УДК 519.23, 519.226

Е. А. Крымова[1,2,3], *Е. О. Черноусова*[1]

[1]Московский физико-технический институт (государственный университет)
[2]Институт проблем передачи информации им. Харкевича РАН
[3]ООО «ДАТАДВАНС»

Оракульное неравенство для метода экспоненциального взвешивания упорядоченных оценок

Рассматривается задача восстановления неизвестного вектора по зашумленным наблюдениям при помощи упорядоченных линейных оценок. Основной целью является вывод оракульного неравенства, позволяющего контролировать риск оценки, полученной агрегацией упорядоченных оценок при использовании метода экспоненциального взвешивания. Предлагается новый метод получения оракульных неравенств, основанный на вероятностных свойствах несмещенной оценки риска. Показано, что экспоненциальное взвешивание позволяет улучшить оракульное неравенство Кнайпа [1].

Ключевые слова: экспоненциальное взвешивание, агрегация оценок, несмещенная оценка риска, упорядоченные оценки, оракульные неравенства.

1. Введение и основные результаты

Рассматривается задача оценивания вектора $\mu = (\mu_1, \dots, \mu_n)^\top$ по наблюдениям

$$Y_i = \mu_i + \sigma\xi_i, \quad i = 1, 2, \dots, n, \tag{1}$$

где ξ_i — стандартный белый гауссовский шум. Для простоты будем считать, что параметр $\sigma > 0$ известен.

Наша цель состоит в том, чтобы построить оценку вектора μ на основе семейства линейных оценок

$$\hat{\mu}_i^h(Y) = h_i Y_i, \ h \in \mathcal{H}, \tag{2}$$

где $\mathcal{H}$ — заданное множество векторов из $\mathbb{R}^n$, которое будет описано ниже.

Риск оценки $\hat{\mu}(Y) = (\hat{\mu}_i(Y), \dots, \hat{\mu}_n(Y))^\top$ измеряется величиной

$$R(\hat{\mu}, \mu) = \mathbf{E}_\mu \|\hat{\mu}(Y) - \mu\|^2,$$

здесь $\mathbf{E}_\mu$ — математическое ожидание по мере $\mathbf{P}_\mu$, порожденной наблюдениями (1), $\|\cdot\|$ и $\langle\cdot,\cdot\rangle$ обозначают норму и скалярное произведение в $\mathbb{R}^n$ соответственно

$$\|x\|^2 = \sum_{i=1}^n x_i^2, \quad \langle x, y\rangle = \sum_{i=1}^n x_i y_i.$$

Нетрудно показать, что среднеквадратичный риск линейной оценки $\hat{\mu}^h(Y)$ вычисляется следующим образом:

$$R(\hat{\mu}^h, \mu) = \|(1-h)\mu\|^2 + \sigma^2\|h\|^2.$$

Очевидно, что этот риск зависит от h и мы можем найти его минимум по $h \in \mathcal{H}$, который называют оракульным риском:

$$r^{\mathcal{H}}(\mu) = \min_{h \in \mathcal{H}} R(\hat{\mu}^h, \mu).$$

Однако мы не можем использовать оценку

$$\mu^*(Y) = h^* \cdot Y, \quad h^* = \arg\min_{h \in \mathcal{H}} R(\hat{\mu}^h, \mu),$$

так как она зависит от неизвестного вектора μ. Поэтому нашей целью является построение оценки $\tilde{\mu}^{\mathcal{H}}(Y)$ на основе заданных оценок $\hat{\mu}^h(Y)$, $h \in \mathcal{H}$, такой, чтобы ее риск был как можно ближе к риску оракула. То есть мы хотим, чтобы равномерно по $\mu \in \mathbb{R}^n$ выполнялось неравенство следующего вида:

$$R(\tilde{\mu}^{\mathcal{H}}, \mu) \leqslant r^{\mathcal{H}}(\mu) + \tilde{\Delta}^{\mathcal{H}}(\mu),$$

где $\tilde{\Delta}^{\mathcal{H}}(\mu)$ — остаточный член, который мы хотели бы сделать меньше риска оракула $r^{\mathcal{H}}(\mu)$. Такого типа неравенства часто называются оракульными. Основной задачей является поиск оценок с минимальным остаточным членом. В общем случае эта задача не имеет решения. Однако в некоторых случаях удается построить оценку $\tilde{\mu}^{\mathcal{H}}(Y)$ такую, что:

- $\tilde{\Delta}^{\mathcal{H}}(\mu) \leqslant \tilde{C} r^{\mathcal{H}}(\mu)$ для всех $\mu \in \mathbb{R}^n$, где $\tilde{C} > 1$ константа,
- $\tilde{\Delta}^{\mathcal{H}}(\mu) \ll r^{\mathcal{H}}(\mu)$ для всех $\mu :\ r^{\mathcal{H}}(\mu) \gg \sigma^2$.

Известно, что можно построить оценку с приведенными выше свойствами, если векторы в $\mathcal{H}$ являются упорядоченными по Кнайпу [1].

Определение 1. *Множество $\mathcal{H}$ состоит из упорядоченных векторов, если*

- $h_i \in [0,1],\ i = 1, \ldots, n$ *для всех* $h \in \mathcal{H}$,
- $h_{i+1} \leqslant h_i,\ i = 1, \ldots, n$ *для всех* $h \in \mathcal{H}$,
- *если для некоторого натурального k и некоторых $h, g \in \mathcal{H}$*

$$h_k < g_k, \quad \text{тогда} \quad h_i \leqslant g_i \quad \text{для} \quad i = 1, \ldots, n.$$

Последнее условие означает, что векторы из $\mathcal{H}$ могут быть естественным образом упорядочены, так как для любых $h, g \in \mathcal{H}$ возможно только два случая: $h_i \leqslant g_i$ или $h_i \geqslant g_i$ для всех $i = 1, \ldots, n$. Поэтому оценки из (2) и называют упорядоченными оценками.

Оценки во многих статистических моделях являются упорядоченными. Стандартные примеры: сглаживающие сплайны [2] и спектральные методы регуляризации [3], [4].

Сглаживающие сплайны. Сглаживающие сплайны используются для восстановления гладкой функции $f(x),\ x \in [0,1]$ по наблюдениям

$$Z_i = f(x_i) + \sigma \xi_i', \quad i = 1, \ldots, n, \tag{3}$$

где $x_i \in (0,1)$ и ξ_i' — стандартный белый гауссовский шум. Сглаживающий сплайн определяется как решение следующей оптимизационной задачи:

$$\hat{f}_\alpha(x, Z) = \arg\min_f \left\{ \sum_{i=1}^n [Z_i - f(x_i)]^2 + \alpha \int_0^1 [f^{(m)}(x)]^2 dx \right\}, \tag{4}$$

где $f^{(m)}(\cdot)$ — производная порядка m и $\alpha > 0$ — параметр сглаживания, значение которого обычно выбирается исходя из оптимизации критерия GCV (см., например, [5]). Для того чтобы преобразовать модель (3) к виду (1), рассмотрим базис Райнша—Деммлера $\psi_k(x),\ x \in [0,1],\ k = 1, \ldots, n,$ со свойством двойной ортогональности [6]:

$$\langle \psi_k, \psi_l \rangle_n = \delta_{kl}, \quad \int_0^1 \psi_k^{(m)}(x) \psi_l^{(m)}(x)\, dx = \delta_{kl} \lambda_k,\ k, l = 1, \ldots, n,$$

где здесь и ниже $\langle u, v \rangle_n$ обозначает скалярное произведение $\langle u, v \rangle_n = \frac{1}{n} \sum_{i=1}^n u(x_i) v(x_i)$ и собственные числа λ_k упорядочены $\lambda_n \geqslant \ldots \geqslant \lambda_1$.

Разлагая функцию $f(x)$ по базису Райнша—Деммлера:

$$f(x) = \sum_{k=1}^{n} \psi_k(x)\mu_k, \tag{5}$$

из (3) получаем

$$Y_k = \langle Y, \psi_k \rangle_n = \mu_k + \sigma\xi_k.$$

Затем, подставив (5) в (4), имеем

$$\hat{f}_\alpha(x, Y) = \sum_{k=1}^{n} \hat{\mu}_k \psi_k(k),$$

где

$$\hat{\mu} = \arg\min_{\mu} \left\{ \sum_{k=1}^{n} [Y_k - \mu_k]^2 + \alpha \sum_{k=1}^{n} \lambda_k \mu_k^2 \right\}.$$

Отсюда

$$\hat{\mu}_k = \frac{Y_k}{1 + \alpha\lambda_k}.$$

Поэтому задачи (1) – (2) и (3) – (4) эквивалентны при

$$h_k = h_k^\alpha = \frac{1}{1 + \alpha\lambda_k}.$$

Заметим, что базис Райнша—Деммлера не используется для вычисления сплайнов, так как для этого существуют быстрые алгоритмы (см. [7]). ■

Известно множество подходов для построения оценки, риск которой близок к риску оракула. Один из первых подходов связан с идеей несмещенного оценивания риска [8]. Упомянем классический в непараметрическом оценивании результат Кнайпа [1]. Обозначим несмещенную оценку риска оценки $\hat{\mu}^h(Y)$:

$$\bar{r}(Y, \hat{\mu}^h) = \|(1-h)Y\|^2 + 2\sigma^2 \sum_{i=1}^{n} h_i - \sigma^2 n. \tag{6}$$

Теорема 1. *Пусть*

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \bar{r}(Y, \hat{\mu}^h),$$

тогда равномерно по $\mu \in \mathbb{R}^n$

$$\mathbf{E}\|\hat{\mu}^{\hat{h}} - \mu\|^2 \leqslant r^{\mathcal{H}}(\mu) + K\sigma^2 \sqrt{1 + \frac{r^{\mathcal{H}}(\mu)}{\sigma^2}}, \tag{7}$$

где K *— универсальная константа.*

Другой подход заключается в построении оценки с помощью выпуклой комбинации заданных оценок с неатомарными весами. Основные результаты в этой области получены Немировским [9] и Катони [10] в предположении, что имеется дополнительная выборка для подбора весов. От такого предположения можно отказаться, если, например, использовать экспоненциальное взвешивание. В работах [11], [12], [13] было показано, что для экспоненциального взвешивания можно получить хорошие оракульные неравенства в различных статистических задачах.

В этом методе оценка строится следующим образом:

$$\bar{\mu}(Y) = \sum_{h \in \mathcal{H}} w^h(Y) \hat{\mu}^h(Y),$$

где $w^h(Y)$ — положительные веса, такие, что $\sum_{h\in\mathcal{H}} w^h(Y) = 1$ и

$$w^h(Y) = \pi^h \exp\left[-\frac{\bar r(Y,\hat\mu^h)}{2\beta\sigma^2}\right] \Big/ \sum_{g\in\mathcal{H}} \pi^g \exp\left[-\frac{\bar r(Y,\hat\mu^g)}{2\beta\sigma^2}\right].$$

Основной целью данной работы является вывод оракульного неравенства для метода экспоненциального взвешивания при использовании упорядоченных оценок.

Для того чтобы оракульное неравенство было верно при неограниченном увеличении числа элементов множества $\mathcal{H}$, в качестве априорных весов π^h выбираются веса следующего вида:

$$\pi^h = \left[1 - \exp\left\{-\frac{\|h^+\|_1 - \|h\|_1}{\beta}\right\}\right], \tag{8}$$

где h^+ — элемент из $\mathcal{H}$, следующий за h:

$$h^+ = \min\{g \in \mathcal{H} : g > h\},$$

$\pi^{h_{\max}} = 1$, где $h^{\max} = \max_{h\in\mathcal{H}} h$, и $\|\cdot\|_1$ обозначает l_1-норму в $\mathbb{R}^n$, то есть

$$\|h\|_1 = \sum_{i=1}^n |h_i|.$$

Далее будем предполагать, что выполнено

Условие 1. *Существуют константы $k_\circ, K$, такие, что*

$$\sum_{i=1}^n (h_i^2 - g_i^2) \geqslant k_\circ\big(\|h\|_1 - \|g\|_1\big) \text{ для всех } h \geqslant g \text{ из } \mathcal{H},$$

$$\|h^+\|^2 \leqslant K\|h\|^2 \text{ для всех } h \in \mathcal{H}.$$

Следующая верхняя граница риска оценки $\bar\mu(Y)$ является основным результатом статьи.

Теорема 2. *Пусть $\beta \geqslant 4$ и справедливо условие 1. Тогда равномерно по $\mu \in \mathbb{R}^n$ выполнено*

$$\mathbf{E}_\mu\|\bar\mu - \mu\|^2 \leqslant r^{\mathcal{H}}(\mu) + 2\beta\sigma^2 \log\left[\frac{r^{\mathcal{H}}(\mu)}{\sigma^2} + \Psi\left(\frac{r^{\mathcal{H}}(\mu)}{\sigma^2}\right)\right], \tag{9}$$

где $\Psi(x) > 0,\ x \geqslant 0$ — неубывающая функция, ограниченная в нуле, такая, что

$$\Psi(x) \leqslant \frac{Cx}{\log(x)}.$$

Заметим, что полученная оценка имеет такую же структуру, как и оракульное неравенство Кнайпа из теоремы 1. Отметим также, что в отличие от результата, приведенного в [12], верхняя оценка (9) не зависит ни от числа элементов множества $\mathcal{H}$, ни от n.

Сравним остаточные члены полученного неравенства и неравенства Кнайпа в теореме 1. Видно, что при $r^{\mathcal{H}}(\mu)\sigma^{-2} \approx 1$ остаточные члены имеют одинаковый порядок, а именно $C\sigma^2$. Однако при $r^{\mathcal{H}}(\mu)\sigma^{-2} \gg 1$

$$2\beta\sigma^2 \log\left[\frac{r^{\mathcal{H}}(\mu)}{\sigma^2} + \Psi\left(\frac{r^{\mathcal{H}}(\mu)}{\sigma^2}\right)\right] \ll K\sigma^2\sqrt{1 + \frac{r^{\mathcal{H}}(\mu)}{\sigma^2}}.$$

Таким образом, оракульное неравенство для экспоненциального взвешивания лучше оракульного неравенства Кнайпа из теоремы 1.

2. Доказательство

Доказательство теоремы 2 основано на комбинации методов для вывода оракульных неравенств, предложенных в [13] и [14]. Основная идея состоит в том, чтобы использовать следующее свойство несмещенной оценки риска: пусть

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \bar{r}(Y, \hat{\mu}^h),$$

тогда для достаточно малого $\varepsilon < 1$ существует $\hat{h}^\varepsilon \geqslant \hat{h}$ такое, что с вероятностью 1

$$\bar{r}(Y, \hat{\mu}^h) - \bar{r}(Y, \hat{\mu}^{\hat{h}}) \geqslant 2\beta\sigma^2\varepsilon\big[\|h\|^2 - \|\hat{h}\|^2\big] - 2\beta\sigma^2$$

для всех $h \geqslant \hat{h}^\varepsilon$. Это свойство означает, что веса $w^h(Y)$ экспоненциально убывают для больших h и поэтому можно получить следующую энтропийную оценку (см. лемму 5 ниже):

$$\sum_{h \in \mathcal{H}} w^h(Y) \log \frac{\pi^k}{w^h} \leqslant \log\Bigg[\sum_{h \leqslant \hat{h}^\varepsilon} \pi^h + C\varepsilon^{-1}\exp\big(C\varepsilon^{-1}\big)\Bigg].$$

Здесь и далее C обозначает универсальную константу.

По лемме 2 из [14] (см. лемму 3 ниже)

$$\mathbf{E}_\mu \|\hat{h}^\varepsilon\|^2 \leqslant \frac{C r^{\mathcal{H}}(\mu)}{(1 - C\varepsilon)\sigma^2} + \frac{C}{\varepsilon}.$$

Наконец, полученные результаты объединяются согласно доказательству теоремы 5 из [13].

2.1. Дополнительные утверждения

Следующие три леммы являются техническими, поэтому мы приводим их без доказательств, которые имеются в [14], [15].

Лемма 1. *При выполнении условий 1 для любых $h \in \mathcal{H}$*

- $$\sum_{g \geqslant h} \pi^g \exp\bigg\{-\frac{\|g\|_1}{\beta}\bigg\} = \exp\bigg\{-\frac{\|h\|_1}{\beta}\bigg\}, \tag{10}$$

- *существует константа $C_\circ$, такая, что*

$$\sum_{g \leqslant h} \pi^g \leqslant C_\circ \|h\|^2 + C_\circ, \tag{11}$$

- *существуют константы $\pi_\circ$ и π°, такие, что*

$$\pi_\circ \leqslant \sum_{g: \|h\|^2 \leqslant \|g\|^2 \leqslant \|h\|^2 + 1} \pi^g \leqslant \pi^\circ. \tag{12}$$

Лемма 2. *Пусть ξ_i — стандартный белый гауссовский шум и $\mathcal{G}$ — множество упорядоченных последовательностей. Тогда для любого $\alpha > 0$*

$$\mathbf{E}\max_{g \in \mathcal{G}}\bigg\{\pm\sum_{i=1}^n (g_i^2 - 2g_i)(\xi_i^2 - 1) - \alpha\sum_{i=1}^n g_i^2\bigg\} \leq \frac{C}{\alpha},$$

$$\mathbf{E}\max_{g \in \mathcal{G}}\left\{\sum_{i=1}^n (1 - g_i)^2 \xi_i \mu_i - \alpha\sum_{i=1}^n (1 - g_i)^2 \mu_i^2\right\} \leq \frac{C}{\alpha}.$$

Лемма 3. *Пусть*

$$\hat{h}^{\varepsilon} = \max\Big\{h : \big[\bar{r}(Y, \hat{\mu}^h) - \bar{r}^{\mathcal{H}}(Y)\big] \leqslant 2\beta\varepsilon\sigma^2\big[\|h\|^2 - \|\hat{h}\|^2\big] + 2\beta\sigma^2\Big\}, \tag{13}$$

где $\varepsilon \in (0, (5\beta)^{-1})$ *и*

$$\bar{r}^{\mathcal{H}}(Y) = \min_{h\in\mathcal{H}} \bar{r}(Y, \hat{\mu}^h), \quad \hat{h} = \arg\min_{h\in\mathcal{H}} \bar{r}(Y, \hat{\mu}^h).$$

Тогда

$$\sigma^2 \mathbf{E}_\mu \|\hat{h}^{\varepsilon}\|^2 \leqslant \frac{1}{1-5\beta\varepsilon} r^{\mathcal{H}}(\mu) + \frac{C\sigma^2}{(1-5\beta\varepsilon)\varepsilon}. \tag{14}$$

Следующая лемма является ключевой в доказательстве теоремы 2.

Лемма 4. *Для* $\beta \geqslant 4$ *риск оценки* $\bar{\mu}(Y)$ *ограничен сверху:*

$$\mathbf{E}_\mu \|\bar{\mu}(Y) - \mu\|^2 \leqslant \mathbf{E}_\mu \sum_{h\in\mathcal{H}} w^h(Y)\bar{r}(Y, \hat{\mu}^h).$$

Доказательство. Напомним, что несмещенная оценка риска для $\bar{\mu}_i(Y)$ и $\hat{\mu}_i^h(Y)$ вычисляется следующим образом (см. [16]):

$$\begin{aligned} \bar{r}(Y_i, \bar{\mu}_i) &= [\bar{\mu}_i(Y) - Y_i]^2 + 2\sigma^2 \frac{\partial \bar{\mu}_i(Y)}{\partial Y_i} - \sigma^2, \\ \bar{r}(Y_i, \hat{\mu}_i^h) &= [\hat{\mu}_i^h(Y) - Y_i]^2 + 2\sigma^2 h_i - \sigma^2. \end{aligned} \tag{15}$$

Так как $\sum_{h\in\mathcal{H}} w^h = 1$, имеем

$$[\bar{\mu}_i(Y) - Y_i]^2 = -\sum_{h\in\mathcal{H}} w^h(Y)[\bar{\mu}_i(Y) - \hat{\mu}_i^h(Y)]^2 + \sum_{h\in\mathcal{H}} w^h(Y)[\hat{\mu}_i^h(Y) - Y_i]^2$$

и

$$\sum_{h\in\mathcal{H}} \frac{\partial w^h(Y)}{\partial Y_i} = \sum_{h\in\mathcal{H}} w^h(Y) \frac{\partial \log w^h(Y)}{\partial Y_i} = 0.$$

Из определения $\bar{\mu}(Y)$ очевидно, что

$$\frac{\partial \bar{\mu}_i(Y)}{\partial Y_i} = \sum_{h\in\mathcal{H}} w^h(Y) \frac{\partial \hat{\mu}_i^h(Y)}{\partial Y_i} + \sum_{h\in\mathcal{H}} \frac{\partial w^h(Y)}{\partial Y_i} \hat{\mu}_i^h(Y).$$

Тогда (15) принимает вид

$$\begin{aligned} \bar{r}(Y_i, \bar{\mu}_i) = \sum_{h\in\mathcal{H}} w^h(Y)\bar{r}(Y_i, \hat{\mu}_i^h) + \sum_{h\in\mathcal{H}} w^h(Y) \Big\{ -[\bar{\mu}_i(Y) - \hat{\mu}_i^h(Y)]^2 + \\ + 2\sigma^2 \frac{\partial \log[w^h(Y)]}{\partial Y_i} [\hat{\mu}_i^h(Y) - \bar{\mu}_i(Y)] \Big\}. \end{aligned} \tag{16}$$

Для оценки второй суммы в (16) используем

$$\log w^h(Y) = -\frac{\bar{r}(Y, \hat{\mu}^h)}{2\beta\sigma^2} + \log(\pi^h) - \log\Big\{\sum_{g\in\mathcal{H}} \pi^g \exp\Big[-\frac{\bar{r}(Y, \hat{\mu}^g)}{2\beta\sigma^2}\Big]\Big\}.$$

Поэтому

$$\sum_{h\in\mathcal{H}} w^h(Y) \frac{\partial \log w^h(Y)}{\partial Y_i} [\hat{\mu}_i^h(Y) - \bar{\mu}_i(Y)] = -\frac{1}{2\beta\sigma^2} \sum_{h\in\mathcal{H}} w^h(Y) \frac{\partial \bar{r}(Y, \hat{\mu}^h)}{\partial Y_i} [\hat{\mu}_i^h(Y) - \bar{\mu}_i(Y)].$$

Очевидно, что

$$\frac{\partial \bar{r}(Y_i, \hat{\mu}_i^h)}{\partial Y_i} = 2(1-h_i)^2 Y_i,$$

тогда

$$\sum_{h\in\mathcal{H}} w^h(Y) \frac{\partial \log w^h(Y)}{\partial Y_i} [\hat{\mu}_i^h(Y) - \bar{\mu}_i(Y)] = -\frac{1}{\beta\sigma^2} Y_i^2 \sum_{h\in\mathcal{H}} w^h(Y)[h_i - 1]^2 [h_i - \bar{h}_i], \tag{17}$$

где

$$\bar{h}_i = \sum_{h\in\mathcal{H}} w^h(Y) h_i.$$

Заметим, что

$$(1-h_i)^2 = (1-\bar{h}_i)^2 + (\bar{h}_i - h_i)^2 + 2(1-\bar{h}_i)(\bar{h}_i - h_i),$$

поэтому

$$-Y_i^2 \sum_{h\in\mathcal{H}} w^h(Y)(h_i-1)^2(h_i - \bar{h}_i) = Y_i^2(1-\bar{h}_i)^2 \sum_{h\in\mathcal{H}} w^h(Y)(\bar{h}_i - h_i) +$$

$$+ Y_i^2 \sum_{h\in\mathcal{H}} w^h(Y)(\bar{h}_i - h_i)^2(\bar{h}_i - h_i + 2 - 2\bar{h}_i) =$$

$$= 2Y_i^2 \sum_{h\in\mathcal{H}} w^h(Y)(\bar{h}_i - h_i)^2 \left(1 - \frac{h_i + \bar{h}_i}{2}\right) \leqslant 2 \sum_{h\in\mathcal{H}} w^h(Y)[\bar{\mu}_i(Y) - \hat{\mu}_i^h(Y)]^2.$$

Утверждение леммы следует из (16), (17) и последнего неравенства. ■

Лемма 5. *Пусть* $\{q^h \leqslant 1,\ h \in \mathcal{H}\}$ *— неотрицательная последовательность, такая, что*

- *для всех* $h \geqslant \tilde{h}$

$$q^h \leqslant \exp\left\{-\varepsilon\left[\sum_{s=1}^n (h_i^2 - \tilde{h}_i^2)\right] - 1\right\}, \quad \varepsilon > 0,$$

- *для некоторых* h^**, таких, что* $\|h^*\|^2 \leqslant \|\tilde{h}\|^2$

$$q^g \geqslant q_\circ, \ \text{для всех} \ \ g \in \mathcal{G}_{h^*} = \Big\{g \in \mathcal{H}:\ \|h^*\|^2 \leqslant \|g\|^2 \leqslant \|h^*\|^2 + 1\Big\}.$$

Пусть

$$W^h = \pi^h q^h \left[\sum_{g\in\mathcal{H}} \pi^g q^g\right]^{-1}.$$

Тогда

$$H(W^h) \stackrel{\text{def}}{=} \sum_{h\in\mathcal{H}} W^h \log\frac{\pi^h}{W^h} \leqslant \log\left[\sum_{h\leqslant\tilde{h}} \pi^h + \exp[R(\varepsilon)]\right],$$

где

$$R(\varepsilon) = \log\frac{C[1+\varepsilon]}{\varepsilon} + \frac{C[1+\varepsilon-\varepsilon\log(q_\circ)]}{\varepsilon q_\circ}.$$

Доказательство. Разобъем $\mathcal{H}$ на два подмножества $\mathcal{Q} = \{h \geqslant \tilde{h}\} \cup \mathcal{G}_{h^*}$ и $\mathcal{P} = \mathcal{H} \setminus \mathcal{Q}$.

Обозначим для краткости

$$P = \sum_{h\in\mathcal{P}} \pi^h q^h, \quad Q = \sum_{h\in\mathcal{Q}} \pi^h q^h.$$

Из выпуклости $\log(x)$

$$H(W^h) = \frac{P}{P+Q}\sum_{h\in\mathcal{P}}\frac{\pi^h q^h}{P}\log\frac{(P+Q)/P}{q^h/P} + \frac{Q}{P+Q}\sum_{h\in\mathcal{Q}}\frac{\pi^h q^h}{Q}\log\frac{(P+Q)/Q}{q^h/Q} \leqslant$$
$$\leqslant \frac{P}{P+Q}\log\frac{P+Q}{P} + \frac{Q}{P+Q}\log\frac{P+Q}{Q} + \frac{P}{P+Q}\log\Bigl(\sum_{h\in\mathcal{P}}\pi^h\Bigr) +$$
$$+\frac{1}{P+Q}\Bigl[\sum_{h\in\mathcal{Q}}\pi^h q^h\log\frac{1}{q^h} + Q\log(Q)\Bigr].$$

Заметим, что $x\log\left(\frac{1}{x}\right)$ является возрастающей функцией при $x\in[0,\mathrm{e}^{-1}]$. Поэтому, используя неравенство $1-\exp(-\varepsilon) > (1+\varepsilon)^{-1}\varepsilon$, условие 1 и (12), получим

$$\sum_{h\in\mathcal{Q}}\pi^h q^h\log\frac{1}{q^h} \leqslant \sum_{h\in\mathcal{G}_{h^*}}\pi^h\log\frac{1}{q_\circ} + \sum_{h\geqslant\tilde h}\pi^h\exp\Bigl[-\varepsilon\sum_{i=1}^n(h_i^2-\tilde h_i^2)-1\Bigr]\Bigl[\varepsilon\sum_{i=1}^n(h_i^2-\tilde h_i^2)+1\Bigr] \leqslant$$
$$\leqslant \pi^\circ\log\frac{1}{q_\circ} + \frac{1}{\mathrm{e}}\sum_{h\geqslant\tilde h}\pi^h\exp\bigl[-\varepsilon K_\circ(\|h\|_1-\|\tilde h\|_1)\bigr]\times\bigl[K_\circ\varepsilon(\|h\|_1-\|\tilde h\|_1)+1\bigr].$$

Для того чтобы продолжить это неравенство, используем тот факт, что из (8) $\pi^h \leqslant \|h^+\|_1 - \|h\|_1$, следовательно:

$$\sum_{h\in\mathcal{Q}}\pi^h q^h\log\frac{1}{q^h} \leqslant C + \frac{1}{\mathrm{e}}\sum_{h\geqslant\tilde h}\exp\bigl[-\varepsilon K_\circ(\|h\|_1-\|\tilde h\|_1)\bigr]\times$$
$$\times\bigl[K_\circ\varepsilon\bigl(\|h\|_1-\|\tilde h\|_1\bigr)+1\bigr]\bigl(\|h^+\|_1-\|h\|_1\bigr). \qquad (18)$$

Для того чтобы ограничить сверху правую часть неравенства, рассмотрим множество $\tilde{\mathcal{H}} = \{h\in\mathcal{H}:\ h\geqslant\tilde h\}$. Предположим, что множество $\tilde{\mathcal{H}} = \{h_k,\ k=1,\ldots\}$ упорядочено по норме l_1, то есть $\|h_{k+1}\|_1 \geqslant \|h_k\|_1$. Обозначим для краткости

$$S_i = \|h_i\|_1 - \|\tilde h\|_1,\quad h_i\in\tilde{\mathcal{H}}.$$

Тогда

$$\sum_{h\geqslant\tilde h}\exp\bigl[-\varepsilon K_\circ(\|h\|_1-\|\tilde h\|_1)\bigr]\bigl[2\varepsilon\bigl(\|h\|_1-\|\tilde h\|_1\bigr)+1\bigr]\bigl(\|h^+\|_1-\|h\|_1\bigr) =$$
$$= \sum_{i\geqslant1}\exp\bigl[-\varepsilon K_\circ S_i\bigr]\bigl[\varepsilon S_i+1\bigr](S_{i+1}-S_i). \qquad (19)$$

Проверим выполнение

$$\max_{S_k,k\geq1}\sum_{i\geqslant1}\exp\bigl[-\varepsilon K_\circ S_i\bigr][S_{i+1}-S_i]_+ \leqslant \frac{C}{\varepsilon K_\circ}, \qquad (20)$$

где $[x]_+ = \max(0,x)$. Из решения уравнения

$$\frac{\partial}{\partial S_i}\sum_{i\geqslant1}\exp\bigl[-\varepsilon K_\circ S_i\bigr][S_{i+1}-S_i]_+ = 0$$

получаем

$$S_{i+1}-S_i = \frac{\exp\bigl[\varepsilon K_\circ(S_i-S_{i-1})\bigr]-1}{\varepsilon K_\circ}.$$

Поэтому

$$\exp(-\varepsilon K_\circ S_i)(S_{i+1}-S_i) = \frac{\exp[-\varepsilon K_\circ S_{i-1}]-\exp[-\varepsilon K_\circ S_i]}{\varepsilon K_\circ},$$

получаем (20).

Аналогично можно показать, что

$$\max_{S_k,k\geq 1}\sum_{i\geqslant 1} S_i\exp\bigl[-\varepsilon K_\circ S_i\bigr][S_{i+1}-S_i]_+\leqslant\frac{C}{\varepsilon K_\circ}. \tag{21}$$

Из (18) – (21) следует

$$\sum_{h\in\mathcal{Q}}\pi^h q^h\log\frac{1}{q^h}\leqslant\frac{C[1+\varepsilon-\varepsilon\log(q_\circ)]}{\varepsilon} \tag{22}$$

и

$$Q=\sum_{h\in\mathcal{Q}}\pi^h q^h\leqslant\frac{C[1+\varepsilon]}{\varepsilon}. \tag{23}$$

Поэтому

$$\log(Q)\leqslant\log\frac{C[1+\varepsilon]}{\varepsilon}. \tag{24}$$

Обозначим для краткости

$$x=\frac{Q}{P+Q}.$$

Тогда из (22) – (24) получаем

$$H(W^h)\leqslant\max_{x\in[0,1]}\left\{-x\log(x)-(1-x)\log(1-x)+(1-x)\log\left(\sum_{h\in\mathcal{P}}\pi^h\right)+xR(\varepsilon)\right\}, \tag{25}$$

где

$$R(\varepsilon)=\log\frac{C[1+\varepsilon]}{\varepsilon}+\frac{C[1+\varepsilon-\varepsilon\log(q_\circ)]}{\varepsilon q_\circ}.$$

Значение x^*, максимизирующее правую часть (25), находится из уравнения

$$\log\frac{1-x^*}{x^*}=\log\left(\sum_{h\in\mathcal{P}}\pi^h\right)-R(\varepsilon),$$

то есть

$$x^*=\left\{1+\left(\sum_{h\in\mathcal{P}}\pi^h\right)\exp[-R(\varepsilon)]\right\}^{-1}.$$

Таким образом, из (25) следует

$$\begin{aligned}H(W^h)\leqslant&\log\left(\sum_{h\in\mathcal{P}}\pi^h\right)-\log(1-x^*)-x^*\left[\log\frac{x^*}{1-x^*}+\log\left(\sum_{h\in\mathcal{P}}\pi^h\right)-R(\varepsilon)\right]=\\ &=\log\left(\sum_{h\in\mathcal{P}}\pi^h\right)-\log(1-x^*)=\log\left[\sum_{h\in\mathcal{P}}\pi^h+\mathrm{e}^{R(\varepsilon)}\right]\leqslant\log\left[\sum_{h<\tilde{h}}\pi^h+\mathrm{e}^{R(\varepsilon)}\right].\quad\blacksquare\end{aligned}$$

2.2. Доказательство теоремы 2

Из определения $w^h(Y)$ имеем

$$\log[w^h(Y)] = -\frac{1}{2\sigma^2\beta}\bar{r}(Y,\hat{\mu}^h) + \log\pi^h - \log\left\{\sum_{g\in\mathcal{H}}\pi^g\exp\left[-\frac{\bar{r}(Y,\hat{\mu}^g)}{2\beta\sigma^2}\right]\right\} =$$

$$= \frac{1}{2\sigma^2\beta}\bar{r}(Y,\hat{\mu}^{\hat{h}}) - \frac{1}{2\sigma^2\beta}\bar{r}(Y,\hat{\mu}^h) + \log\pi^h - \log\left\{\sum_{g\in\mathcal{H}}\pi^g\exp\left[-\frac{\bar{r}(Y,\hat{\mu}^g)-\bar{r}(Y,\hat{\mu}^{\hat{h}})}{2\beta\sigma^2}\right]\right\},$$

где $\hat{h} = \arg\min_{h\in\mathcal{H}}\bar{r}(Y,\hat{\mu}^h)$. Поэтому

$$\sum_{h\in\mathcal{H}} w^h(Y)\bar{r}(Y,\hat{\mu}^h) = \bar{r}(Y,\hat{\mu}^{\hat{h}}) + 2\beta\sigma^2\sum_{h\in\mathcal{H}} w^h(Y)\log\frac{\pi^h}{w^h(Y)} - \\ -2\beta\sigma^2\log\left\{\sum_{g\in\mathcal{H}}\pi^g\exp\left[-\frac{\bar{r}(Y,\hat{\mu}^g)-\bar{r}(Y,\hat{\mu}^{\hat{h}})}{2\beta\sigma^2}\right]\right\}. \tag{26}$$

Заметим, что для любого h, такого, что $\|\hat{h}\|^2 \leqslant \|h\|^2 \leqslant \|\hat{h}\|^2+1$, из условия 1 следует

$$\exp\left[-\frac{\bar{r}(Y,\hat{\mu}^h)-\bar{r}(Y,\hat{\mu}^{\hat{h}})}{2\beta\sigma^2}\right] = \exp\left[-\frac{\|(1-h)Y\|^2-(1-\hat{h})Y\|^2}{2\beta\sigma^2} - \\ -\frac{1}{\beta}\sum_{i=1}^n(h_i-\hat{h}_i)\right] \geqslant \exp\left[-\frac{K_\circ}{\beta}\right]. \tag{27}$$

Преобразуем последний член в правой часть неравенства (26), используя упорядоченность векторов из $\mathcal{H}$:

$$\log\left\{\sum_{g\in\mathcal{H}}\pi^g\exp\left[-\frac{\bar{r}(Y,\hat{\mu}^g)-\bar{r}(Y,\hat{\mu}^{\hat{h}})}{2\beta\sigma^2}\right]\right\} \geqslant \log\left\{\sum_{g\geqslant\hat{h}}\pi^g\exp\left[-\frac{\bar{r}(Y,\hat{\mu}^g)-\bar{r}(Y,\hat{\mu}^{\hat{h}})}{2\beta\sigma^2}\right]\right\} = \\ = \log\left\{\sum_{g\geqslant\hat{h}}\pi^g\exp\left[-\frac{\|(1-g)Y\|^2-\|(1-\hat{h})Y\|^2}{2\beta\sigma^2} - \frac{1}{\beta}\sum_{i=1}^n[g_i-\hat{h}_i]\right]\right\} \geqslant \\ \geqslant \log\left\{\sum_{g\geqslant\hat{h}}\pi^g\exp\left[-\frac{1}{\beta}\sum_{i=1}^n[g_i-\hat{h}_i]\right]\right\}. \tag{28}$$

Тогда из (28) и (10) следует

$$\log\left\{\sum_{g\in\mathcal{H}}\pi^g\exp\left[-\frac{\bar{r}(Y,\hat{\mu}^g)-\bar{r}(Y,\hat{\mu}^{\hat{h}})}{2\beta\sigma^2}\right]\right\} \geqslant 0.$$

Ограничим сверху второй член справа в неравенстве (26), используя леммы 3 и 5.

Пусть $\hat{h}^\varepsilon$ определено в (13). Тогда для любого $h > \hat{h}^\varepsilon$

$$\left[\bar{r}(Y,\hat{\mu}^h) - r^{\mathcal{H}}(Y)\right] \geqslant 2\beta\varepsilon\sigma^2\left[\|h\|^2 - \|\hat{h}\|^2\right] + 2\beta\sigma^2.$$

Из (11), (27) и леммы 5 получаем

$$\mathbf{E}_\mu\sum_{h\in\mathcal{H}} w^h(Y)\log\frac{\pi^h}{w^h(Y)} \leqslant \mathbf{E}_\mu\log\left\{\sum_{h\leqslant\hat{h}^\varepsilon}\pi^h + \exp[R(\varepsilon)]\right\} \leqslant \log\left\{\mathbf{E}_\mu\|\hat{h}^\varepsilon\|^2 + 1 + \exp[R(\varepsilon)]\right\}.$$

Затем, огранивая $\mathbf{E}_\mu \|\hat{h}_\varepsilon\|^2$ с помощью (14), получаем

$$\mathbf{E}_\mu \sum_{h\in\mathcal{H}} w^h(Y)\bar{r}(Y,\hat{\mu}^h) \leqslant \mathbf{E}_\mu \bar{r}^{\mathcal{H}}(Y) + 2\beta\sigma^2 \log\biggl\{\frac{1}{1-5\beta\varepsilon}\cdot\frac{r^{\mathcal{H}}(\mu)}{\sigma^2} + \\ + \frac{C}{(1-5\beta\varepsilon)\varepsilon} + \exp[R(\varepsilon)]\biggr\}. \tag{29}$$

Для завершения доказательства теоремы остается минимизировать правую часть (29) по ε. Полагая $\varepsilon \leqslant 1/(5\beta)$, имеем

$$\frac{1+\beta\varepsilon}{1-5\beta\varepsilon}\cdot\frac{r^{\mathcal{H}}(\mu)}{\sigma^2} + \frac{C}{(1-5\beta\varepsilon)\varepsilon} + \exp[R(\varepsilon)] \leqslant \frac{r^{\mathcal{H}}(\mu)}{\sigma^2} + \frac{C\varepsilon r^{\mathcal{H}}(\mu)}{\sigma^2} + \frac{C}{\varepsilon} + \exp\left[\frac{C}{\varepsilon}\right].$$

Поэтому, выбирая

$$\Psi(x) = C \min_{\varepsilon\in[0,1/(5\beta)]}\left[\varepsilon x + \frac{1}{\varepsilon} + \exp\left(\frac{C}{\varepsilon}\right)\right],$$

из леммы 4 и неравенства (29) следует теорема 2, так как

$$\mathbf{E}_\mu \bar{r}^{\mathcal{H}}(Y) \leqslant r^{\mathcal{H}}(\mu).$$

Ясно, что $\Psi(0)$ ограничено сверху. Также нетрудно проверить, что при $\rho\to 0$,

$$\varepsilon^*(\rho) = \arg\min_{\varepsilon\in[0,1/(5\beta)]}\left\{C\varepsilon + \rho\left[\frac{C}{\varepsilon} + \exp\left(\frac{C}{\varepsilon}\right)\right]\right\} \approx \log^{-1}\frac{C}{\rho} + 2\log^{-2}\frac{C}{\rho}\log\left(\log\frac{C}{\rho}\right).$$

Отсюда $\Psi(x) \leqslant Cx/\log(Cx)$. ■

Работа выполнена при поддержке Лаборатории структурных методов анализа данных в предсказательном моделировании, МФТИ, грант правительства РФ дог. 11.G34.31.0073

Литература

1. *Kneip A.* Ordered linear smoothers // Annals of Stat. — 1994. — V. 22. — P. 835–866.
2. *Speckman P.* Spline smoothing and optimal rates of convergence in nonparametric regression // Ann. Statist. — 1985. — V. 13. — P. 970–983.
3. *Тихонов А. Н., Арсенин В. Я.* Методы решения некорректных задач. — М.: Наука, 1974.
4. *Engl H. W., Hanke M., Neubauer A.* Regularization of Inverse Problems. Mathematics and its Applications. V. 375. — Springer, 2000.
5. *Wahba G.* Spline Models for Observational Data. — Philadelphia: SIAM, 1990.
6. *Demmler A., Reinsch C.* Oscillation matrices with spline smoothing // Numer. Math. — 1975. — V. 24. — P. 375–382.
7. *Green P. J., Silverman B. V.* Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach. — Chapman & Hall, 1993.
8. *Akaike H.* Information theory and an extension of the maximum likelihood principle // Proc. 2nd Intern. Symp. Inf. Theory. — 1973. — P. 267–281.
9. *Nemirovski A. S.* Topics in non-parametric statistics. Lectures Notes in Math. V. 1738. — Berlin: Springer-Verlag, 2000.
10. *Catoni O.* Statistical learning theory and stochastic optimization Lectures Notes in Math. V. 1851. — Berlin: Springer-Verlag, 2004.

11. *Rigolet Ph., Tsybakov A.* Sparse estimation by exponential weighting // arXiv:1108.5116v1 [math.ST].

12. *Dalalyan A., Salmon J.* Sharp oracle inequalities for aggregation of affine estimators // arXiv:1104.3969v2 [math.ST].

13. *Leung G., Barron A.* Information theory and mixing least-squares regressions // IEEE Transactions on Information Theory. — 2006. — V. 52. N. 8. — P. 3396–3410.

14. *Golubev Yu.* On universal oracle inequalities related to high dimensional linear models // Ann. Statist. — 2010. — V. 38, N. 5. — P. 2751–2780.

15. *Chernousova E., Golubev Yu., Krymova E.* Ordered smoothers with exponential weighting // arXiv:1211.4207 [Probab. Theory Relat. Fields].

16. *Stein, C.* Estimation of the mean of a multivariate normal distribution // Annals of Stat. — 1981. — V. 9. — P. 1135–1151.

Поступила в редакцию 05.12.2012.