и представлены в виде

$$\omega^K = \{\omega_k = (\eta_k, y_k), 1 \leqslant k \leqslant K\}, \tag{8}$$

где ω_k, $1 \leqslant k \leqslant K$, — независимые одинаково распределенные реализации случайной пары (η, y) с распределением P_x таким, что

• *регрессор* η является случайной $(n \times m)$-матрицей с некоторым вероятностным распределением Q, *не зависящим от* x;

• *отклик* ("метка") y является m-мерным случайным вектором таким, что его условное по η распределение, индуцированное распределением P_x, имеет математическое ожидание $f(\eta^{\mathrm{T}} x)$:

$$\mathbf{E}^x_{|\eta}\{y\} = f(\eta^{\mathrm{T}} x), \tag{9}$$

где $\mathbf{E}^x_{|\eta}$ — условное по η математическое ожидание метки y, задаваемое распределением P_x величины $\omega = (\eta, y)$, а $f(\cdot) : \mathbf{R}^m \to \mathbf{R}^m$ — заданное отображение.

Ниже будут сформулированы предположения о параметрах обобщенной линейной модели (а именно, условия на $f(\cdot)$ и на распределение P_x, $x \in \mathcal{X}$ пары (η, y)), необходимые для обоснования предлагаемого подхода.

[2] Например, в нелинейной схеме наименьших квадратов с $m = 1$ и в неэкспоненциальной логистической регрессии от функции f требовалась лишь непрерывная дифференцируемость и положительность производной, а от множества $\mathcal{X}$ сигналов — выпуклость.

[3] Несмотря на простоту приводимых ниже выкладок, авторам не удалось найти упоминание о предлагаемом подходе в статистической литературе.

2.1. Предварительные сведения: монотонные векторные поля

Монотонное векторное поле на $\mathbf{R}^m$ — это однозначное всюду определенное отображение $g(\cdot) : \mathbf{R}^m \to \mathbf{R}^m$, обладающее *свойством монотонности*:

$$[g(z) - g(z')]^{\mathrm{T}}[z - z'] \geqslant 0 \quad \forall z,\ z' \in \mathbf{R}^m.$$

Скажем, что такое поле *монотонно с модулем* $\varkappa \geqslant 0$ *на замкнутом выпуклом множестве* $Z \subset \mathbf{R}^m$, если

$$[g(z) - g(z')]^{\mathrm{T}}[z - z'] \geqslant \varkappa \|z - z'\|_2^2 \quad \forall z,\ z' \in Z;$$

назовем g *сильно монотонным* на Z, если модуль монотонности $\varkappa$ поля g на Z положителен. Очевидно, что для монотонного векторного поля f, непрерывно дифференцируемого на замкнутом выпуклом множестве Z с непустой внутренностью, условие

$$d^{\mathrm{T}} f'(z) d \geqslant \varkappa d^{\mathrm{T}} d \quad \forall (d \in \mathbf{R}^n, z \in Z) \tag{10}$$

является необходимым и достаточным условием монотонности с модулем $\varkappa$. Стандартными примерами монотонных векторных полей являются

- градиентные поля $\nabla\phi(x)$ непрерывно дифференцируемых выпуклых функций m переменных, а также векторные поля $[\nabla_x \phi(x,y); -\nabla_y \phi(x,y)]$, порожденные непрерывно дифференцируемыми функциями $\phi(x,y)$, выпуклыми по x и вогнутыми по y;
- "диагональные" векторные поля $f(x) = [f_1(x_1); \ldots; f_m(x_m)]$ с монотонно неубывающими компонентами $f_i(\cdot)$ — функциями одного аргумента. Если, кроме того, $f_i(\cdot)$ непрерывно дифференцируемы с положительными производными, то соответствующее поле f сильно монотонно на любом компактном выпуклом подмножестве в $\mathbf{R}^m$ с модулем монотонности, зависящим от подмножества.

Монотонные векторные поля на $\mathbf{R}^n$ допускают выполнение простых операций, включающих, например, следующие:

I. [аффинная подстановка аргумента]: Если $f(\cdot)$ — монотонное векторное поле на $\mathbf{R}^m$ и A есть $(n \times m)$-матрица, то векторное поле

$$g(x) = Af(A^{\mathrm{T}}x + a)$$

также монотонно на $\mathbf{R}^n$; если, кроме того, поле f монотонно с модулем $\varkappa \geqslant 0$ на замкнутом выпуклом множестве $Z \subset \mathbf{R}^m$, а множество $X \subset \mathbf{R}^n$ замкнуто, выпукло и обладает свойством $A^{\mathrm{T}}x + a \in Z$ для всех $x \in X$, то поле g монотонно с модулем $\sigma^2\varkappa$ на X, где σ — минимальное сингулярное значение матрицы A.

II. [суммирование]: Пусть S — польское пространство, $f(x,s) : \mathbf{R}^m \times S \to \to \mathbf{R}^m$ — борелевская векторнозначная функция, монотонная по x при каждом $s \in S$, а $\mu(ds)$ — борелевская вероятностная мера на S такая, что векторное поле

$$F(x) = \int\limits_S f(x,s)\mu(ds)$$

определено для всех x; тогда $F(\cdot)$ монотонно. Если, кроме того, X — замкнутое выпуклое множество в $\mathbf{R}^m$, а $f(\cdot, s)$ монотонно на X с борелевским по s модулем $\varkappa(s)$ для всякого $s \in S$, то F монотонно на X с модулем $\int_S \varkappa(s)\mu(ds)$.

2.2. Предположения

В дальнейшем потребуются следующие предположения об основных компонентах задачи оценивания, описанной во введении.

A.1. Векторное поле $f(\cdot)$ непрерывно и монотонно, а векторное поле

$$F(z) = \mathbf{E}_{\eta\sim Q}\left\{\eta f(\eta^{\mathrm{T}} z)\right\}$$

корректно определено (следовательно, монотонно, как и поле f, см. **I**, **II**).

A.2. Множество $\mathcal{X}$ сигналов непусто, выпукло и компактно, а векторное поле F монотонно с положительным модулем $\varkappa$ на $\mathcal{X}$.

A.3. Для соответствующим образом выбранного $M < \infty$ и всякого $x \in \mathcal{X}$ верно

$$\mathbf{E}_{(\eta,y)\sim P_x}\left\{\|\eta y\|_2^2\right\} \leqslant M^2. \tag{11}$$

Простое *достаточное* условие выполнения предположений **A.1**–**A.3** с соответствующим образом выбранным $M < \infty$ и $\varkappa > 0$ заключается в следующем:

- распределение Q регрессора η имеет конечные моменты всех порядков и $\mathbf{E}_{\eta\sim Q}\{\eta\eta^{\mathrm{T}}\} \succ 0$;
- функция f непрерывно дифференцируема, и $d^{\mathrm{T}} f'(z) d > 0$ для всех $d \neq 0$ и всех z. Кроме того, f имеет полиномиальный рост, т.е. для некоторых констант $C \geqslant 0$ и $p \geqslant 0$ и всех z выполнено $\|f(z)\|_2 \leqslant C(1 + \|z\|_2^p)$.

Проверка достаточности осуществляется непосредственно.

3. Построения и основной результат

Сформулируем ключевое наблюдение, лежащее в основе представляемых ниже конструкций.

Предложение 1. Пусть выполнены предположения **A.1**–**A.3**. *Сопоставим паре* $(\eta, y) \in \mathbf{R}^{n\times m} \times \mathbf{R}^m$ *векторное поле*

$$G_{(\eta,y)}(z) = \eta f(\eta^{\mathrm{T}} z) - \eta y : \mathbf{R}^n \to \mathbf{R}^n. \tag{12}$$

Тогда для любого $x \in \mathcal{X}$ *имеем*

$$\begin{aligned}
&(a) \qquad \mathbf{E}_{(\eta,y)\sim P_x}\left\{G_{(\eta,y)}(z)\right\} = F(z) - F(x) \quad \forall z \in \mathbf{R}^n \\
&(b) \qquad \|F(z)\|_2 \leqslant M \quad \forall z \in \mathcal{X} \\
&(c) \quad \mathbf{E}_{(\eta,y)\sim P_x}\left\{\|G_{(\eta,y)}(z)\|_2^2\right\} \leqslant 4M^2 \quad \forall z \in \mathcal{X}.
\end{aligned} \tag{13}$$

Доказательство. Пусть $x \in \mathcal{X}$. Тогда, пользуясь соотношением (9) и определением поля F, имеем

$$\mathbf{E}_{(\eta,y)\sim P_x}\{\eta y\} = \mathbf{E}_{\eta\sim Q}\left\{\mathbf{E}^x_{|\eta}\{\eta y\}\right\} = \mathbf{E}_\eta\left\{\eta f(\eta^{\mathrm{T}}x)\right\} = F(x).$$

Отсюда получаем

$$\begin{aligned}\mathbf{E}_{(\eta,y)\sim P_x}\left\{G_{(\eta,y)}(z)\right\} &= \mathbf{E}_{(\eta,y)\sim P_x}\left\{\eta f(\eta^{\mathrm{T}}z) - \eta y\right\} = \\ &= \mathbf{E}_{(\eta,y)\sim P_x}\left\{\eta f(\eta^{\mathrm{T}}z)\right\} - F(x) = \\ &= \mathbf{E}_{\eta\sim Q}\left\{\eta f(\eta^{\mathrm{T}}z)\right\} - F(x) = F(z) - F(x);\end{aligned}$$

последнее неравенство следует из (13.a). Кроме того, для $x, z \in \mathcal{X}$ обозначим через $P^z_{|\eta}$ условное по η распределение величины z, индуцированное распределением P_z величины $(\eta, y) = \eta^{\mathrm{T}}y$. С учетом того, что маргинальное распределение величины η, индуцированное распределением P_z, есть не что иное как Q, приходим к

$$\begin{aligned}&\mathbf{E}_{(\eta,y)\sim P_x}\left\{\|\eta f(\eta^{\mathrm{T}}z)\|_2^2\right\} = \mathbf{E}_{\eta\sim Q}\left\{\|\eta f(\eta^{\mathrm{T}}z)\|_2^2\right\} = \\ &= \mathbf{E}_{\eta\sim Q}\left\{\|\mathbf{E}_{y\sim P^z_{|\eta}}\{\eta y\}\|_2^2\right\} \leqslant && [\text{поскольку } \mathbf{E}_{y\sim P^z_{|\eta}}\{y\} = f(\eta^{\mathrm{T}}z)] \\ &\leqslant \mathbf{E}_{\eta\sim Q}\left\{\mathbf{E}_{y\sim P^z_{|\eta}}\left\{\|\eta y\|_2^2\right\}\right\} = && [\text{по неравенству Йенсена}] \\ &= \mathbf{E}_{(\tilde\eta,y)\sim P_z}\left\{\|\eta y\|_2^2\right\} \leqslant M^2 && [\text{по } \mathbf{A.3}, \text{ поскольку } z \in \mathcal{X}].\end{aligned}$$

В совокупности с соотношением $\mathbf{E}_{(\eta,y)\sim P_x}\{\|\eta y\|_2^2\} \leqslant M^2$, справедливым в соответствии с **A.3** (поскольку $x \in \mathcal{X}$), это влечет выполнение (13.b) и (13.c).

3.1. Основной результат

Напомним, что целью статьи является восстановление сигнала $x \in \mathcal{X}$ по наблюдениям (8). В предположениях **A.1**–**A.3** точка x есть корень монотонного векторного поля

$$G(z) = F(z) - F(x), \quad F(z) = \mathbf{E}_{\eta\sim Q}\left\{\eta f(\eta^{\mathrm{T}}z)\right\}; \tag{14}$$

причем известно, что этот корень принадлежит множеству $\mathcal{X}$ и он единствен, поскольку поле $G(\cdot)$ сильно монотонно на $\mathcal{X}$, равно как и поле $F(\cdot)$. Известно, что для заданного выпуклого компактного множества $\mathcal{X}$ задача отыскания такого корня эффективно разрешима при наличии "оракула", который, имея на входе точку $z \in \mathcal{X}$, выдает значение $G(z)$ поля в этой точке. Это не совсем та ситуация, которая здесь описана, поскольку поле G есть математическое ожидание случайного поля:

$$G(z) = \mathbf{E}_{(\eta,y)\sim P_x}\left\{\eta f(\eta^{\mathrm{T}}z) - \eta y\right\}$$

и заранее не известно, каково то распределение, по которому берется ожидание. Однако выборка из этого распределения доступна, причем выборочные

значения в точности и есть наблюдения (8); эти выборочные значения могут использоваться для аппроксимации поля G тем или иным способом с последующим применением этой аппроксимации для восстановления сигнала x. Два стандартных пути реализации этой простой идеи суть ВСА и СА, упоминавшиеся выше. Обсудим эти два метода применительно к ситуации, в которой находимся.

3.1.1. Оценивание: аппроксимация выборочным средним. Идея, лежащая в основе аппроксимации по выборочному среднему (ВСА), совершенно прозрачна: имея наблюдения (8), будем аппроксимировать неизвестное поле G его эмпирическим аналогом

$$G_{\omega^K}(z) = \frac{1}{K}\sum_{k=1}^{K}\left[\eta_k f(\eta_k^{\mathrm{T}} z) - \eta_k y_k\right].$$

По закону больших чисел при $K \to \infty$ эмпирическое поле G_{ω^K} сходится к интересующему нас полю G, поэтому при некоторых необременительных условиях регулярности с большой вероятностью поле G_{ω^K} будет равномерно на $\mathcal{X}$ близкó к полю G для достаточно больших значений K. Вследствие сильной монотонности G отсюда следует, что множество "почти нулей" поля G_{ω^K} на $\mathcal{X}$ будет близкó к нулю x поля G, т.е. к сигналу, который и подлежит восстановлению. Вопрос теперь заключается в том, как правильно определить понятие "почти нуля" поля G_{ω^K} на $\mathcal{X}$[4]. В данной ситуации это удобно сделать исходя из понятия *слабого решения* вариационного неравенства (ВН) с монотонным оператором, определяемого (в данном конкретном случае) следующим образом.

Пусть $\mathcal{X} \subset \mathbf{R}^n$ — непустое выпуклое компактное множество и пусть $H(z)$: $\mathcal{X} \to \mathbf{R}^n$ — монотонное векторное поле (т.е. $[H(z)-H(z')]^{\mathrm{T}}[z-z'] \geqslant 0$ для всех $z, z' \in \mathcal{X}$). Вектор $z_* \in \mathcal{X}$ называется *слабым решением* вариационного неравенства, связанного с $H, \mathcal{X}$, если

$$H(z)^{\mathrm{T}}(z - z_*) \geqslant 0 \quad \forall z \in \mathcal{X}.$$

Пусть $\mathcal{X} \subset \mathbf{R}^n$ — непустое выпуклое компактное множество и пусть H монотонно на $\mathcal{X}$. Хорошо известно, что

- ВН, связанное с $H, \mathcal{X}$ (обозначим его через $\mathrm{VI}(H, \mathcal{X})$), всегда имеет слабое решение. Ясно, что если $\bar{z} \in \mathcal{X}$ — корень H, то $\bar{z}$ — слабое решение для $\mathrm{VI}(H, \mathcal{X})$.[5]
- Если поле H непрерывно на $\mathcal{X}$, то любое слабое решение $\bar{z}$ неравенства $\mathrm{VI}(H, \mathcal{X})$ также является и *сильным решением* в том смысле, что

$$H^{\mathrm{T}}(\bar{z})(z - \bar{z}) \geqslant 0 \ \forall z \in \mathcal{X}. \tag{15}$$

[4] Заметим, что "почти нуль" поля G_{ω^K} на $\mathcal{X}$ не может быть определен как корень G_{ω^K} на этом множестве. Если у G действительно имеется корень, принадлежащий множеству $\mathcal{X}$, то у G_{ω^K} такого корня может и не быть.

[5] Если $\bar{z} \in \mathcal{X}$ и $H(\bar{z}) = 0$, то монотонность H влечет $H(z)^{\mathrm{T}}[z-\bar{z}] = [H(z)-H(\bar{z})]^{\mathrm{T}}[z-\bar{z}] \geqslant$ $\geqslant 0$ для всех $z \in \mathcal{X}$, т.е. $\bar{z}$ — слабое решение для $\mathrm{VI}(H, \mathcal{X})$.

Действительно, соотношение (15) очевидным образом выполнено для $z = \bar{z}$. При $z \neq \bar{z}$, положив $z_t = \bar{z} + t(z - \bar{z})$, $0 < t \leqslant 1$, имеем $H^{\mathrm{T}}(z_t)(z_t - \bar{z}) \geqslant 0$ (так как $\bar{z}$ является слабым решением), откуда $H^{\mathrm{T}}(z_t)(z - \bar{z}) \geqslant 0$ (поскольку $z - \bar{z}$ кратно $z_t - \bar{z}$ с положительным множителем). Переходя к пределу при $t \to +0$ и вспоминая про непрерывность H, получаем желаемое $H^{\mathrm{T}}(\bar{z})(z - \bar{z}) \geqslant 0$.

- Если H — градиентное поле непрерывно дифференцируемой выпуклой на $\mathcal{X}$ функции (такое поле монотонно), слабые (или сильные, что в случае непрерывного H одно и то же) решения неравенства $\mathrm{VI}(H, \mathcal{X})$ являются точками минимума функции на $\mathcal{X}$.

  Заметим еще, что сильное решение неравенства $\mathrm{VI}(H, \mathcal{X})$ с монотонным H всегда является и слабым решением: если $\bar{z} \in \mathcal{X}$ удовлетворяет условию $H^{\mathrm{T}}(\bar{z})(z - \bar{z}) \geqslant 0$ при всех $z \in \mathcal{X}$, то и $H(z)^{\mathrm{T}}(z - \bar{z}) \geqslant 0$ при всех $z \in \mathcal{X}$, поскольку вследствие монотонности имеем $H(z)^{\mathrm{T}}(z - \bar{z}) \geqslant H^{\mathrm{T}}(\bar{z})(z - \bar{z})$.

В дальнейшем будем существенно использовать следующий простой и хорошо известный факт.

Лемма 1. Пусть $\mathcal{X}$ — выпуклое компактное множество и пусть H — монотонное векторное поле на $\mathcal{X}$ с модулем монотонности $\varkappa > 0$, т.е.

$$[H(z) - H(z')]^{\mathrm{T}}[z - z'] \geqslant \varkappa \|z - z'\|_2^2 \quad \forall z, z' \in X.$$

Пусть также $\bar{z}$ — слабое решение неравенства $\mathrm{VI}(H, \mathcal{X})$. Тогда слабое решение $\mathrm{VI}(H, \mathcal{X})$ единственно. Кроме того,

$$H^{\mathrm{T}}(z)[z - \bar{z}] \geqslant \varkappa \|z - \bar{z}\|_2^2. \tag{16}$$

Доказательство. В условиях леммы положим $z \in \mathcal{X}$, и пусть $\bar{z}$ есть слабое решение неравенства $\mathrm{VI}(H, \mathcal{X})$ (помним, что оно существует!). Положив $z_t = \bar{z} + t(z - \bar{z})$ при $t \in (0, 1)$, получаем цепочку неравенств

$$H^{\mathrm{T}}(z)[z - z_t] \geqslant H^{\mathrm{T}}(z_t)[z - z_t] + \varkappa \|z - z_t\|^2 \geqslant \varkappa \|z - z_t\|^2,$$

первое из которых справедливо в силу сильной монотонности H, а второе следует из того, что величина $H^{\mathrm{T}}(z_t)[z - z_t]$ кратна $H^{\mathrm{T}}(z_t)[z_t - \bar{z}]$ с положительным коэффициентом. Последняя величина неотрицательна, поскольку $\bar{z}$ является слабым решением исследуемого ВН. Таким образом, имеем $H^{\mathrm{T}}(z)(z - z_t) \geqslant \varkappa \|z - z_t\|_2^2$ и, переходя к пределу при $t \to +0$, приходим к (16).

Для доказательства единственности слабого решения предположим, что помимо слабого решения $\bar{z}$ имеется отличное от него слабое решение $\widetilde{z}$, и положим $z' = \frac{1}{2}[\bar{z} + \widetilde{z}]$. Поскольку и $\bar{z}$, и $\widetilde{z}$ — слабые решения, то обе величины $H^{\mathrm{T}}(z')[z' - \bar{z}]$ и $H^{\mathrm{T}}(z')[z' - \widetilde{z}]$ должны быть неотрицательны, а поскольку в сумме они равны нулю, то каждое из них равно нулю. Поэтому, применяя (16) к $z = z'$, получаем $z' = \bar{z}$, откуда также и $\widetilde{z} = \bar{z}$.

Вернемся теперь к исходной задаче оценивания. Пусть предположения **A.1**–**A.3** выполнены, так что векторные поля $G_{(\eta_k, y_k)}(z)$, определенные в (12),

а следовательно, и векторное поле $G_{\omega^K}(z)$ непрерывны и монотонны. При использовании ВСА было найдено слабое решение $\widehat{x}(\omega^K)$ вариационного неравенства $\mathrm{VI}(G_{\omega^K}, \mathcal{X})$ и принято в качестве ВСА-оценки сигнала x по наблюдениям (8). Поскольку векторное поле $G_{\omega^K}(\cdot)$ монотонно и его значения эффективно вычислимы — при условии, что таковым является f, — задача вычисления слабого решения ВН $\mathrm{VI}(G_{\omega^K}, \mathcal{X})$ (точнее, очень хорошего приближения к нему) также является эффективно разрешимой (например, см. [17]). Хотя авторы не ставили перед собой такой задачи в настоящей работе, используя методологию из [16, 18–20] и дополняя предположения **A.1**–**A.3** необременительными условиями регулярности, можно доказать неасимптотическую верхнюю границу, например, для ожидаемой ошибки $\|\cdot\|_2^2$ ВСА-оценки как функцию размера выборки K и найти скорость, с которой эта верхняя граница сходится к нулю с ростом $K \to \infty$.

Рассмотрим ВСА-оценки в схеме логистической регрессии. Имеем $f(u) = (1 + \mathrm{e}^{-u})^{-1}$ и

$$G_{(\eta_k, y_k)}(z) = \left[\frac{\exp\{\eta_k^{\mathrm{T}} z\}}{1 + \exp\{\eta_k^{\mathrm{T}} z\}} - y_k\right]\eta_k,$$

$$G_{\omega^K}(z) = \frac{1}{K}\sum_{k=1}^{K}\left[\frac{\exp\{\eta_k^{\mathrm{T}} z\}}{1 + \exp\{\eta_k^{\mathrm{T}} z\}} - y_k\right]\eta_k =$$

$$= \frac{1}{K}\nabla_z\left[\sum_k \left(\ln\left(1 + \exp\{\eta_k^{\mathrm{T}} z\}\right) - y_k \eta_k^{\mathrm{T}} z\right)\right].$$

Иначе говоря, $G_{\omega^K}(z)$ является градиентным полем отрицательного логарифма эмпирического МП $\ell(z, \omega^K)$, см. (5). В результате слабые решения ВН $\mathrm{VI}(G_{\omega^K}, \mathcal{X})$ дают в точности оптимальные решения задаче (5), т.е. *в схеме логистической регрессии ВСА-оценки есть оценки максимального правдоподобия* $\widehat{x}_{\mathrm{ML}}(\omega^K)$[6]. С другой стороны, в примере для нелинейной схемы наи-

[6] Заметим, что это явление специфично для модели логистической регрессии. То, что в этом случае ВСА-оценки совпадают с МП-оценками, объясняется тем, что логистическая сигмоида $f(s) = \exp\{s\}/(1 + \exp\{s\})$ удовлетворяет тождеству $f'(s) = f(s)(1 - f(s))$. При ее замене на $f(s) = \phi(s)/(1 + \phi(s))$ с дифференцируемой монотонно неубывающей положительной $\phi(\cdot)$ ВСА-оценка доствляет слабое решение вариационному неравенству $\mathrm{VI}(\Phi, \mathcal{X})$ с

$$\Phi(z) = \sum_k \left[\frac{\phi(\eta_k^{\mathrm{T}} z)}{1 + \phi(\eta_k^{\mathrm{T}} z)} - y_k\right]\eta_k.$$

С другой стороны, градиентное поле *отрицательного* логарифма МП

$$-\frac{1}{K}\sum_k \left[y_k \ln(f(\eta_k^{\mathrm{T}} z)) + (1 - y_k)\ln(1 - f(\eta_k^{\mathrm{T}} z))\right],$$

которое должно минимизироваться при отыскании МП-оценок, имеет вид

$$\Psi(z) = \sum_k \frac{\phi'(\eta_k^{\mathrm{T}} z)}{\phi(\eta_k^{\mathrm{T}} z)}\left[\frac{\phi(\eta_k^{\mathrm{T}} z)}{1 + \phi(\eta_k^{\mathrm{T}} z)} - y_k\right]\eta_k.$$

При $k > 1$ и неэкспоненциальной ϕ поля Φ и Ψ "существенно различны"; соответственно, ВСА-оценка, как правило, отлична от МП-оценки.

меньших квадратов, приведенном во введении, с монотонной (для простоты скалярной) функцией $f(\cdot)$ векторное поле $G_{\omega^K}(\cdot)$ имеет вид

$$G_{\omega^K}(z) = \frac{1}{K}\sum_{k=1}^{K}\left[f(\eta_k^{\mathrm{T}}z) - y_k\right]\eta_k,$$

"существенно отличный" (при условии нелинейности f) от градиентного поля

$$\Psi(z) = \frac{2}{K}\sum_{k=1}^{K} f'(\eta_k^{\mathrm{T}}z)\left[f(\eta_k^{\mathrm{T}}z) - y_k\right]\eta_k$$

для МП (7). В результате в этом случае МП-оценка (7), вообще говоря, отличается от ВСА-оценки, но в отличие от МП-оценки ВСА-оценка гораздо проще с вычислительной точки зрения.

3.1.2. Оценивание: стохастическая аппроксимация. Оценки метода *стохастической аппроксимации* (СА) генерируются простым алгоритмом *субградиентного спуска* решения вариационного неравенства $\mathrm{VI}(G, \mathcal{X})$. Если бы значения векторного поля $G(\cdot)$ были доступны, можно было бы аппроксимировать корень $x \in \mathcal{X}$ этого ВН, пользуясь рекурсией

$$z_k = \mathrm{Proj}_{\mathcal{X}}[z_{k-1} - \gamma_k G(z_{k-1})], \quad k = 1, \ldots, K,$$

где

- $\mathrm{Proj}_{\mathcal{X}}[z]$ — метрическая проекция $\mathbf{R}^n$ на $\mathcal{X}$:

$$\mathrm{Proj}_{\mathcal{X}}[z] = \operatorname*{argmin}_{u\in\mathcal{X}} \|z - u\|_2;$$

- $\gamma_k > 0$ — заданная длина шага;
- начальное приближение z_0 — произвольная точка из $\mathcal{X}$.

Хорошо известно, что в предположениях **A.1**–**A.3** такая рекуррентная процедура с надлежащим выбором длины шага и начальной точкой из $\mathcal{X}$ позволяет аппроксимировать корень поля G (на самом деле, также и любое слабое решение $\mathrm{VI}(G, \mathcal{X})$) с произвольно высокой точностью, если K достаточно велико. Здесь, однако, имеем ситуацию, когда истинные значения поля G недоступны; стандартный способ преодоления этой трудности заключается в замене "ненаблюдаемых" значений $G(z_{k-1})$ поля G, присутствующих в рекуррентной процедуре, их несмещенными случайными оценками $G_{(\eta_k, y_k)}(z_{k-1})$. Такая модификация приводит к процедуре *стохастической аппроксимации* (восходящей к [21]) — рекурсии вида

$$z_k = \mathrm{Proj}_{\mathcal{X}}[z_{k-1} - \gamma_k G_{(\eta_k, y_k)}(z_{k-1})], \quad 1 \leqslant k \leqslant K, \tag{17}$$

где z_0 — раз и навсегда выбранная точка из $\mathcal{X}$, а $\gamma_k > 0$ — детерминированные скалярные множители.

Анализ сходимости. Приводимый ниже результат очень хорошо известен; для полноты изложения приводим стандартное доказательство этого утверждения в Приложении.

Предложение 2. В предположениях **A.1**–**A.3** *при выборе длины шага*

$$\gamma_k = [\varkappa(k+1)]^{-1}, \quad k = 1, 2, \dots \tag{18}$$

для последовательности оценок $\widehat{x}_k(\omega^k) = z_k$ *любого сигнала* $x \in \mathcal{X}$*, даваемой процедурой СА* (17) *с* $\omega^k = (\eta_k, y_k)$*, определенными в* (8) *для любого* k*, выполняется следующая оценка:*

$$\mathbf{E}_{\omega^k \sim P_x^k} \left\{ \|\widehat{x}_k(\omega^k) - x\|_2^2 \right\} \leqslant \frac{4M^2}{\varkappa^2(k+1)}, \quad k = 0, 1, \dots, \tag{19}$$

где P_x *— распределение пары* (η, y)*, обусловленное сигналом* x*.*

3.2. Численный пример

Для иллюстрации изложенных выше построений приведем результаты некоторых численных экспериментов. Для наглядности рассматриваем простейшие ситуации, а именно:

- $\mathcal{X} = \{x \in \mathbf{R}^n : \|x\|_2 \leqslant R\}$;
- распределение Q регрессора η есть $\mathcal{N}(0, I_n)$;
- f — монотонное векторное поле на $\mathbf{R}$, задаваемое одним из следующих четырех способов:
  - A. $f(s) = \exp\{s\}/(1 + \exp\{s\})$;
  - B. $f(s) = s$;
  - C. $f(s) = \max[s, 0]$;
  - D. $f(s) = \min[1, \max[s, 0]]$.
- условное по η распределение y, индуцированное распределением P_x, есть
  - — распределение Бернулли с вероятностью $f(\eta^{\mathrm{T}} x)$ исхода единица в случае A (т.е. этот случай соответствует логистической модели),
  - — гауссовское распределение $\mathcal{N}(f(\eta^{\mathrm{T}} x), I_n)$ в случаях B–D.

Обратим внимание, что в рассматриваемом примере поле $F(z)$ легко вычисляется. Действительно, для всех $z \in \mathbf{R}^n$ имеем

$$\eta = \frac{zz^{\mathrm{T}}}{\|z\|_2^2}\eta + \underbrace{\left(I_n - \frac{zz^{\mathrm{T}}}{\|z\|_2^2}\right)\eta}_{\eta_\perp}$$

и в силу независимости $\eta^{\mathrm{T}} z$ и $\eta_\perp$ получаем

$$F(z) = \mathbf{E}_{\eta \sim \mathcal{N}(0,I)}\{\eta f(\eta^{\mathrm{T}} z)\} = \mathbf{E}_{\eta \sim \mathcal{N}(0,I)} \left\{ \frac{zz^{\mathrm{T}}\eta}{\|z\|_2^2} f(\eta^{\mathrm{T}} z) \right\} =$$
$$= \frac{z}{\|z\|_2} \mathbf{E}_{\zeta \sim \mathcal{N}(0,1)}\{\zeta f(\|z\|_2 \zeta)\},$$

т.е. $F(z)$ пропорционально $z/\|z\|_2$ с коэффициентом

$$h(\|z\|_2) = \mathbf{E}_{\zeta \sim \mathcal{N}(0,1)}\{\zeta f(\|z\|_2 \zeta)\}.$$

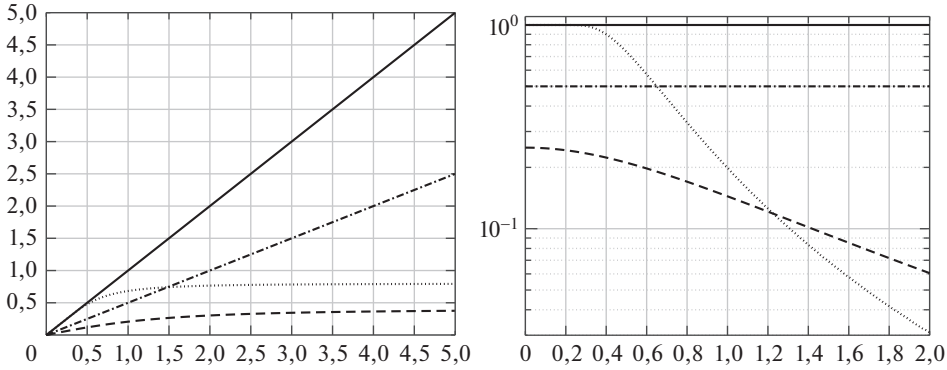

Рис. 1. Слева: функции h; справа: модули сильной монотонности операторов $F(\cdot)$ на $\mathcal{X} = \{z : \|z\|_2 \leqslant R\}$ как функции R. Пунктирные линии — случай A, сплошные линии — случай B, штрих-пунктирные линии — случай C, точечные линии — случай D.

На рис. 1 представлены графики функции $h(t)$ для случаев A–D, а также зависимости модулей сильной монотонности соответствующих отображений F на шаре $\mathcal{X}$ от радиуса R шара. Размерность n во всех экспериментах полагалась равной 100, а количество K наблюдений полагалось 400, 10^3, $4\cdot10^3$, 10^4 и $4\cdot10^4$. Для каждой комбинации параметров проводилось 10 моделирований сигнала x, порождающего наблюдения (8), генерируемого случайно равномерно на единичной сфере (границе множества $\mathcal{X}$).

В каждом эксперименте вычислялись BCA-оценки и CA-оценки (заметим, что в случаях A и B оценки, даваемые BCA, совпадают с МП-оценками). Длина γ_k шага в CA выбиралась в соответствии с (18) при "эмпирическом" выборе величины $\varkappa$[7]. Точнее, полученные наблюдения $\omega_k = (\eta_k, y_k)$, $k \leqslant K$ (8) использовались для построения CA-оценок в два этапа:
— на *этапе настройки* генерировался случайный "обучающий сигнал" $x' \in \mathcal{X}$, после чего генерировались метки y'_k так, как если бы x' являлся истинным сигналом. Например, в случае A метке y'_k придавалось значение единица с вероятностью $q = f(\eta_k^{\mathrm{T}} x')$ и нуль с вероятностью $1 - q$. После того как сгенерированы "обучающий сигнал" и соответствующие метки, к полученным искусственным наблюдениям применялась схема CA с различными значениями $\varkappa$, вычислялась точность полученных оценок и выбиралось то значение $\varkappa$, которое приводит к наилучшему восстановлению сигнала;
— на *этапе исполнения* алгоритм CA прогонялся на фактических данных с шагом (18), определяемым тем значением $\varkappa$, которое было найдено на этапе настройки.

Результаты некоторых численных экспериментов представлены на рис. 2.

Подчеркнем, что время вычислений включает в себя оба этапа. Вывод из результатов экспериментов таков: будучи лишь немногим лучше CA по ка-

[7] Для величины модуля сильной монотонности векторных полей $F(\cdot)$ могут быть аналитически получены нижние границы, однако это требует немалых усилий, а границы оказываются консервативными.

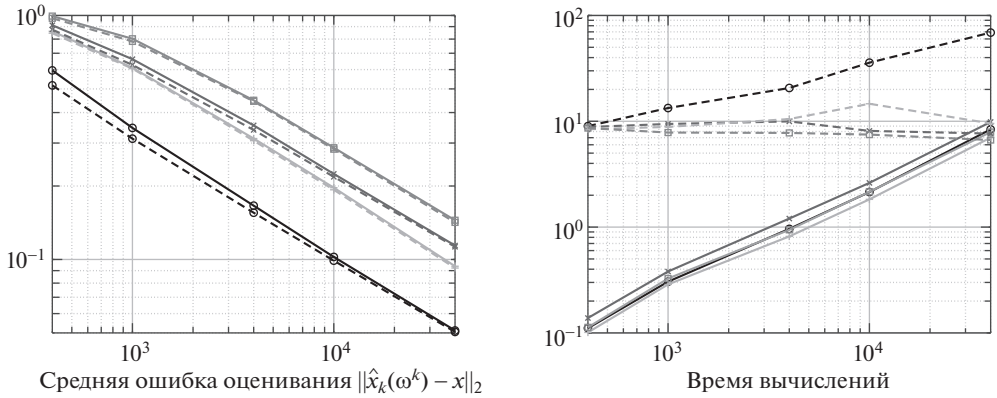

Рис. 2. Численные результаты: средние ошибки и время вычислений для CA-оценок (сплошная линия) и BCA-оценок (пунктирная линия). o — случай A, × — случай B, + — случай C, □ — случай D.

честву оценивания, BCA значительно проигрывает ему по времени исполнения. Заметим также, что наблюдаемая в экспериментах зависимость ошибок восстановления от размера K выборки соответствует скорости сходимости $O(1/\sqrt{K})$, установленной в предложении 2.

4. Случай "одного наблюдения"

Исследуем один специальный случай задачи оценивания, в котором последовательность $\eta_1, \ldots, \eta_K$ регрессоров в (8) детерминированная. На первый взгляд, такая формулировка не укладывается в рамки рассматриваемой здесь общей постановки, в которой регрессоры предполагаются случайными независимыми, одинаково распределенными, выбираемыми из некоторого распределения Q. Однако имеется возможность обойти это "противоречие", считая, что находимся в ситуации с *одним наблюдением*, где регрессором является матрица $[\eta_1, \ldots, \eta_K]$, а Q — вырожденное одноточечное распределение. Конкретно, пусть имеем наблюдение

$$\omega = (\eta, y) \in \mathbf{R}^{n \times mK} \times \mathbf{R}^{mK} \tag{20}$$

(m, n, K — известные натуральные числа), а распределение P_x наблюдений, порожденных сигналом $x \in \mathbf{R}^n$, имеет следующую форму:

- η — заданная детерминированная матрица, не зависящая от x;
- метка y случайна, и распределение y, индуцированное распределением P_x, имеет среднее $\phi(\eta^{\mathrm{T}} x)$, где $\phi : \mathbf{R}^{mK} \to \mathbf{R}^{mK}$ — некоторое заданное отображение.

В качестве информативного примера, связывающего рассматриваемую постановку с предыдущей, приведем конструкцию, в которой $\eta = [\eta_1, \ldots, \eta_K]$ с $n \times m$ детерминированными "индивидуальными регрессорами" η_k, $y = [y_1; \ldots; y_K]$, имеющими "индивидуальные метки" $y_k \in \mathbf{R}^m$, которые условно (при данном x) независимы по k и такими, что математические ожидания y_k,

индуцированные сигналом x, имеют вид $f(\eta_k^{\mathrm{T}} x)$ для некоторой $f : \mathbf{R}^m \to \mathbf{R}^m$. Положим $\phi([u_1; \ldots; u_K]) = [f(u_1); \ldots; f(u_K)]$. Полученная модель "одного наблюдения" представляет собой естественный аналог модели K наблюдений, которая и рассматривалась до сих пор. Единственная разница состоит в том, что индивидуальные регрессоры теперь образуют детерминированную последовательность, не являясь выборкой некоторой случайной матрицы.

Как и повсюду, в настоящей работе цель — использовать наблюдения (20) для восстановления неизвестного сигнала x, порождающего, как объяснено выше, распределение наблюдения. Формально находимся в ситуации $K = 1$ в исходной задаче восстановления, в которой носителем распределения Q является точка $\{\eta\}$, поэтому можем пользоваться всеми предложенными ранее конструкциями. Конкретно:

- векторное поле $F(z)$, соответствующее задаче, имеет вид

$$F(z) = \eta\phi(\eta^{\mathrm{T}} z)$$

(ранее было $\mathbf{E}_{\eta\sim Q}\{\eta f(\eta^{\mathrm{T}} z)\}$), а векторное поле $G(z) = F(z) - F(x)$, где x — сигнал, породивший наблюдение (20), имеет вид

$$G(z) = \mathbf{E}_{(\eta,y)\sim P_x}\{F(z) - \eta y\}$$

(ср. с (14)). Как и ранее, восстанавливаемый сигнал является нулем этого поля $G(\cdot)$. Обратим внимание на то, что теперь векторное поле $F(z)$ наблюдаемо, а векторное поле G по-прежнему является математическим ожиданием (по P_x) наблюдаемого векторного поля:

$$G(z) = \mathbf{E}_{(\eta,y)\sim P_x}\{\underbrace{\eta\phi(\eta^{\mathrm{T}} z) - \eta y}_{G_y(z)}\},$$

ср. с леммой 1.

- предположения **A.1**–**A.2** теперь принимают следующий вид.

A.1′. Векторное поле $\phi(\cdot) : \mathbf{R}^{mK} \to \mathbf{R}^{mK}$ непрерывно и монотонно, так что $F(\cdot)$ также непрерывно и монотонно.

A.2′. Множество $\mathcal{X}$ непусто, компактно и выпукло, а поле F сильно монотонно на $\mathcal{X}$ с модулем $\varkappa > 0$.

Простым достаточным условием выполнения приведенных выше условий монотонности является положительная определенность матрицы $\eta\eta^{\mathrm{T}}$ в совокупности с сильной монотонностью ϕ на любом ограниченном множестве.

- Предположение **A.3** удобно переписать в следующей эквивалентной форме:

A.3′. Для надлежащим образом выбранного $\sigma \geqslant 0$ и всякого $x \in \mathcal{X}$ справедливо

$$\mathbf{E}_{(\eta,y)\sim P_x}\left\{\|\eta[y - \phi(\eta^{\mathrm{T}} x)]\|_2^2\right\} \leqslant \sigma^2.$$

Теперь в новой постановке ВСА-оценка $\widehat{x}(y)$ является единственным слабым решением ВН $\mathrm{VI}(G_y, \mathcal{X})$, а качество этой оценки дается следующим утверждением.

Предложение 3. Пусть в рассматриваемой ситуации выполнены предположения **A.1′–A3′**. *Тогда для всякого* $x \in \mathcal{X}$ *и любой реализации соответствующих наблюдений* (η, y) (20) *справедливо*

$$\|\widehat{x}(y) - x\|_2 \leqslant \varkappa^{-1} \|\underbrace{\eta[y - \phi(\eta^{\mathrm{T}} x)]}_{\Delta(x,y)}\|_2, \tag{21}$$

откуда также следует

$$\mathbf{E}_{(\eta,y)\sim P_x}\{\|\widehat{x}(y) - x\|_2^2\} \leqslant \sigma^2/\varkappa^2. \tag{22}$$

Доказательство. Пусть $x \in \mathcal{X}$ — сигнал, породивший наблюдения (20) и пусть $G(z) = F(z) - F(x)$ — соответствующее векторное поле G. Имеем

$$\begin{aligned} G_y(z) = F(z) - \eta y = F(z) - F(x) + [F(x) - \eta y] = \\ = G(z) - \eta[y - \phi(\eta^{\mathrm{T}} x)] = G(z) - \Delta(x, y). \end{aligned}$$

При фиксированном y точка $\bar{z} = \widehat{x}(y)$ является слабым, а, следовательно, и сильным (так как $G_y(\cdot)$ непрерывно) решением ВН $\mathrm{VI}(G_y, \mathcal{X})$, что в силу $x \in \mathcal{X}$, влечет

$$0 \leqslant G_y^{\mathrm{T}}(\bar{z})[x - \bar{z}] = G^{\mathrm{T}}(\bar{z})[x - \bar{z}] - \Delta^{\mathrm{T}}(x, y)[x - \bar{z}],$$

откуда

$$-G^{\mathrm{T}}(\bar{z})[x - \bar{z}] \leqslant -\Delta^{\mathrm{T}}(x, y)[x - \bar{z}].$$

Кроме того, $G(x) = 0$, откуда $G^{\mathrm{T}}(x)[x - \bar{z}] = 0$, так что приходим к

$$[G(x) - G(\bar{z})]^{\mathrm{T}}[x - \bar{z}] \leqslant -\Delta^{\mathrm{T}}(x, y)[x - \bar{z}],$$

откуда также следует

$$\varkappa\|x - \bar{z}\|_2^2 \leqslant -\Delta^{\mathrm{T}}(x, y)[x - \bar{z}],$$

поскольку наряду с F поле G сильно монотонно на $\mathcal{X}$ с модулем $\varkappa$, а $x, \bar{z} \in \mathcal{X}$. Применяя неравенство Коши, приходим к (21).

Пример 1. Пусть $m = 1$, ϕ сильно монотонна на всем $\mathbf{R}^K$ с модулем $\varkappa_\phi > 0$, а η в (20) выбирается случайно из "гауссовского ансамбля", т.е. столбцы η_k матрицы η размера $n \times K$ — независимые случайные векторы с распределением $\mathcal{N}(0, I_n)$. Пусть шум в наблюдениях также гауссовский:

$$y = \phi(\eta^{\mathrm{T}} x) + \lambda\xi, \;\; \xi \sim \mathcal{N}(0, I_K).$$

Хорошо известно, что при $K/n \to \infty$ минимальное сингулярное значение $(n \times n)$-матрицы $\eta\eta^{\mathrm{T}}$ с большой вероятностью имеет порядок по крайней

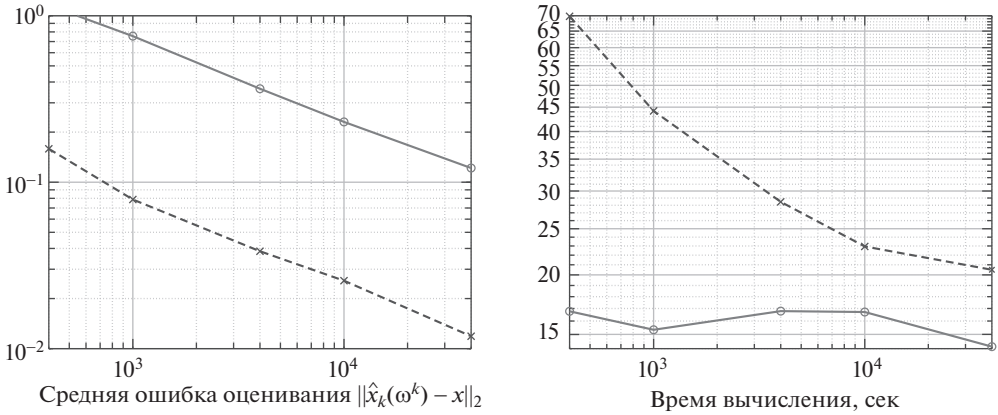

Рис. 3. Средние ошибки и время вычисления для стандартного уклонения $\lambda = 1$ (пунктир) и $\lambda = 0{,}1$ (сплошная линия).

мере $O(1)K$, откуда следует, что при $K/n \gg 1$ для величины модуля сильной монотонности поля $F(\cdot)$, "как правило", верно $\varkappa \geqslant O(1)K\varkappa_\phi$. Более того, в ситуации при $K/n \to \infty$ фробениусова норма матрицы η с большой вероятностью имеет порядок не более $O(1)\sqrt{nK}$. Иными словами, для больших значений K/n задача восстановления по описанному выше ансамблю, "как правило", удовлетворяет предположению 3 с $\varkappa = O(1)K\varkappa_\phi$ и $\sigma^2 = O(\lambda^2 nK)$. В результате (22) дает

$$\mathbf{E}_{(\eta,y)\sim P_x}\left\{\|\widehat{x}(y) - x\|_2^2\right\} \leqslant O(1)\frac{\lambda^2 n}{\varkappa_\phi^2 K}, \qquad [K \gg n].$$

Известно, что для стандартного случая линейной регрессии, где $\phi(x) = \varkappa_\phi x$, получающаяся граница почти оптимальна при условии, что $\mathcal{X}$ достаточно велико.

Численная иллюстрация. В условиях рассмотренного выше примера положим $m = 1$, $n = 100$ и примем

$$\phi(u) = \arctan[u] := [\arctan(u_1); \ldots; \arctan(u_K)] : \ \mathbf{R}^K \to \mathbf{R}^K.$$

Множество $\mathcal{X}$ — единичный шар $\{x \in \mathbf{R}^n : \|x\|_2 \leqslant 1\}$. В конкретном эксперименте η выбиралась случайно из гауссовского ансамбля, как описано выше, сигнал $x \in \mathcal{X}$, порождающий наблюдения (20), выбирался случайно, помеха $y - \phi(\eta^{\mathrm{T}}x)$ в наблюдении имеет распределение $\mathcal{N}(0, \lambda^2 I_K)$. Результаты 10 экспериментов для каждой комбинации размера выборки и дисперсии λ^2 шума представлены на рис. 3.

5. Заключение

Рассмотрен подход к восстановлению сигналов в обобщенных линейных моделях, основанный на сведении к решению стохастических вариационных

неравенств. Рассмотрены условия, при которых соответствующие ВН сильно монотонны. Показано, что в этом случае полученные оценки допускают верхнюю границу на конечном времени для ожидаемой ошибки в ℓ_2-норме, сходящуюся к нулю со скоростью $O(1/K)$ с ростом числа K наблюдений.

ПРИЛОЖЕНИЕ

Доказательство предложения 2. Во-первых, заметим, что z_k являются детерминированными функциями $z_k = Z_k(\omega^k)$ "начальных фрагментов" $\omega^k = \{\omega_t, 1 \leqslant t \leqslant k\} \sim \underbrace{P_x \times \ldots \times P_x}_{P_x^k}$ последовательности наблюдений $\omega^K = \{\omega_k = (\eta_k, y_k), 1 \leqslant k \leqslant K\}$. Введем

$$D_k(\omega^k) = \frac{1}{2}\|Z_k(\omega^k) - x\|_2^2 = \frac{1}{2}\|z_k - x\|_2^2, \quad d_k = \mathbf{E}_{\omega^k \sim P_x^k}\{D_k(\omega^k)\},$$

где $x \in \mathcal{X}$ — сигнал, породивший наблюдения (8). Напомним, что метрическая проекция на замкнутое выпуклое множество $\mathcal{X}$ является сжимающим отображением:

$$\forall (z \in \mathbf{R}^n, u \in \mathcal{X}): \ \|\mathrm{Proj}_{\mathcal{X}}[z] - u\|_2 \leqslant \|z - u\|_2.$$

Соответственно, при $1 \leqslant k \leqslant K$ имеем

$$\begin{aligned} D_k(\omega^k) &= \frac{1}{2}\|\mathrm{Proj}_{\mathcal{X}}[z_{k-1} - \gamma_k G_{\omega_k}(z_{k-1}) - x]\|_2^2 \leqslant \\ &\leqslant \frac{1}{2}\|z_{k-1} - \gamma_k G_{\omega_k}(z_{k-1}) - x\|_2^2 = \\ &= \frac{1}{2}\|z_{k-1} - x\|_2^2 - \gamma_k G_{\omega_k}^{\mathrm{T}}(z_{k-1})(z_{k-1} - x) + \frac{1}{2}\gamma_k^2\|G_{\omega_k}(z_{k-1})\|_2^2. \end{aligned}$$

Беря математическое ожидание относительно $\omega^k \sim P_x^k$ от обеих частей полученного неравенства и учитывая соотношения (13) вместе с тем фактом, что $z_{k-1} \in \mathcal{X}$, получаем

$$d_k \leqslant d_{k-1} - \gamma_k \mathbf{E}_{\omega^{k-1} \sim P_x^{k-1}}\left\{G(z_{k-1})^{\mathrm{T}}(z_{k-1} - x)\right\} + 2\gamma_k^2 M^2. \tag{П.1}$$

Поскольку в данном случае поле G сильно монотонно на $\mathcal{X}$ с модулем $\varkappa > 0$, x является слабым решением ВН $\mathrm{VI}(G, \mathcal{X})$, а z_{k-1} принимает значения в $\mathcal{X}$, то, привлекая (16), получаем, что величина математического ожидания в (П.1) равна по крайней мере $2\varkappa d_k$, и приходим к соотношению

$$d_k \leqslant (1 - 2\varkappa\gamma_k)d_{k-1} + 2\gamma_k^2 M^2. \tag{П.2}$$

Положим

$$S = \frac{2M^2}{\varkappa^2}, \quad \gamma_k = \frac{\varkappa S}{4M^2(k+1)} = \frac{1}{\varkappa(k+1)};$$

заметим, что γ_k есть в точности длины шагов в (18). Проверим по индукции по k, что соотношение

$$d_k \leqslant (k+1)^{-1} S \qquad (*_k)$$

выполнено для $k = 0, 1, \ldots, K$.

Основание индукции $k = 0$**.** Через D обозначим $\|\cdot\|_2$-диаметр множества $\mathcal{X}$, и пусть $z_\pm \in \mathcal{Z}$ таково, что $\|z_+ - z_-\|_2 = D$. Согласно (13) имеем $\|F(z)\|_2 \leqslant M$ для всех $z \in \mathcal{X}$, а сильная монотонность поля $G(\cdot)$ на $\mathcal{X}$ влечет

$$[G(z_+) - G(z_-)]^{\mathrm{T}}[z_+ - z_-] = [F(z_+) - F(z_-)][z_+ - z_-] \geqslant \varkappa \|z_+ - z_-\|_2^2 = \varkappa D^2.$$

По неравенству Коши левая часть последнего неравенства не превосходит $2MD$, поэтому

$$D \leqslant \frac{2M}{\varkappa},$$

откуда $S \geqslant D^2/2$. С другой стороны, по определению d_0 имеем $d_0 \leqslant D^2/2$. Таким образом, $(*_0)$ выполнено.

Шаг индукции $(*_{k-1}) \Rightarrow (*_k)$**.** Предположим теперь, что $(*_{k-1})$ выполнено при некотором k, $1 \leqslant k \leqslant K$, и докажем, что $(*_k)$ также выполнено. Поскольку $\varkappa\gamma_k = (k+1)^{-1} \leqslant 1/2$, то

$$\begin{aligned}
d_k &\leqslant d_{k-1}(1 - 2\varkappa\gamma_k) + 2\gamma_k^2 M^2 \leqslant \qquad [\text{по (П.2)}] \\
&\leqslant \frac{S}{k}(1 - 2\varkappa\gamma_k) + 2\gamma_k^2 M^2 = \qquad [\text{по } (*_{k-1}) \text{ и вследствие } \varkappa\gamma_k \leqslant 1/2] \\
&= \frac{S}{k}\left(1 - \frac{2}{k+1}\right) + \frac{S}{(k+1)^2} = \frac{S}{k+1}\left(\frac{k-1}{k} + \frac{1}{k+1}\right) \leqslant \frac{S}{k+1},
\end{aligned}$$

поэтому $(*_k)$ справедливо. Индукция закончена. Остается заметить, что по определению величины d_k имеем $d_k = \frac{1}{2}\mathbf{E}\{\|\widehat{x}_k - x\|_2^2\}$.

СПИСОК ЛИТЕРАТУРЫ

1. *Devroye L., Györfi L., Lugosi G.* A Probabilistic Theory of Pattern Recognition (Stochastic Modelling and Applied Probability No. 31). N.Y.: Springer Sci. & Business Media, 2013.
2. *Nelder J.A., Wedderburn R.W.M.* Generalized Linear Models // J. Royal Statist. Soc., Ser. A (General). 1972. V. 135. No. 3. P. 370–384.
3. *McCullagh P., Nelder J.A.* Generalized Linear Models. Boca Raton: CRC Press, 1989.
4. *Shapiro A., Dentcheva D., Ruszczyński A.* Lectures on Stochastic Programming: Modeling and Theory. 2nd ed. Philadelphia: SIAM, 2014.
5. *Robbins H., Monro S.* A Stochastic Approximation Method // Ann. Math. Statist. 1951. V. 22. No. 3. P. 400–407.
6. *Wolfowitz J.* On the Stochastic Approximation Method of Robbins and Monro // Ann. Math. Statist. 1952. V. 23. No. 3. P. 457–461.

7. *Barndorff-Nielsen O.* Information and Exponential Families in Statistical Theory. N.Y.: Wiley, 1978.
8. *Feigin P.D.* Conditional Exponential Families and a Representation Theorem for Asympotic Inference // Ann. Statist. 1981. V. 9. No. 3. P. 597–603.
9. *Rosenblatt F.* The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain // Psychol. Rev. 1958. V. 65. No. 6. P. 386–408.
10. *Block H.-D.* The perceptron: A model for brain functioning. I // Rev. Modern Phys. 1962. V. 134. No. 1. P. 123.
11. *Helmbold D.P., Warmuth M.K.* On Weak Learning // J. Comput. Syst. Sci. 1995. V. 50. No. 3. P. 551–573.
12. *Айзерман М.А., Браверман Э.М., Розоноэр Л.И.* Теоретические основы метода потенциальных функций в задаче об обучении автоматов разделению входных ситуаций на классы // АиТ. 1964. Т. 25. № 6. С. 917–936.

    Aizerman M.A., Braverman E.M., Rozonoer L.I. Theoretical Foundations of the Potential Function Method in Pattern Recognition Learning // Autom. Remote Control. 1964. V. 25. No. 6. P. 821–837.
13. *Девятериков И.П., Пропой А.И., Цыпкин Я.З.* О рекуррентных алгоритмах обучения распознаванию образов // АиТ. 1967. № 1. С. 122–132.

    Devyaterikov I.P., Propoi A.I., Tsypkin Ya.Z. Iterative Learning Algorithms for Pattern Recognition // Autom. Remote Control. 1967. V. 28. No. 1. P. 108–117.
14. *Айзерман М.А., Браверман Э.М., Розоноэр Л.И.* Метод потенциальных функций в теории обучения машин. М.: Наука, 1970.
15. *Bousquet O., Boucheron S., Lugosi G.* Introduction to Statistical Learning Theory / Bousquet O., von Luxburg U., Rätsch G. (eds) Advanced Lectures on Machine Learning. ML 2003. Lecture Notes in Computer Science, vol 3176. Springer, Berlin, Heidelberg. P. 169–207.
16. *Sridharan K., Shalev-Shwartz S., Srebro N.* Fast Rates for Regularized Objectives // Advances in Neural Information Proc. Syst. No. 21. 2009. P. 1545–1552.
17. *Nemirovski A., Onn S., Rothblum U.* Accuracy Ccertificates for Computational Problems with Convex Structure // Math. Oper. Res. 2010. V. 35. No. 1. P. 52–78.
18. *Bousquet O., Elisseeff A.* Stability and Generalization // J. Machine Learning Res. 2002. V. 2. P. 499–526.
19. *Rakhlin A., Mukherjee S., Poggio T.* Stability Results in Learning Theory // Anal. Appl. 2005. V. 3. No. 4. P. 397–417.
20. *Shalev-Shwartz S., Shamir O., Srebro N., Sridharan K.* Stochastic Convex Optimization // Conf. Learning Theory. 2009.
21. *Kiefer J., Wolfowitz J.* Stochastic Estimation of the Maximum of a Regression Function // Ann. Mat. Statist. 1952. V. 23. No. 3. P. 462–466.

Статья представлена к публикации членом редколлегии А.В. Назиным.

Поступила в редакцию 19.07.2018

После доработки 12.09.2018

Принята к публикации 08.11.2019