© 2005 г. А.Б. Юдицкий, А.В. Назин [1], А.Б. Цыбаков, Н. Ваятис

РЕКУРРЕНТНОЕ АГРЕГИРОВАНИЕ ОЦЕНОК МЕТОДОМ ЗЕРКАЛЬНОГО СПУСКА С УСРЕДНЕНИЕМ [2]

Рассматривается рекуррентный метод построения агрегированной оценки на конечном классе базовых решающих правил в задаче классификации. Оценка приближенно минимизирует выпуклый функционал риска при ℓ_1-ограничении. Она задается стохастическим вариантом метода зеркального спуска, осуществляющего спуск градиентного типа в двойственном пространстве с дополнительным усреднением. Основной результат настоящей статьи — верхняя граница для средней точности предложенного алгоритма, имеющая порядок $C\sqrt{(\ln M)/t}$, с явным выражением малого постоянного множителя C, где M — размерность задачи, t — число наблюдений. Аналогичная граница получена и для более общей постановки, охватывающей, в частности, модель регрессии при квадратичных потерях.

§ 1. Введение

Методы обобщенного портрета (далее SVM — Support Vector Machines) и бустинга в последнее время широко используются в практике классификации (см., например, [1 – 4]). Эти методы основаны на минимизации выпуклых функционалов эмпирического риска со штрафом. Их статистический анализ дается, например, в работах [5 – 8] (см. также библиографию в них). Отметим, что этот анализ является лишь приближенным, поскольку численные алгоритмы бустинга и SVM не обязательно достигают точного минимума эмпирического риска. Отметим также, что для реализации этих алгоритмов необходимо иметь в наличии всю выборку наблюдений. Вместе с тем, интересна и постановка, в которой наблюдения поступают по одному и требуется настраивать решающие правила рекуррентно, в текущем режиме.

По тематике рекуррентной классификации существует обширная литература, начиная с перцептрона и различных его вариантов (см., например, [9 – 11 и библиографию в них], а также [12, 13]). Упомянем здесь лишь методы, в которых используются те же функции потерь, что в бустинге и SVM, и которые, таким образом, можно рассматривать как аналоги последних, работающие в режиме текущего поступления данных. По-видимому, первым таким примером является метод потенциальных функций, некоторые из вариантов которых представляют собой аналоги SVM в текущем режиме (см.

[1] Исследования проведены во время визитов в Университеты Париж–VI и Гренобль–I (Франция) в 2004–2005 гг.

[2] Работа выполнена в рамках проектов ACI NIM “BIOCLASSIF” и ACI MD “OPSYC”, Франция.

[10, 11], а также [12, гл.10]). Недавно аналоги SVM и методов минимизации эмпирического риска с более общими выпуклыми потерями, работающие в текущем режиме, были предложены в [14]. Укажем также работу [15], в которой для общего класса функций потерь исследуется стохастический градиентный метод с усреднением (ср. с [16]). Во всех этих работах используется обычный стохастический градиентный метод, т.е. спуск осуществляется в исходном пространстве оцениваемых параметров.

В данной статье тоже строятся аналоги бустинга и SVM, работающие в текущем режиме, но по иному принципу: градиентный спуск осуществляется в двойственном пространстве. Алгоритмы такого типа известны для задач детерминированной оптимизации и носят название методов зеркального спуска [17]. Их преимущество по сравнению с обычными градиентными методами состоит в том, что их скорость сходимости зависит лишь логарифмически от размерности задачи. Поэтому они гораздо эффективнее в пространствах большой размерности [18].

Варианты метода зеркального спуска, предложенного в [17] (см. также [19]), были выведены независимо и применялись для задачи классификации и для общей задачи обучения в работах [20 – 22], где получены границы для критерия относительного риска. Однако эти результаты сформулированы в детерминированной постановке и не допускают непосредственного обобщения на случай стандартной стохастической постановки с критерием среднего риска (для понимания взаимосвязи результатов этих двух типов см. [20, 23, 24]). Ниже предлагается новый вариант метода зеркального спуска, достигающий оптимальных границ точности по среднему риску. Основным его отличием от известных является дополнительный шаг усреднения результатов итераций.

Цель данной статьи состоит в построении агрегированного решающего правила: вводится фиксированный конечный базовый класс решающих функций и оптимальным образом подбираются веса в их выпуклой или линейной комбинации. Оптимальность весов понимается в смысле минимума выпуклой функции риска при ℓ_1-ограничении на веса. Эта задача агрегирования подобна тем, которые были рассмотрены, например, в [25 и 26] для модели регрессии при квадратичных потерях. Для ее решения ниже предлагается алгоритм типа зеркального спуска с усреднением результатов итераций, работающий в режиме текущего поступления данных. Доказывается, что алгоритм сходится со скоростью порядка $C\sqrt{(\ln M)/t}$ с явным выражением малого постоянного множителя C, где M — размерность задачи, а t — число наблюдений.

Статья построена следующим образом. В § 2 приводится постановка задачи и формулируется основной результат о скорости сходимости, в § 3 описывается алгоритм, а в § 4 приводится доказательство основного результата. В § 5 результат переносится на общие функции потерь и общую задачу оценивания, в § 6 даются заключительные замечания.

§ 2. Постановка задачи и основной результат

Рассматривается задача классификации на два класса. Пусть (X, Y) — пара случайных величин со значениями в $\mathcal{X}\times\{-1, +1\}$, где $\mathcal{X}$ — пространство признаков. Решающее правило $g_f : \mathcal{X} \to \{-1, +1\}$, соответствующее измеримой функции $f : \mathcal{X} \to \mathbb{R}$, определяется как $g_f(x) = 2\mathbb{I}_{[f(x)>0]} - 1$, где $\mathbb{I}_{[\cdot]}$ - индикаторная функция. Стандартной мерой качества решающего правила g_f является его риск, равный вероятности ошибочной

классификации $R(g_f) = \mathbb{P}\{Y \neq g_f(X)\}$. Оптимальное решающее правило определяется как g_{f^*}, где f^* — функция, минимизирующая риск $R(g_f)$ по всем измеримым f. Оно не реализуемо на практике, поскольку для его построения нужно знать распределение (X, Y). Чтобы аппроксимировать g_{f^*}, строят эмпирические решающие правила $\widehat{g}_n$, основанные на выборке $(X_1, Y_1), \ldots, (X_n, Y_n)$, где (X_i, Y_i) — независимые случайные пары, распределенные так же, как (X, Y).

Абстрактный подход к построению эмпирических решающих правил [4, 27] предписывает искать $\widehat{g}_n$ в виде $\widehat{g}_n = g_{\widehat{f}_n}$, где $\widehat{f}_n$ — функция, минимизирующая эмпирический риск (эмпирическую ошибку классификации)

$$R_n(g_f) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{[Y_i \neq g_f(X_i)]} \tag{1}$$

по всем f из некоторого заданного класса функций. Условия статистической оптимальности метода минимизации эмпирической ошибки классификации (1) широко изучались в литературе (см., в частности, [4, 12, 27]). Однако численный поиск этого минимума, как правило, невозможен, поскольку функционал $R_n(g_f)$ не является ни выпуклым, ни непрерывным по f. Это обстоятельство служило до последнего времени причиной некоторого несоответствия между теорией классификации, обосновывавшей оптимальность метода минимизации эмпирической ошибки классификации (1) (см. [4, 12, 27]), и практикой, где применялись совершенно другие подходы, такие как, например, SVM и бустинг, основанные на численной минимизации выпуклых функционалов эмпирического риска, отличных от (1), как отмечено в [28] и [29]. В работах [5, 6, 8] показано, что это несоответствие можно устранить, а именно, предложена схема доказательства того, что многие из таких практических методов имеют малую ошибку классификации. Ключевым фактом, используемым в этих работах, является то, что при очень общих предположениях оптимальное решающее правило g_{f^*} совпадает с g_{f^A} для функции f^A, доставляющей минимум выпуклому функционалу риска, называемого φ-риском и определяемого выражением

$$A(f) = \mathbf{E}\, \varphi(Yf(X)),$$

где $\varphi : \mathbb{R} \to \mathbb{R}_+$ — выпуклая функция потерь, а $\mathbf{E}$ — математическое ожидание. Типичными примерами функций потерь являются функция $\varphi(x) = (1 - x)_+$, используемая в SVM, а также используемые в бустинге экспоненциальные и логит-потери: $\varphi(x) = \exp(-x)$ и $\varphi(x) = \log_2(1 + \exp(-x))$ соответственно.

Таким образом, для поиска эмпирического решающего правила $\widehat{g}_n$, аппроксимирующего оптимальное g_{f^*}, не обязательно минимизировать по f ошибку классификации (1), а можно рассмотреть минимизацию эмпирического φ-риска

$$A_n(f) = \frac{1}{n} \sum_{i=1}^{n} \varphi(Y_i f(X_i)),$$

являющийся несмещенной оценкой $A(f)$. Эта задача минимизации уже гораздо проще и может быть решена стандартными численными методами, так как функционал A_n выпуклый. При добавлении подходящих штрафных функций она приводит к алгоритмам

типа бустинга и SVM. Вместе с тем, для их реализации необходимо иметь в наличии всю выборку $(X_1, Y_1), \ldots, (X_n, Y_n)$, т.е. это *бэтч-методы*.

В настоящей статье рассматривается задача минимизации φ-риска A по конечно-параметрическому классу функций f в случае, когда данные (X_i, Y_i) поступают последовательно (в текущем режиме).

Введем параметрический класс функций, среди которых мы выбираем f. Допустим, что задан конечный набор базовых функций $\{h_1, \ldots, h_M\}$, где $h_j : \mathcal{X} \to [-K, K]$, $j = 1, \ldots, M$, постоянная $K \in (0, \infty)$, и $M \geq 2$. Через H будем обозначать вектор-функцию, компонентами которой являются эти базовые элементы:

$$H(x) = (h_1(x), \ldots, h_M(x))^T \ . \tag{2}$$

Типичным примером служит модель, в которой функции h_j являются решающими правилами, т.е. принимают значения в $\{-1, 1\}$. Далее, для фиксированного $\lambda > 0$ через $\Theta_{M,\lambda}$ обозначим λ-симплекс в $\mathbb{R}^M$:

$$\Theta_{M,\lambda} = \left\{ \theta = (\theta^{(1)}, \ldots, \theta^{(M)})^T \in \mathbb{R}_+^M \ : \ \sum_{i=1}^{M} \theta^{(i)} = \lambda \right\} \ .$$

Введем семейство λ-выпуклых комбинаций функций $h_1, \ldots, h_M$, т.е.

$$\mathcal{F}_{M,\lambda} = \left\{ f_\theta = \theta^T H \ : \ \theta \in \Theta_{M,\lambda} \right\} \ .$$

Это и есть класс функций, по которому мы хотим минимизировать $A(f)$. Минимизация $A(f)$ по всем $f \in \mathcal{F}_{M,\lambda}$ эквивалентна минимизации $A(f_\theta)$ по всем $\theta \in \Theta_{M,\lambda}$, так что мы упростим обозначения и будем далее писать

$$A(\theta) \triangleq A(f_\theta).$$

Определим вектор оптимальных весов λ-выпуклой комбинации базовых функций как решение задачи на минимум

$$\min_{\theta \in \Theta_{M,\lambda}} A(\theta). \tag{3}$$

Предполагается, что распределение (X, Y) не известно, поэтому функция $A(\cdot)$ тоже не известна, и прямая ее минимизация невозможна. Однако имеется обучающая выборка, т.е. набор независимых случайных пар (X_i, Y_i), распределенных так же, как (X, Y), по которой можно построить оценку оптимальных весов.

В § 3 предложен стохастический алгоритм, основанный на принципе зеркального спуска, который выдает на t-й итерации оценку $\widehat{\theta}_t = \widehat{\theta}_t((X_1, Y_1), \ldots, (X_{t-1}, Y_{t-1}))$ решения задачи (3). Оценка $\widehat{\theta}_t$ измерима относительно совокупности $(\widehat{\theta}_{t-1}, X_{t-1}, Y_{t-1})$, т.е. алгоритм применим в режиме текущего поступления данных. Для использования алгоритма достаточно располагать случайными реализациями субградиента A, имеющими вид

$$u_i(\theta) = \varphi'(Y_i \theta^T H(X_i))\, Y_i H(X_i) \in \mathbb{R}^M, \quad i = 1, 2, \ldots, \tag{4}$$

где φ' — произвольная монотонная версия производной от φ (можно взять, например, непрерывную справа версию).

По $\widehat{\theta}_t$ строится λ-выпуклая комбинация базовых функций $\widehat{\theta}_t^T H(\cdot)$, которая определяет агрегированное решающее правило

$$\widetilde{g}_t(x) = 2\mathbb{I}_{[\widehat{\theta}_t^T H(x)>0]} - 1\,.$$

Статистические свойства этого решающего правила описываются следующим результатом о скорости сходимости оценки $\widehat{\theta}_t$ по φ-риску.

Т е о р е м а 1. *Пусть при заданной выпуклой функции потерь φ, фиксированном наборе (2) из $M \geq 2$ базовых элементов и фиксированном $\lambda > 0$ оценка $\widehat{\theta}_t$ определяется алгоритмом из раздела 3.4 (см. ниже). Тогда для любого целого числа $t \geq 1$ имеем*

$$\mathbf{E}\, A(\widehat{\theta}_t) - \min_{\theta\in\Theta_{M,\,\lambda}} A(\theta) \leq C\,\frac{(\ln M)^{1/2}\sqrt{t+1}}{t}\,, \tag{5}$$

где $C = C(\varphi,\lambda) = 2\,\lambda\,L_\varphi(\lambda)$ и $L_\varphi(\lambda) = K \sup\limits_{|x|\leq K\lambda} |\varphi'(x)|$.

Например, теорема 1 справедлива с константой $C = 2$ в типичном случае, когда рассматривается выпуклое ($\lambda = 1$) агрегирование базовых решающих правил h_j, принимающих значения в $\{-1, 1\}$, и берутся потери $\varphi(x) = (1-x)_+$, используемые в SVM. Заметим также, что теорема 1 справедлива при любом распределении пары (X, Y): в ней нет ограничений на это распределение за исключением условия, что Y принимает значение в $\{-1, 1\}$, которое диктуется самой задачей классификации.

Замечание 1 (эффективность). Сходимость со скоростью порядка $\sqrt{(\ln M)/t}$ типична в отсутствии предположений о малом уровне шума (введенных в [30]). Бэтч-методы, основанные на минимизации эмпирического выпуклого функционала, сходятся с такой же скоростью. Поэтому со статистической точки зрения нет сколько-нибудь значительного различия между бэтч-методами и нашим алгоритмом зеркального спуска. С другой стороны, с вычислительной точки зрения наш алгоритм вполне сравним с прямым методом градиентного спуска. Однако алгоритм зеркального спуска имеет два важных преимущества: (i) его ошибка гораздо слабее растет в зависимости от размерности M базового класса, чем у прямого метода градиентного спуска (зависимость порядка $\sqrt{\ln M}$ в теореме 1 вместо M или $\sqrt{M}$ для прямого метода стохастического градиента); (ii) в численном отношении рекуррентный зеркальный спуск более эффективен, чем бэтч-методы, особенно при данных большой размерности, поскольку его алгоритмическая сложность и требуемый объем памяти строго меньше по порядку, чем для соответствующих бэтч-методов (см. сравнение бэтч-методов и методов оценивания в текущем режиме, данное в [25]).

Замечание 2 (оптимальность скорости сходимости). Используя методы из [25 и 26], нетрудно доказать минимаксную нижнюю границу для разности $\mathbf{E}\, A(\widehat{\theta}_t) - \min\limits_{\theta\in\Theta_{M,\,\lambda}} A(\theta)$, имеющую порядок $\sqrt{(\ln M)/t}$, если $M \geq t^{1/2+\delta}$ при некотором $\delta > 0$. Это показывает, что верхняя граница теоремы 1 оптимальна по порядку для таких значений M.

Замечание 3 (выбор базового класса). Хорошее поведение метода существенно зависит от того, как выбран класс базовых функций $\{h_j\}_{1\leq j\leq M}$. Естественно, например, рассмотреть симметричный класс в том смысле, что если элемент h принадлежит классу, то $-h$ также находится в классе. При практической реализации необходимо иметь в распоряжении некоторый начальный набор данных для предварительного выбора M

функций h_j. Другой распространенный на практике способ: выбирать в качестве h_j очень простые и сами по себе неэффективные решающие правила, такие как покоординатные решения (decision stumps, см., например, [3]); тем не менее агрегирование может приводить к хорошему результату, если их число M очень большое. Что касается теории, то чтобы провести полный статистический анализ, необходимо установить границу для ошибки аппроксимации $\inf\limits_{f\in\mathcal{F}_{M,\,\lambda}} A(f) - \inf\limits_{f} A(f)$, зависящую от массивности базового класса, которая определяется как разнообразием (ортогональностью, приближенной ортогональностью или независимостью) функций h_j, так и размерностью M. Например, можно взять в качестве h_j собственные функции некоторого положительно определенного ядра. Соответствующие результаты можно найти в [31] (см. также [7]). Выбор λ мотивируется сходными соображениями. На самом деле, если учитывать ошибку аппроксимации, имеет смысл выбирать число λ, зависящее от объема выборки t и стремящееся к бесконечности с некоторой небольшой скоростью (ср. с [6]). При этом оптимальное λ, по-видимому, должно определяться из условия баланса между стохастической ошибкой, даваемой теоремой 1, и ошибкой аппроксимации. Эти вопросы здесь не рассматриваются, так как данная статья посвящена только задаче агрегирования.

§ 3. Определение и обсуждение алгоритма

В этом параграфе дается определение предлагаемого алгоритма. В его основе лежит идея зеркального спуска, восходящая к монографии [17]. Введем сначала ряд определений и напомним некоторые факты из выпуклого анализа.

3.1. Прокси-функции. Обозначим через $E = \ell_1^M$ пространство $\mathbb{R}^M$, оснащенное 1-нормой

$$\|z\|_1 = \sum_{j=1}^{M} |z^{(j)}|,$$

где $z = (z^{(1)}, \ldots, z^{(M)})^T$, а через $E^* = \ell_\infty^M$ — двойственное пространство, представляющее собой $\mathbb{R}^M$, оснащенное супремум-нормой

$$\|z\|_\infty = \max_{\|\theta\|_1=1} z^T\theta = \max_{1\le j\le M} |z^{(j)}|, \quad \forall\, z \in E^*.$$

Пусть Θ — выпуклое замкнутое подмножество E, и пусть заданы параметр $\beta > 0$ и выпуклая функция $V : \Theta \to \mathbb{R}$. Назовем *$\beta$-сопряженной* функцией к V преобразование типа Лежандра–Фенхеля W_β от произведения βV:

$$\forall\, z \in E^*, \quad W_\beta(z) = \sup_{\theta\in\Theta} \left\{-z^T\theta - \beta V(\theta)\right\}. \tag{6}$$

Введем теперь ключевое предположение (условие Липшица в сопряженных нормах $\|\cdot\|_1$ и $\|\cdot\|_\infty$), которое будет использовано при доказательстве теоремы 1.

Предположение (L). *Выпуклая функция $V : \Theta \to \mathbb{R}$ такова, что ее β-сопряженная W_β непрерывно дифференцируема на E^* с градиентом ∇W_β, удовлетворяющим*

неравенству

$$\|\nabla W_\beta(z) - \nabla W_\beta(\tilde{z})\|_1 \le \frac{1}{\alpha\beta}\|z - \tilde{z}\|_\infty, \quad \forall\, z, \tilde{z} \in E^*,\ \beta > 0,$$

где α — положительная постоянная, не зависящая от β.

Как известно (см., например, [19, 32]), это предположение связано со следующим понятием сильной выпуклости функции V относительно нормы $\|\cdot\|_1$.

О п р е д е л е н и е 1. *Пусть $\alpha > 0$. Выпуклая функция $V : \Theta \to \mathbb{R}$ называется* α-сильно выпуклой *относительно нормы $\|\cdot\|_1$, если*

$$V(sx + (1-s)y) \le sV(x) + (1-s)V(y) - \frac{\alpha}{2}s(1-s)\|x - y\|_1^2 \tag{7}$$

при любых $x, y \in \Theta$ и $s \in [0, 1]$.

В следующем предложении приводятся свойства β-сопряженных функций, и в частности, дается достаточное условие для справедливости предположения (L).

П р е д л о ж е н и е 1. *Пусть функция $V : \Theta \to \mathbb{R}$ выпукла, а параметр β положителен. Тогда β-сопряженная к V функция W_β обладает следующими свойствами:*

1. *Функция $W_\beta : E^* \to \mathbb{R}$ выпукла и имеет сопряженную βV, т.е.*

$$\forall\, \theta \in \Theta, \quad \beta V(\theta) = \sup_{z \in E^*} \left\{-z^T\theta - W_\beta(z)\right\}.$$

2. *Если функция V является α-сильно выпуклой относительно нормы $\|\cdot\|_1$, то*

   (i) *выполнено предположение (L),*

   (ii) $\operatorname{argmax}_{\theta \in \Theta} \left\{-z^T\theta - \beta V(\theta)\right\} = -\nabla W_\beta(z) \in \Theta$.

Доказательство этого предложения дано в [19, 32].

О п р е д е л е н и е 2. *Назовем функцию $V : \Theta \to \mathbb{R}_+$* прокси-функцией, *если она выпукла и*

(i) *существует такая точка $\theta_* \in \Theta$, что $\min_{\theta \in \Theta} V(\theta) = V(\theta_*)$,*

(ii) *выполнено предположение (L).*

Пример. Пусть $\Theta = \Theta_{M,\lambda}$. Рассмотрим прокси-функцию энтропийного типа

$$\forall\, \theta \in \Theta_{M,\lambda}, \quad V(\theta) = \lambda \ln\left(\frac{M}{\lambda}\right) + \sum_{j=1}^{M} \theta^{(j)} \ln \theta^{(j)} \tag{8}$$

(где $0 \ln 0 \triangleq 0$), имеющую единственную точку минимума $\theta_* = (\lambda/M, \ldots, \lambda/M)^T$, причем $V(\theta_*) = 0$. Нетрудно проверить, что эта функция является α-сильно выпуклой относительно нормы $\|\cdot\|_1$ с параметром $\alpha = 1/\lambda$. Значит, в силу предложения 1 выполняется предположение (L) (его прямое доказательство приводится в приложении).

Важная особенность прокси-функции (8) состоит в том, что для нее задача оптимизации в (6) может быть решена явно. При этом $W_\beta(z)$ и $\nabla W_\beta(z)$ задаются следующими формулами:

$$\forall\, z \in E^*, \quad W_\beta(z) = \lambda\beta \ln\left(\frac{1}{M}\sum_{k=1}^{M} e^{-z^{(k)}/\beta}\right), \tag{9}$$

$$\frac{\partial W_\beta(z)}{\partial z^{(j)}} = -\lambda e^{-z^{(j)}/\beta}\left(\sum_{k=1}^{M} e^{-z^{(k)}/\beta}\right)^{-1}, \; j = 1, \ldots, M. \tag{10}$$

Заметим, что при $\lambda = 1$ верно следующее:

– прокси-функция (8) равна информационному расхождению Кульбака между равномерным распределением на множестве $\{1, \ldots, M\}$ и распределением на этом множестве, заданным вероятностями $\theta^{(j)}$, $j = 1, \ldots, M$;

– в силу (10), компоненты вектора $-\nabla W_\beta(z)$ задают на координатах вектора z распределение Гиббса, в котором β играет роль параметра температуры.

3.2. Алгоритм. Алгоритмы зеркального спуска суть рекуррентные процедуры оптимизации, осуществляющие градиентный спуск в двойственном пространстве. Предлагаемый алгоритм принадлежит к их числу, с той особенностью, что в нем используются стохастические субградиенты и производится усреднение результатов итераций. На каждой итерации i наблюдается новая пара данных (X_i, Y_i) и пересчитываются две переменные: одна — значение ζ_i, определяемое стохастическими субградиентами $u_k(\theta_{k-1})$, $k = 1, \ldots, i$, как результат спуска в двойственном пространстве E^*, вторая — параметр θ_i, который представляет собой "зеркальное изображение" ζ_i в исходном пространстве. Чтобы должным образом настроить алгоритм, необходимо задать две положительные последовательности: $(\gamma_i)_{i\geq 1}$ (размер шага) и $(\beta_i)_{i\geq 1}$ ("температура"), причем $\beta_i \geq \beta_{i-1}$, $\forall\, i \geq 1$. Алгоритм определяется следующим образом:

- Фиксируются начальные значения $\theta_0 \in \Theta$ и $\zeta_0 = 0 \in \mathbb{R}^M$.
- Для $i = 1, \ldots, t-1$ выполняется рекуррентный пересчет

$$\begin{aligned} \zeta_i &= \zeta_{i-1} + \gamma_i u_i(\theta_{i-1}), \\ \theta_i &= -\nabla W_{\beta_i}(\zeta_i)\,. \end{aligned} \tag{11}$$

- Выходом t-й итерации является выпуклая комбинация

$$\widehat{\theta}_t = \frac{\sum\limits_{i=1}^{t} \gamma_i \theta_{i-1}}{\sum\limits_{i=1}^{t} \gamma_i}\,. \tag{12}$$

Заметим, что компоненты $\theta_i^{(j)}$ вектора θ_i из (11) можно записать в виде

$$\theta_i^{(j)} = \frac{\lambda \exp\left(-\beta_i^{-1} \sum_{m=1}^{i} \gamma_m u_{m,j}(\theta_{m-1})\right)}{\sum_{k=1}^{M} \exp\left(-\beta_i^{-1} \sum_{m=1}^{i} \gamma_m u_{m,k}(\theta_{m-1})\right)},$$

где $u_{m,j}(\theta)$ — j-я компонента вектора $u_m(\theta)$, $j = 1, \ldots, M$.

3.3. Эвристические соображения. Допустим, что мы хотим минимизировать выпуклую функцию $\theta \mapsto A(\theta)$ на выпуклом множестве Θ. Если $\theta_0, \ldots, \theta_{t-1}$ — полученные к итерации t точки поиска, можно построить аффинные приближения ϕ_i функции A, определяемые при $\theta \in \Theta$ следующим образом:

$$\phi_i(\theta) = A(\theta_{i-1}) + (\theta - \theta_{i-1})^T \nabla A(\theta_{i-1}), \quad i = 1, \ldots, t\,.$$

Здесь $\theta \mapsto \nabla A(\theta)$ — вектор-функция, принадлежащая субдифференциалу $A(\cdot)$. Взяв выпуклую комбинацию функций ϕ_i, получим усредненное приближение для $A(\theta)$:

$$\bar{\phi}_t(\theta) = \frac{\sum_{i=1}^{t} \gamma_i \left(A(\theta_{i-1}) + (\theta - \theta_{i-1})^T \nabla A(\theta_{i-1})\right)}{\sum_{i=1}^{t} \gamma_i}.$$

На первый взгляд, естественно выбрать в качестве следующей точки поиска вектор $\theta_t \in \Theta$, минимизирующий приближение $\bar{\phi}_t$, т.е.

$$\theta_t = \operatorname*{argmin}_{\theta \in \Theta} \bar{\phi}_t(\theta) = \operatorname*{argmin}_{\theta \in \Theta} \theta^T \left(\sum_{i=1}^{t} \gamma_i \nabla A(\theta_{i-1})\right).$$

Однако это не приводит к успеху, так как приближения хороши только в окрестности точек поиска $\theta_0, \ldots, \theta_{t-1}$. Поэтому следует изменить критерий, например, добавить к целевой функции некоторый штраф $B_t(\theta, \theta_{t-1})$ для того, чтобы удерживать последующую точку θ_t в окрестности предыдущей θ_{t-1}. Таким образом, выбирается точка

$$\theta_t = \operatorname*{argmin}_{\theta \in \Theta} \left[\theta^T \left(\sum_{i=1}^{t} \gamma_i \nabla A(\theta_{i-1})\right) + B_t(\theta, \theta_{t-1})\right]. \tag{13}$$

Наш алгоритм соответствует специальному виду штрафа $B_t(\theta, \theta_{t-1}) = \beta_t V(\theta)$, где V — прокси-функция. Заметим также, что в нашей задаче вектор-функция $\nabla A(\cdot)$ не известна. Поэтому мы заменяем в (13) ненаблюдаемые градиенты $\nabla A(\theta_{i-1})$ наблюдаемыми стохастическими субградиентами $u_i(\theta_{i-1})$. Это приводит к новому определению t-й точки поиска:

$$\theta_t = \operatorname*{argmin}_{\theta \in \Theta} \left[\theta^T \left(\sum_{i=1}^{t} \gamma_i u_i(\theta_{i-1})\right) + \beta_t V(\theta)\right] = \operatorname*{argmax}_{\theta \in \Theta} \left[-\zeta_t^T \theta - \beta_t V(\theta)\right], \tag{14}$$

где

$$\zeta_t = \sum_{i=1}^{t} \gamma_i u_i(\theta_{i-1})\,.$$

Заметим теперь, что в силу предложения 1 значение (14) совпадает с $-\nabla W_{\beta_t}(\zeta_t)$. Отсюда легко выводится итеративная схема (11).

3.4. Частный случай алгоритма. Частный случай метода зеркального спуска с усреднением, для которого доказана теорема 1, определяется следующим образом: это алгоритм, описанный в разделе 3.2, с прокси-функцией V энтропийного типа, определенной в (8), и с последовательностями $(\gamma_i)_{i\geq 1}$ и $(\beta_i)_{i\geq 1}$ вида

$$\gamma_i \equiv 1\,, \quad \beta_i = \beta_0\sqrt{i+1}\,, \quad i = 1, 2, \ldots, \tag{15}$$

где

$$\beta_0 = L_\varphi(\lambda)(\ln M)^{-1/2}. \tag{16}$$

Таким образом, алгоритм упрощается и реализуется в следующем рекуррентном виде:

$$\zeta_i = \zeta_{i-1} + u_i(\theta_{i-1})\,, \tag{17}$$

$$\theta_i = -\nabla W_{\beta_i}(\zeta_i)\,, \tag{18}$$

$$\widehat{\theta}_i = \widehat{\theta}_{i-1} - \frac{1}{i}\left(\widehat{\theta}_{i-1} - \theta_{i-1}\right), \quad i = 1, 2, \ldots, \tag{19}$$

с начальными значениями $\zeta_0 = 0\,,\ \theta_0 \in \Theta$ и с $(\beta_i)_{i\geq 1}$ из (15), (16).

3.5. Сравнение с другими вариантами метода зеркального спуска. Варианты метода зеркального спуска, предложенные в [17], несколько отличаются от нашей итеративной схемы (11). Один из них, наиболее близкий к (11), подробно исследуется в [19]. Он базируется на рекуррентном соотношении

$$\theta_i = -\nabla W_1\Big(-\nabla V(\theta_{i-1}) + \gamma_i u_i(\theta_{i-1})\Big), \quad i = 1, 2, \ldots, \tag{20}$$

где функция V — сильно выпуклая относительно нормы исходного пространства E (которое не обязательно является пространством ℓ_1^M) и W_1 — обычная сопряженная к V. Если $\Theta = \mathbb{R}^M$ и $V(\theta) = \frac{1}{2}\|\theta\|_2^2$, то (20) совпадает с обычным градиентным методом. Если же $\Theta = \Theta_{M,1}$ — единичный симплекс, а V — энтропийная прокси-функция (8), то компоненты $\theta_i^{(j)}$ вектора θ_i из (20) записываются в следующем явном виде:

$$\theta_i^{(j)} = \frac{\theta_{i-1}^{(j)} \exp\left(-\gamma_i u_{i,\,j}(\theta_{i-1})\right)}{\sum\limits_{k=1}^{M} \theta_{i-1}^{(k)} \exp\left(-\gamma_i u_{i,\,k}(\theta_{i-1})\right)} = \frac{\theta_0^{(j)} \exp\left(-\sum\limits_{m=1}^{i} \gamma_m u_{m,\,j}(\theta_{m-1})\right)}{\sum\limits_{k=1}^{M} \theta_0^{(k)} \exp\left(-\sum\limits_{m=1}^{i} \gamma_m u_{m,\,k}(\theta_{m-1})\right)}\,, \tag{21}$$

$j = 1, \ldots, M$. Алгоритм (21) известен также под названием экспонентного градиентного метода (Exponentiated Gradient (EG) method) [20].

Отличиями алгоритма (20) от нашего являются:

– начальная итеративная схема (11) — иная, чем (20), в частности, она содержит второй настроечный параметр β_i; кроме того, в алгоритме (21) по-другому используется начальное значение θ_0;

– помимо схемы (11) наш алгоритм содержит дополнительный этап усреднения результатов итераций (12).

Свойства сходимости EG метода (21) исследованы в детерминированной постановке в работах [21, 22]. Именно, в них показано, что при некоторых предположениях разность $A_t(\theta_t) - \min\limits_{\theta\in\Theta_{M,1}} A_t(\theta)$ ограничена некоторой постоянной, зависящей от M и t. Если эта постоянная достаточно мала, эти результаты показывают, что EG метод обеспечивает хорошую численную минимизацию эмпирического риска A_t. Однако они ничего не говорят о величине среднего риска. В частности, из них не следует, что средний риск $\mathbf{E}A(\theta_t)$ близок к минимально возможной величине $\min\limits_{\theta\in\Theta_{M,1}} A(\theta)$, как это доказывается нами для предложенного здесь алгоритма с усреднением.

Наконец, отметим, что алгоритм (20) можно вывести из соображений близких к тем, которые приведены в разделе 3.3 и которые подробно обсуждаются в литературе по проксимальным методам выпуклой оптимизации (см., например, [33, 34 и библиографию в них]). А именно, при достаточно общих условиях величина θ_i из (20) есть решение задачи на минимум

$$\theta_i = \operatorname*{argmin}_{\theta\in\Theta} \left(\theta^T\gamma_i u_i(\theta_{i-1}) + B(\theta,\theta_{i-1})\right),$$

где штраф $B(\theta,\theta_{i-1}) = V(\theta) - V(\theta_{i-1}) - (\theta-\theta_{i-1})^T\nabla V(\theta_{i-1})$ — расхождение Брегмана между θ и θ_{i-1}, соответствующее сильно выпуклой функции V.

§ 4. Доказательства

В этом параграфе даются выкладки, приводящие к результату теоремы 1. Они проводятся для более общей постановки, чем в теореме 1. Именно, рассматривается произвольная прокси-функция V и используются обозначения и предположения из раздела 3.2. При этом в доказываемых ниже предложениях 2 и 3 множество Θ — произвольное выпуклое замкнутое подмножество E, последовательности оценок (θ_i) и $(\widehat{\theta}_i)$ порождены алгоритмом (11)–(12), а в доказательстве теоремы 1 все выкладки вплоть до формулы (26) верны в предположении, что Θ выпуклый компакт в E.

Введем обозначения

$$\begin{aligned}\nabla A(\theta) &= \mathbf{E}\, u_i(\theta),\\ \xi_i(\theta) &= u_i(\theta) - \nabla A(\theta), \quad \forall\,\theta\in\Theta,\end{aligned}$$

где $u_i(\theta)$ — случайные векторы, определенные в (4). Заметим, что отображение $\theta \mapsto \mathbf{E}\, u_i(\theta)$ принадлежит субдифференциалу $A(\cdot)$ (что объясняет обозначение ∇A). Этот факт и неравенство $\mathbf{E}\|u_i(\theta)\|_\infty^2 \le L_\varphi^2(\lambda)$, справедливое для всех $\theta\in\Theta$, — единственные свойства u_i, которые будут использованы при доказательстве; другие же особенности определения (4) не имеют значения.

П р е д л о ж е н и е 2. *Для любого $\theta \in \Theta$ и любого целого $t \geq 1$ справедливо неравенство*

$$\sum_{i=1}^{t} \gamma_i (\theta_{i-1} - \theta)^T \nabla A(\theta_{i-1}) \leq$$
$$\leq \beta_t V(\theta) - \beta_0 V(\theta_*) - \sum_{i=1}^{t} \gamma_i (\theta_{i-1} - \theta)^T \xi_i(\theta_{i-1}) + \sum_{i=1}^{t} \frac{\gamma_i^2}{2\alpha\beta_{i-1}} \|u_i(\theta_{i-1})\|_\infty^2 .$$

Д о к а з а т е л ь с т в о. В силу непрерывной дифференцируемости $W_{\beta_{i-1}}$ имеем

$$W_{\beta_{i-1}}(\zeta_i) = W_{\beta_{i-1}}(\zeta_{i-1}) + \int_0^1 (\zeta_i - \zeta_{i-1})^T \nabla W_{\beta_{i-1}}(\tau\zeta_i + (1-\tau)\zeta_{i-1})\, d\tau .$$

Положим $v_i = u_i(\theta_{i-1})$. Тогда $\zeta_i - \zeta_{i-1} = \gamma_i v_i$, и в силу предположения (L)

$$\begin{aligned} W_{\beta_{i-1}}(\zeta_i) &= W_{\beta_{i-1}}(\zeta_{i-1}) + \gamma_i v_i^T \nabla W_{\beta_{i-1}}(\zeta_{i-1}) + \\ &\quad + \gamma_i \int_0^1 v_i^T \Big[\nabla W_{\beta_{i-1}}(\tau\zeta_i + (1-\tau)\zeta_{i-1}) - \nabla W_{\beta_{i-1}}(\zeta_{i-1})\Big] d\tau \leq \\ &\leq W_{\beta_{i-1}}(\zeta_{i-1}) + \gamma_i v_i^T \nabla W_{\beta_{i-1}}(\zeta_{i-1}) + \\ &\quad + \gamma_i \|v_i\|_\infty \int_0^1 \|\nabla W_{\beta_{i-1}}(\tau\zeta_i + (1-\tau)\zeta_{i-1}) - \nabla W_{\beta_{i-1}}(\zeta_{i-1})\|_1 d\tau \leq \\ &\leq W_{\beta_{i-1}}(\zeta_{i-1}) + \gamma_i v_i^T \nabla W_{\beta_{i-1}}(\zeta_{i-1}) + \frac{\gamma_i^2 \|v_i\|_\infty^2}{2\alpha\beta_{i-1}} . \end{aligned}$$

Используя последнее неравенство, тот факт, что $(\beta_i)_{i\geq 1}$ — неубывающая последовательность, и что для фиксированного z отображение $\beta \mapsto W_\beta(z)$ представляет собой невозрастающую функцию, получаем

$$W_{\beta_i}(\zeta_i) \leq W_{\beta_{i-1}}(\zeta_i) \leq W_{\beta_{i-1}}(\zeta_{i-1}) - \gamma_i \theta_{i-1}^T v_i + \frac{\gamma_i^2 \|v_i\|_\infty^2}{2\alpha\beta_{i-1}} .$$

Суммируя, получаем

$$\sum_{i=1}^{t} \gamma_i \theta_{i-1}^T v_i \leq W_{\beta_0}(\zeta_0) - W_{\beta_t}(\zeta_t) + \sum_{i=1}^{t} \frac{\gamma_i^2 \|v_i\|_\infty^2}{2\alpha\beta_{i-1}} .$$

Используя представление $\zeta_t = \sum_{i=1}^{t} \gamma_i v_i$, получим, что для всякого $\theta \in \Theta$

$$\sum_{i=1}^{t} \gamma_i (\theta_{i-1} - \theta)^T v_i \leq W_{\beta_0}(\zeta_0) - W_{\beta_t}(\zeta_t) - \zeta_t^T \theta + \sum_{i=1}^{t} \frac{\gamma_i^2 \|v_i\|_\infty^2}{2\alpha\beta_{i-1}} .$$

Наконец, поскольку $v_i = \nabla A(\theta_{i-1}) + \xi_i(\theta_{i-1})$, находим

$$\sum_{i=1}^{t} \gamma_i (\theta_{i-1} - \theta)^T \nabla A(\theta_{i-1}) \le$$

$$\le W_{\beta_0}(\zeta_0) - W_{\beta_t}(\zeta_t) - \zeta_t^T \theta - \sum_{i=1}^{t} \gamma_i (\theta_{i-1} - \theta)^T \xi_i(\theta_{i-1}) + \sum_{i=1}^{t} \frac{\gamma_i^2 \|v_i\|_\infty^2}{2\alpha\beta_{i-1}} \,.$$

Таким образом, требуемое неравенство вытекает из того, что

$$W_{\beta_0}(\zeta_0) = W_{\beta_0}(0) = \beta_0 \sup_{\theta\in\Theta}\{-V(\theta)\} = -\beta_0 V(\theta_*)$$

и $\beta V(\theta) \ge -W_\beta(\zeta) - \zeta^T\theta$ для всех $\zeta \in \mathbb{R}^M$. ▲

Получим теперь основной результат данного параграфа.

П р е д л о ж е н и е 3. *Для любого целого $t \ge 1$ имеет место неравенство*

$$\mathbf{E}\, A(\widehat{\theta}_t) \le \inf_{\theta\in\Theta} \left[A(\theta) + \frac{\beta_t V(\theta) - \beta_0 V(\theta_*)}{\sum\limits_{i=1}^{t} \gamma_i} \right] + \; L_\varphi^2(\lambda) \left(\sum_{i=1}^{t} \gamma_i \right)^{-1} \sum_{i=1}^{t} \frac{\gamma_i^2}{2\alpha\beta_{i-1}} \,, \qquad (22)$$

и средняя точность оценки $\widehat{\theta}_t$ по φ-риску допускает следующую верхнюю границу:

$$\mathbf{E}\, A(\widehat{\theta}_t) - \min_{\theta\in\Theta} A(\theta) \le \frac{1}{\sum\limits_{i=1}^{t} \gamma_i} \left(\beta_t V(\theta_A^*) - \beta_0 V(\theta_*) \; + \; L_\varphi^2(\lambda) \sum_{i=1}^{t} \frac{\gamma_i^2}{2\alpha\beta_{i-1}} \right) , \qquad (23)$$

где $\theta_A^* \in \operatorname*{Argmin}\limits_{\theta\in\Theta} A(\theta)$.

Д о к а з а т е л ь с т в о. При любом $\theta \in \Theta$, в силу выпуклости $A(\theta)$ получаем

$$\begin{aligned} \mathbf{E}\, A(\widehat{\theta}_t) - A(\theta) \;&\le\; \frac{\sum\limits_{i=1}^{t} \gamma_i \left(\mathbf{E}\, A(\theta_{i-1}) - A(\theta) \right)}{\sum\limits_{i=1}^{t} \gamma_i} \le \\ &\le\; \frac{\sum\limits_{i=1}^{t} \gamma_i \mathbf{E} \left[(\theta_{i-1} - \theta)^T \nabla A(\theta_{i-1}) \right]}{\sum\limits_{i=1}^{t} \gamma_i} \,. \end{aligned} \qquad (24)$$

Взяв условное математическое ожидание относительно θ_{i-1} и используя определение $\xi_i(\theta_{i-1})$ и независимость между θ_{i-1} и (X_i, Y_i), находим $\mathbf{E}\, \xi_i(\theta_{i-1}) = 0$. Теперь объединяем (24) и неравенство из предложения 2, где используем границу $\sup\limits_{\theta\in\Theta} \mathbf{E}\|u_i(\theta)\|_\infty^2 \le L_\varphi^2(\lambda)$. Это приводит к неравенству (22), которое очевидным образом влечет (23). ▲

Заметим, что одновременное (и одинаковое) изменение масштаба последовательностей (β_i) и (γ_i), т.е. умножение их на одну и ту же положительную постоянную, не изменяет полученных в предложениях 2 и 3 верхних оценок; более того, при этом и последовательности оценок (θ_i) и $(\widehat{\theta}_i)$ алгоритма (11)–(12) не изменяются (при $\zeta_0 = 0$).

Доказательство теоремы 1. Имеем $V(\theta_*) = 0$ и

$$V(\theta_A^*) \le \max_{\theta\in\Theta} V(\theta) \triangleq V^*.$$

Используя (23) при $\gamma_i \equiv 1$ и $\beta_i = \beta_0\sqrt{i+1}$ с произвольным $\beta_0 > 0$, получаем

$$\mathbf{E}\, A(\widehat{\theta}_t) - A(\theta_A^*) \le \frac{\sqrt{t+1}}{t}\left(\beta_0 V^* + \frac{L_\varphi^2(\lambda)}{\alpha\beta_0}\right). \tag{25}$$

Минимум этой границы по β_0 достигается при

$$\beta_0 = \frac{L_\varphi(\lambda)}{\sqrt{\alpha V^*}}\ ,$$

что дает оценку

$$\mathbf{E}\, A(\widehat{\theta}_t) - A(\theta_A^*) \le \frac{2L_\varphi(\lambda)}{t}\sqrt{\frac{V^*}{\alpha}\,(t+1)}\,. \tag{26}$$

Пусть теперь $\Theta = \Theta_{M,\lambda}$. Напомним, что параметр сильной выпуклости α прокси-функции V, определенной в (8), равен λ^{-1}. Кроме того, эта прокси-функция достигает своего максимума в каждой вершине λ-симплекса $\Theta_{M,\lambda}$ и такова, что

$$V^* = \max_{\theta\in\Theta_{M,\lambda}} V(\theta) = \lambda \ln M\,.$$

Поэтому оптимальное значение β_0 равно $L_\varphi(\lambda)(\ln M)^{-1/2}$. Это дает границу точности, указанную в формулировке теоремы 1. ▲

§ 5. Обобщение

В приведенном выше доказательстве, по существу, не используется тот факт, что функция потерь и прокси-функция имеют специальный вид. Необходимые для доказательства свойства этих функций подытожены в начале § 4. Следовательно, результат типа теоремы 1 может быть сформулирован при более общих условиях. Сделаем это в настоящем параграфе.

Пусть случайная величина Z принимает значения в некотором множестве $\mathcal{Z}$, Θ — выпуклое замкнутое подмножество $\mathbb{R}^M$, а функция потерь $Q : \Theta \times \mathcal{Z} \to \mathbb{R}_+$ такова, что случайная функция $Q(\cdot\,, Z) : \Theta \to \mathbb{R}_+$ выпукла почти наверное. Определим выпуклую функцию риска $A : \Theta \to \mathbb{R}_+$ следующим образом:

$$A(\theta) = \mathbf{E}\, Q(\theta, Z)\,.$$

Предположим, что обучающая выборка задана в виде последовательности $(Z_1, \ldots, Z_{t-1})$ независимых случайных величин, где все Z_i имеют то же распределение, что и Z. Наша

цель состоит в *критериальной минимизации* $A(\cdot)$ на Θ (см., например, [35]), т.е. мы характеризуем точность оценки $\widehat{\theta}_t = \widehat{\theta}_t(Z_1, \ldots, Z_{t-1}) \in \Theta$ точки минимума A разностью

$$\mathbf{E}\, A(\widehat{\theta}_t) - \min_{\theta\in\Theta} A(\theta)$$

(считаем, что $\min\limits_{\theta\in\Theta} A(\theta)$ достигается). Пусть определены стохастические субградиенты

$$u_i(\theta) = \nabla_\theta\, Q(\theta, Z_i)\,, \quad i = 1, 2, \ldots, \tag{27}$$

— такие измеримые функции на $\Theta\times\mathcal{Z}$, что при каждом $\theta \in \Theta$ математическое ожидание $\mathbf{E}\, u_i(\theta)$ принадлежит субдифференциалу функции $A(\theta)$.

Т е о р е м а 2. *Пусть Θ — выпуклое замкнутое подмножество в $\mathbb{R}^M$, а функция потерь $Q(\cdot,\cdot)$ удовлетворяет приведенным выше условиям, причем*

$$\sup_{\theta\in\Theta} \mathbf{E}\|\nabla_\theta\, Q(\theta, Z)\|_\infty^2 \le L_{\Theta,Q}^2\,, \tag{28}$$

где постоянная $L_{\Theta,Q}$ конечна. Пусть V — прокси-функция на Θ с параметром $\alpha > 0$ из предположения (L)*, и пусть существует $\theta_A^* \in \operatorname{Argmin}\limits_{\theta\in\Theta} A(\theta)$. Тогда для любого целого $t \ge 1$ оценка $\widehat{\theta}_t$, определенная в разделе 3.2 со стохастическими субградиентами (27) и с последовательностями $(\gamma_i)_{i\ge1}$ и $(\beta_i)_{i\ge1}$ из (15) с произвольным $\beta_0 > 0$, удовлетворяет неравенству*

$$\mathbf{E}\, A(\widehat{\theta}_t) - \min_{\theta\in\Theta} A(\theta) \le \left(\beta_0 V(\theta_A^*) + \frac{L_{\Theta,Q}^2}{\alpha\beta_0}\right)\frac{\sqrt{t+1}}{t}\,.$$

Если, кроме того, $\overline{V}$ — постоянная, такая что $V(\theta_A^*) \le \overline{V}$, и $\beta_0 = L_{\Theta,Q}\,(\alpha\,\overline{V}\,)^{-1/2}$, то

$$\mathbf{E}\, A(\widehat{\theta}_t) - \min_{\theta\in\Theta} A(\theta) \le 2\, L_{\Theta,Q}\left(\alpha^{-1}\overline{V}\right)^{1/2}\frac{\sqrt{t+1}}{t}\,. \tag{29}$$

В частности, если Θ — выпуклый компакт, то можно взять $\overline{V} = \max\limits_{\theta\in\Theta} V(\theta)$.

Эта теорема вытекает из доказательств, приведенных в § 4 (ср. с (23), (25) и (26)), где вместо постоянной $L_\varphi(\lambda)$ следует писать теперь $L_{\Theta,Q}$. Она является обобщением теоремы 1 и охватывает различные статистические модели, в том числе и описанную в § 2, где роль Z играет пара случайных величин (X, Y), множества $\Theta = \Theta_{M,\lambda}$ и $\mathcal{Z} = \mathcal{X}\times\{-1, +1\}$, а функция потерь $Q(\theta, Z) = \varphi(Y\theta^T H(X))$. Аналогичным образом теорема 2 применима и к стандартной регрессионной модели с квадратичными потерями, для которой сходный результат был получен в [25] для другого метода; в этом случае $Q(\theta, Z) = (Y - \theta^T H(X))^2$.

Заметим наконец, что константу $L_{\Theta,Q}$ из теоремы 2 можно завысить с целью более простого ее вычисления, если верхнюю грань в (28) брать по более широкому множеству, содержащему Θ, например, в случае $\Theta \subset \{\theta : \|\theta\|_1 \le \lambda\}$ при некотором $\lambda > 0$; аналогично при этом можно поступить и с вычислением постоянной $\overline{V} \ge \max\limits_{\theta\in\Theta} V(\theta)$.

Замечание 4 (зависимые данные). Анализ доказательств показывает, что теорему 2 нетрудно распространить на случай зависимых данных Z_i. Фактически, вместо предположения независимости и одинаковой распределенности величин Z_i достаточно было

бы предположить, что они образуют стационарную последовательность и распределения Z_i и Z совпадают. Тогда теорема 2 остается справедливой, если дополнительно предположить, что условные математические ожидания $\mathbf{E}(\xi_i(\theta_{i-1})|\theta_{i-1}) = 0$ п.н.

§ 6. Заключительные замечания

В заключение обсудим выбор прокси-функции V, множества Θ и последовательностей $(\beta_i)_{i\geq 1}$, $(\gamma_i)_{i\geq 1}$.

6.1. Выбор прокси-функции V. Выбор энтропийной прокси-функции, определенной в (8), — не единственно возможный. Ключевым условием на V является сильная выпуклость в норме $\|\cdot\|_1$, гарантирующая выполнение предположения (L). Можно рассматривать и другие прокси-функции, удовлетворяющие этому условию, например,

$$\forall\, \theta \in \mathbb{R}^M\,, \quad V(\theta) = \frac{1}{2\lambda^2}\|\theta\|_p^2 = \frac{1}{2\lambda^2}\left(\sum_{j=1}^{M}\left(\theta^{(j)}\right)^p\right)^{2/p}\,, \tag{30}$$

где $p = 1 + 1/\ln M$ (см. [19]). В отличие от (8), прокси-функция (30) применима для любого выпуклого замкнутого $\Theta \subseteq \mathbb{R}^M$. В случае симплекса $\Theta_{M,\lambda}$ можно рассмотреть функции вида

$$\forall\, \theta \in \Theta_{M,\lambda}\,, \quad V(\theta) = C_0 + C_1\sum_{j=1}^{M}\left(\theta^{(j)}\right)^{s+1}\,, \quad s = \frac{1}{\ln M}\,, \tag{31}$$

где постоянные $C_0 = -\lambda^2/(\mathrm{e}s(s+1))$ и $C_1 = \lambda^{1-s}/(s(s+1))$ таковы, что $\min\limits_{\theta\in\Theta_{M,\lambda}} V(\theta) = 0$. Легко видеть, что функция (31) α-сильно выпукла в норме $\|\cdot\|_1$. При $\lambda = 1$ значение (31) равно частному случаю f-расхождения Чисара (см. определение в [36]) между равномерным распределением на множестве $\{1,\ldots,M\}$ и распределением на этом множестве, заданным вероятностями $\theta^{(j)}$. Напомним, что при $\lambda = 1$ прокси-функция (8) равна расхождению Кульбака между этими распределениями. По-видимому, можно строить прокси-функции, используя другие подходящие f-расхождения Чисара.

Вместе с тем, при выборе прокси-функции, для которой градиент β-сопряженной ∇W_β не выписывается в явном виде, численная реализация нашего алгоритма может оказаться трудоемкой в вычислительном отношении. Из верхней границы (29) видно, что важной характеристикой V с точки зрения точности алгоритма является отношение $\overline{V}/\alpha$ (в частности, при ограниченном множестве Θ, — отношение $\max\limits_{\theta\in\Theta} V(\theta)/\alpha$). Можно определить оптимальную прокси-функцию V как такую, для которой это отношение минимально. По-видимому, при $\Theta = \Theta_{M,\lambda}$ таковой является функция (8), однако у нас нет строгого доказательства этого факта. Для прокси-функции (8) имеем

$$\frac{1}{\alpha}\max_{\theta\in\Theta_{M,\lambda}} V(\theta) = \lambda^2\ln M\,.$$

Для других прокси-функций это отношение имеет примерно такой же порядок. Например, в [19, лемма 6.1] показано, что прокси-функция (30) удовлетворяет соотношению

$$\frac{1}{\alpha}\max_{\theta\in\Theta_{M,\lambda}} V(\theta) = O(1)\lambda^2\ln M\,.$$

Это соотношение также верно для функции (31). Наконец заметим, что такая широко используемая штрафная функция, как $\|\cdot\|_1$, прокси-функцией в смысле определения 2 не является, так как она не сильно выпуклая относительно $\|\cdot\|_1$. Другая часто используемая штрафная функция $\|\cdot\|_2^2$ сильно выпукла. Однако, как нетрудно установить, ее "показатель качества" крайне плох при больших M: эта функция удовлетворяет соотношению

$$\frac{1}{\alpha}\max_{\theta\in\Theta_{M,\lambda}} V(\theta) = \frac{1}{2}\lambda^2 M\,.$$

6.2. Выбор множества Θ. Теорема 2 верна для любого выпуклого замкнутого множества Θ, содержащегося в $\mathbb{R}^M$. Однако для множеств общего вида, как правило, нельзя вычислить явно градиент ∇W_β и количество вычислений на одну итерацию при реализации нашего алгоритма может стать непомерно большим. Поэтому имеет смысл рассматривать только такие множества Θ, для которых решение $\theta_*(z) = -\nabla W_\beta(z)$ задачи оптимизации (6) может быть легко вычисляемым. Некоторыми примерами таких "простых" множеств являются: (i) λ-симплекс $\Theta_{M,\lambda}$, (ii) полноразмерный симплекс $\left\{\theta\in\mathbb{R}_+^M : \|\theta\|_1\le\lambda\right\}$ и (iii) его симметризованный вариант — гипероктаэдр $\{\theta\in\mathbb{R}^M : \|\theta\|_1\le\lambda\}$.

6.3. Выбор последовательностей (β_i) и (γ_i). Постоянный множитель верхней границы из Теоремы 1 может быть лишь немного уменьшен. Указанные в (15) последовательности (β_i) и (γ_i) близки к оптимальным в смысле значения верхней границы (23). Действительно, заменяя в (23) $V(\theta_A^*)$ на верхнюю грань $V^* = \max\limits_{\theta\in\Theta} V(\theta)$ и минимизируя по (γ_i) и по (β_i), удовлетворяющим условию монотонности $\beta_i\ge\beta_{i-1}$, получим, что минимум доставляют любые (γ_i) и (β_i), не зависящие от i и такие, что $\beta_i/\gamma_i \equiv L_\varphi(\lambda)\sqrt{t/2\alpha V^*}$. Например, можно взять

$$\gamma_i \equiv \frac{1}{\sqrt{t}}\,,\quad \beta_i \equiv \frac{L_\varphi(\lambda)}{\sqrt{2\alpha V^*}}\,, \tag{32}$$

при которых граница (5) теоремы 1 заменяется на лучшую:

$$\mathbf{E}\,A(\widehat{\theta}_t) - \min_{\theta\in\Theta_{M,\lambda}} A(\theta) \le \lambda\, L_\varphi(\lambda)\sqrt{\frac{2\ln M}{t}}\,.$$

Таким образом, можно уменьшить множитель в верхней границе с 2 до $\sqrt{2}$. Однако для этого улучшения нужно знать объем выборки t, что исключает возможность применения алгоритма в текущем режиме поступления данных.

П Р И Л О Ж Е Н И Е

Приведем доказательство того, что предположение (L) справедливо для функции W_β, β-сопряженной к функции V энтропийного типа (8). Очевидно, функция W_β, заданная равенством (9), дважды непрерывно дифференцируема в $E^* = \ell_\infty^M$. Положим

$$L = \sup_{z_1,z_2\in E^*,\, z_1\ne z_2} \frac{\|\nabla W(z_1) - \nabla W(z_2)\|_1}{\|z_1 - z_2\|_\infty} \le \sup_{\|x\|_\infty=1,\,\|y\|_\infty\le 1}\ \sup_{z\in E^*} x^T\nabla^2 W(z)y\,,$$

где матрица вторых производных $\nabla^2 W(z)$ имеет элементы

$$\frac{\partial^2 W(z)}{\partial z_i \partial z_j} = \frac{\lambda}{\beta}\left(\frac{e^{-z_i/\beta}\delta_{ij}}{\sum\limits_k e^{-z_k/\beta}} - \frac{e^{-z_j/\beta}e^{-z_i/\beta}}{\left(\sum\limits_k e^{-z_k/\beta}\right)^2}\right).$$

Здесь δ_{ij} — символ Кронекера. Обозначим $a_i = e^{-z_i/\beta}/\sum\limits_k e^{-z_k/\beta}$ — положительные величины с $\sum\limits_i a_i = 1$. Далее,

$$\begin{aligned}\frac{\beta}{\lambda} x^T \nabla^2 W(z) y &= \sum_i x_i y_i a_i - \sum_i a_i x_i \sum_j a_j y_j = \sum_i x_i a_i \left(y_i - \sum_j a_j y_j\right) = \\ &= \sum_i x_i a_i \sum_{j\neq i} a_j (y_i - y_j) \leq \sum_i \sum_{j\neq i} a_i a_j |y_i - y_j| \,. \end{aligned} \qquad (\text{П.1})$$

Наконец, последняя сумма не превышает 1 для любых $|y_i| \leq 1$ и $a_i \geq 0$, $\sum\limits_i a_i = 1$. Чтобы это увидеть, заметим, что максимум выпуклой (по $y \in R^M$) функции, стоящей в правой части неравенства (П.1), на выпуклом множестве $\{y \in R^M : \|y\|_\infty \leq 1\}$ всегда достигается в экстремальных точках этого множества, т.е. в вершинах гиперкуба $\{y \in R^M : y_i = \pm 1,\ i = 1, \ldots, M\}$. Обозначим какую-нибудь из этих экстремальных точек $y^* = (y_1^*, \ldots, y_M^*)^T$. Разобьем множество индексов $\{1, \ldots, M\}$ на $I_+ = \{i : y_i^* = 1\}$ и $I_- = \{i : y_i^* = -1\}$. Тогда наибольшее значение суммы можно разложить, и мы получаем

$$\frac{\beta}{\lambda} L \leq 2 \sum_{i\in I_+, j\in I_-} a_i a_j + 2 \sum_{j\in I_+, i\in I_-} a_i a_j = 4 \sum_{i\in I_+} a_i \sum_{j\in I_-} a_j = 4 a_+ (1 - a_+) \leq 1 \,,$$

где $a_+ = \sum\limits_{i=I_+} a_i$. Следовательно, $\alpha = 1/\lambda$, что не зависит от β.

СПИСОК ЛИТЕРАТУРЫ

1. *Schapire R.E.* The Strength of Weak Learnability // Machine Learning. 1990. V. 5. № 2. P. 197–227.

2. *Freund Y.* Boosting a Weak Learning Algorithm by Majority // Inform. Comput. 1995. V. 121. № 2. P. 256–285.

3. *Schapire R.E., Freund Y., Bartlett P.L., Lee W.S.* Boosting the Margin: a New Explanation for the Effectiveness of Voting Methods // Ann. Stat. 1998. V. 26. № 5. P. 1651–1686.

4. *Vapnik V.N.* Statistical Learning Theory. New York: Wiley, 1998.

5. *Bartlett P.L, Jordan M.I., McAuliffe J.D.* Convexity, Classification, and Risk Bounds. Technical Report 638. Berkeley: Department of Statistics, U.C., 2003.

6. *Lugosi G., Vayatis N.* On the Bayes-Risk Consistency of Regularized Boosting Methods (With Discussion) // Ann. Stat. 2004. V. 32. № 1. P. 30–55.

7. *Scovel J.C., Steinwart I.* Fast Rates for Support Vector Machines. Technical Report LA-UR-03-9117. Los Alamos: National Laboratory, 2003.

8. *Zhang T.* Statistical Behavior and Consistency of Classification Methods Based on Convex Risk Minimization (With Discussion) // Ann. Stat. 2004. V. 32. № 1. P. 56–85.

9. *Цыпкин Я.З.* Основы теории обучающихся систем. М.: Наука, 1970.

10. *Айзерман М.А., Браверман Э.М., Розоноэр Л.И.* Метод потенциальных функций в теории обучения машин. М.: Наука, 1970.

11. *Aizerman M., Braverman E., Rozonoer L.* Extrapolative Problems in Automatic Control and the Method of Potential Functions // American Mathematical Society Translations. 1970. V. 87. P. 281–303.

12. *Devroye L., Györfi L., Lugosi G.* A Probabilistic Theory of Pattern Recognition. New York – Berlin – Heidelberg: Springer, 1996.

13. *Cesa-Bianchi N., Conconi A., Gentile C.* A Second-Order Perceptron Algorithm // SIAM J. Comput. 2005. V. 34. № 3. P. 640–668.

14. *Kivinen J., Smola A.J., Williamson R.C.* Online Learning with Kernels // IEEE Trans. Signal Proc. 2004. V. 52. № 8. P. 2165–2176.

15. *Zhang T.* Solving Large Scale Linear Prediction Problems Using Stochastic Gradient Descent Algorithms // Proc. 21st Int. Conf. on Machine Learning, ICML'04. Banff, Alberta, Canada. New York: ACM Press, 2004.

16. *Polyak B.T., Juditsky A.B.* Acceleration of Stochastic Approximation by Averaging // SIAM J. Control Optim. 1992. V. 30. № 4. P. 838–855.

17. *Немировский А.С., Юдин Д.Б.* Сложность задач и эффективность методов оптимизации. М.: Наука, 1979.

18. *Ben-Tal A., Margalit T., Nemirovski A.* The Ordered Subsets Mirror Descent Optimization Method with Applications to Tomography // SIAM J. Optim. 2001. V. 12. № 1. P. 79–108.

19. *Ben-Tal A., Nemirovski A.S.* The Conjugate Barrier Mirror Descent Method for Non-Smooth Convex Optimization // MINERVA Optim. Center Report. Haifa: Faculty of Industrial Engineering and Management, Technion – Israel Institute of Technology, 1999. http://iew3.technion.ac.il/Labs/Opt/opt/Pap/CP_MD.pdf

20. *Kivinen J., Warmuth M.K.* Additive Versus Exponentiated Gradient Updates for Linear Prediction // Inform. Comput. 1997. V. 132. № 1. P. 1–64.

21. *Helmbold D.P., Kivinen J., Warmuth M.K.* Relative Loss Bounds for Single Neurons // IEEE Trans. Neural Networks. 1999. V. 10. № 6. P. 1291–1304.

22. *Kivinen J., Warmuth M.K.* Relative Loss Bounds for Multidimensional Regression Problems // Machine Learning. 2001. V. 45. № 3. P. 301–329.

23. *Cesa-Bianchi N., Gentile C.* Improved Risk Tail Bounds for On-Line Algorithms // Neural Information Processing Systems, NIPS 2004 Workshop on (Ab)Use of Bounds, Whistler, BC, Canada, December 18, 2004.
http://mercurio.srv.dsi.unimi.it/ cesabian/Pubblicazioni/iada.pdf

24. *Cesa-Bianchi N., Conconi A., Gentile C.* On the Generalization Ability of On-Line Learning Algorithms // IEEE Trans. Inform. Theory. 2004. V. 50. № 9. P. 2050–2057.

25. *Juditsky A., Nemirovski A.* Functional Aggregation for Nonparametric Estimation // Ann. Stat. 2000. V. 28. № 3. P. 681–712.

26. *Tsybakov A.* Optimal Rates of Aggregation // Computational Learning Theory and Kernel Machines. B. Scholkopf and M. Warmuth, eds. Lecture Notes in Artificial Intelligence. Heidelberg: Springer, 2003. V. 2777. P. 303–313.

27. *Вапник В.Н., Червоненкис А.Я.* Теория распознавания образов. М.: Наука, 1974.

28. *Breiman L.* Arcing the edge. Technical Report 486. Berkeley: Statistics Department, University of California, 1997.

29. *Friedman J., Hastie T., Tibshirani R.* Additive Logistic Regression: a Statistical View of Boosting (With Discussion and a Rejoinder by the Authors) // Ann. Stat. 2000. V. 28. № 2. P. 337–407.

30. *Tsybakov A.* Optimal Aggregation of Classifiers in Statistical Learning // Ann. Stat. 2004. V. 32. № 1. P. 135–166.

31. *Tarigan B., van de Geer S.A.* Adaptivity of Support Vector Machines with ℓ_1 Penalty. Technical Report MI 2004-14. Leiden: Mathematical Institute, University of Leiden, 2004. http://www.math.leidenuniv.nl/ geer/svm4.pdf

32. *Rockafellar R.T., Wets R.J.B.* Variational Analysis. New York: Springer, 1998.

33. *Kiwiel K.C.* Proximal Minimization Methods with Generalized Bregman Functions // SIAM J. Control Optim. 1997. V. 35. № 4. P. 1142–1168.

34. *Beck A., Teboulle M.* Mirror Descent and Nonlinear Projected Subgradient Methods for Convex Optimization // Oper. Research Letters. 2003. V. 31. № 3. P. 167–175.

35. *Поляк Б.Т., Цыпкин Я. З.* Критериальные алгоритмы стохастической оптимизации // АиТ. 1984. № 6. С. 95–104.

36. *Vajda I.* Theory of Statistical Inference and Information. Dordrecht: Kluwer, 1986.

Юдицкий Анатолий Борисович
Лаборатория моделирования и вычислительной математики,
Университет Жозефа Фурье, Гренобль, Франция
anatoli.iouditski@imag.fr
Назин Александр Викторович
Институт проблем управления им. В.А. Трапезникова РАН
nazine@ipu.rssi.ru
Цыбаков Александр Борисович
Ваятис Николас
Лаборатория теории вероятностей и стохастического моделирования,
Университет Пьера и Мари Кюри, Париж, Франция
tsybakov@ccr.jussieu.fr, vayatis@ccr.jussieu.fr