Московский физико-технический институт (государственный университет)

Факультет радиотехники и кибернетики

Кафедра проблем передачи и обработки информации

Работа допущена к защите

зав. кафедрой

_____ акад. РАН Кулешов А. П.

«____» _____ 2014 г.

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
на соискание ученой степени
МАГИСТРА

Тема: **Неравенства концентрации для метода экспоненциального взвешивания.**

Направление: 010600 – Прикладные математика и физика

Выполнил студент гр. 811 _____ Островский Д. М.

Научный руководитель,

д. ф.-м. н. _____ Голубев Г. К.

Москва – 2014

Содержание

Введение

Важным в непараметрической математической статистике (см., например, монографии [1], [2]) является следующий вопрос. Пусть $f : [0,1] \to \mathbb{R}$ – неизвестная функция регрессии из пространства $L^2[0,1]$ функций, квадратично интегрируемых на $[0,1]$. Наблюдаются значения этой функции на фоне аддитивного белого гауссовского шума с известной дисперсией

$$y(t)\,dt = f(t)\,dt + \sigma\,dW_t, \quad t \in [0,1], \tag{1}$$

где W_t – стандартный винеровский процесс. Пусть качество оценок $\hat{f}(\cdot)$ функции $f(\cdot)$ измеряется с помощью квадратичного риска

$$R(\hat{f}, f) = \mathbf{E}_f \left\| \hat{f}(\cdot) - f(\cdot) \right\|_2^2, \tag{2}$$

где $\mathbf{E}_f$ – математическое ожидание по мере (1), связанной с настоящей функцией регрессии f. Требуется построить «хорошую» оценку $\hat{f}$ с низким квадратичным риском (2).

Часто бывает удобно работать со следующей переформулировкой данной задачи. Пусть $\boldsymbol{\mu} = \{\mu_k\}_{k=1}^{\infty}$ – последовательность коэффициентов Фурье функции f в некотором базисе $\{\phi_k\}_{k=1}^{\infty}$, ортонормированном на $[0,1]$. В силу равенства Парсеваля наблюдения (1) и квадратичный риск (2) переходят, соответственно, в

$$Y_k = \mu_k + \sigma\xi_k,\ k \in \mathbb{N}, \tag{3}$$

$$R(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) = \mathbf{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2^2; \tag{4}$$

здесь ξ_k – независимые стандартные гауссовские случайные величины, $\mathbf{E}_{\boldsymbol{\mu}}$ — усреднение по мере, порождающей наблюдения (3), и $\|\cdot\|_2$ – евклидова норма в пространстве квадратично суммируемых последовательностей ℓ^2. Отождествляя функции из $L^2[0,1]$, отличающиеся лишь на множестве лебеговой меры нуль, легко видеть, что $f \in L^2[0,1]$ взаимно однозначно соответствует последовательности $\boldsymbol{\mu} \in \ell^2$, и оценивание f в модели (1) сводится к оцениванию квадратично суммируемой последовательности $\boldsymbol{\mu} \in \ell^2$ в модели (3).

Итак, цель статистика – построить по наблюдениям $\mathbf{Y} = \{Y_k\}_{k=1}^{\infty}$ из (3) оценку $\boldsymbol{\mu}$ с как можно меньшим квадратичным риском (4). Возникает, однако, следующая пробле-

ма: никто не гарантирует, что существует оценка, равномерно лучшая для всех допустимых $\boldsymbol{\mu}$ (функций f). Широко распространены два подхода к решению этой проблемы, и оба они связаны с заданием некоторой априорной информации о задаче.

В рамках первого – байесовского – подхода эта информация формулируется в виде априорного распределения π параметра $\boldsymbol{\mu}$. Например, можно предположить, что компоненты $\boldsymbol{\mu}$ – независимые гауссовские величины с нулевыми средними и дисперсиями

$$\mathbf{E}_\pi \mu_k^2 = \Sigma_k^2, \tag{5}$$

где $\mathbf{E}_\pi$ —усреднение по априорному распределению. Как только априорное распределение фиксировано, можно вычислить усредненный по нему риск

$$r_\pi(\hat{\boldsymbol{\mu}}) = \mathbf{E}_\pi R(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu})$$

и искать оценку, которая минимизирует такой риск. Эта задача уже имеет хорошо определенное решение: минимизировать усредненный риск r_π будет байесовская оценка $\hat{\boldsymbol{\mu}}_\pi(\mathbf{Y}) = \mathbf{E}(\boldsymbol{\mu}|\mathbf{Y})$, то есть апостериорное среднее оцениваемого параметра.

Существует и другой широко распространенный подход к заданию априорной информации – минимаксный. Согласно ему, априорная информация о задаче формулируется в виде предположения, что параметр $\boldsymbol{\mu}$ принадлежит некоторому заданному множеству Θ. В качестве такого множества часто выступает эллипсоид

$$\Theta = \left\{ \boldsymbol{\mu} : \sum_{k=1}^{\infty} \theta_k^2 \mu_k^2 \le 1 \right\}.$$

Например, в случае, если выбран тригонометрический базис

$$\begin{aligned} &\phi_1(x) = 1, \\ &\phi_{2k}(x) = 2^{-1/2}\cos(2\pi k x), \quad \phi_{2k+1}(x) = 2^{-1/2}\sin(2\pi k x), \quad k \in \mathbb{N}, \end{aligned}$$

эллипсоид

$$\Theta = \left\{ \boldsymbol{\mu} : \sum_{j=1}^{\infty} (2\pi j)^{2m} (\mu_{2j}^2 + \mu_{2j+1}^2) \le \mathrm{M} \right\} \tag{6}$$

соответствует, как легко проверить, априорной информации о том, что функция f принадлежит периодическому соболевскому классу $\tilde{\mathcal{W}}_2^m(\mathrm{M})$ функций, для которых выполнено ограничение

$$\left\| f^{(m)}(\cdot) \right\|_2 \le \mathrm{M}$$

вместе с граничными условиями $f_+^{(k)}(0) = f_-^{(k)}(1), \quad k = 0, \ldots, m-1$.

Как только априорная информация в виде множества $\boldsymbol{\Theta}$ задана, в рамках минимаксного подхода вычисляется максимальный риск оценки $\hat{\boldsymbol{\mu}}(\mathbf{Y})$

$$r_{\Theta}(\hat{\boldsymbol{\mu}}) = \sup_{\boldsymbol{\mu} \in \Theta} R(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu})$$

и ищется так называемая минимаксная оценка, которая его минимизирует:

$$\hat{\boldsymbol{\mu}}_{\Theta} = \operatorname*{argmin}_{\hat{\boldsymbol{\mu}}} r_{\Theta}(\hat{\boldsymbol{\mu}}).$$

Отметим, что поиск минимаксной оценки является существенно более сложной задачей, чем вычисление байесовской оценки; чаще всего точно найти ее не представляется возможным. Впрочем, в некоторых важных случаях минимаксный риск может быть эффективно оценен. Первые результаты для минимаксных оценок на эллипсоидах использовали теоретико-информационные идеи и связаны с именами И. А. Ибрагимова и Р. З. Хасьминского [3], см. также монографию [4]. В 80-х гг. настоящий прорыв в этой области был осуществлен М. С. Пинскером [5]: им было доказано, что при мягких ограничениях на эллипсоид (6) минимаксный риск среди всех возможных оценок асимптотически (при $\sigma \to 0$) эквивалентен минимаксному риску лишь среди покомпонентных линейных оценок вида

$$\hat{\mu}_k = h_k Y_k, \quad 0 \le h_k \le 1. \tag{7}$$

Доказательство этого результата построено на следующей идее: для параметра $\boldsymbol{\mu}$ вводится фиктивное гауссовское априорное распределение, которое благодаря условию $\sigma \to 0$ концентрируется на параметрическом множестве $\boldsymbol{\Theta}$, после чего используется тот факт, что для гауссовского распределения байесовская оценка является линейной. Позднее похожие результаты, связывающие минимаксный риск среди всех и среди линейных оценок, были получены и для других статистических моделей (см. обзорную статью [6] и приведенные в ней источники).

Основная проблема, возникающая при применении как байесовского, так и минимаксного подхода, связана с невозможностью полностью задать априорную информацию: как правило, точно неизвестны ни априорное распределение $\boldsymbol{\pi}$ в байесовском подходе, ни множество $\boldsymbol{\Theta}$ в минимаксном подходе. В случае байесовского подхода эту трудность обычно преодолевают, вводя параметрическую модель уже для априорного распределения. Параметры этой модели принято называть гиперпараметрами, и их

значения оценивают на основании наблюдений. Для этого либо максимизируют маргинальное, то есть усредненное по априорному распределению, правдоподобие (на практике такой подход используется, к примеру, в методах регрессии на основе гауссовских процессов [7]), либо используют кросс-валидацию.

В то же время ситуация с мимимаксным подходом существенно сложнее. В частности, если множество $\mathbf{\Theta}$ имеет вид (6), то никаких достаточно хорошо аргументированных подходов к оцениванию параметров m, M уже не существует.

Простая идея, позволяющая решить проблему неполной априорной информации, состоит в том, чтобы смотреть на баейсовский и минимаксный подходы просто как на методы, позволяющие генерировать семейства хороших оценок. Зачастую эти оценки имеют вид (7) и семейство описывается небольшим числом параметров. Например, при априорной информации вида (6) мы получим следующее семейство минимаксных оценок [5]:

$$\hat{\mu}_k^{m,\Lambda}(\mathbf{Y}) = \left[1 - \left(\frac{k}{\Lambda}\right)^m\right]_+ Y_k, \quad \Lambda > 0,\ m \in \mathbb{N}, \tag{8}$$

где Λ определяется по m и M; здесь и далее $[\cdot]_+$ означает $\max(\cdot, 0)$. Это семейство имеет вид (7) и описывается всего двумя параметрами. Другой пример: при часто используемом априорном предположении, что μ_k независимы и $\mu_k \sim \mathcal{N}\left(0, \Sigma_k^2\right)$ с $\Sigma_k^2 = \mathrm{A}\exp(-\alpha k^2)$, семейство байесовских оценок имеет вид

$$\hat{\mu}_k^{\alpha,\mathrm{A}}(\mathbf{Y}) = \frac{\Sigma_k^2}{\Sigma_k^2 + \sigma^2} Y_k = \frac{\mathrm{A}}{\mathrm{A} + \sigma^2 \exp(\alpha k^2)} Y_k.$$

Это семейство также описывается всего двумя параметрами и имеет вид (7).

В данной перспективе естественно возникает следующая задача. Пусть для оценивания $\boldsymbol{\mu}$ на основании наблюдений (3) задано некоторое семейство оценок следующего вида

$$\hat{\mu}_k^{\mathbf{h}}(\mathbf{Y}) = h_k Y_k, \quad \mathbf{h} = \{h_k\}_{k=1}^{\infty} \in \mathcal{H}, \tag{9}$$

причем $\forall\, \mathbf{h} \in \mathcal{H}$ все компоненты $0 \le h_k \le 1$ и $\sum_{k=1}^{\infty} h_k < \infty$; кроме того, множество $\mathcal{H}$ для простоты будем считать дискретным. Требуется построить как можно более хорошую финальную оценку $\boldsymbol{\mu}$ с помощью оценок из этого семейства. Заметим, что при фиксированном векторе $\mathbf{h}$ квадратичный риск оценки $\hat{\boldsymbol{\mu}}^{\mathbf{h}}(\mathbf{Y})$ вычисляется очень просто, а именно как

$$R\left(\hat{\boldsymbol{\mu}}^{\mathbf{h}}, \boldsymbol{\mu}\right) = \|(\mathbf{1} - \mathbf{h}) \cdot \boldsymbol{\mu}\|_2^2 + \sigma^2 \|\mathbf{h}\|_2^2,$$

где $\mathbf{1}$ – последовательность из единиц; здесь и далее в тексте символ $\cdot$ используется для обозначения операции покомпонентного произведения последовательностей, то есть $\mathbf{a} \cdot \mathbf{b} = \{a_k\, b_k\}_{k=1}^{\infty}$.

Итак, для задачи комбинирования оценок цель статистика состоит в поиске метода комбинации оценок из данного семейства, который давал бы финальную оценку с как можно меньшим риском. При этом естественно сравнивать риск найденной оценки с наименьшим возможным риском среди оценок в исходном семействе:

$$r^{\mathcal{H}}(\boldsymbol{\mu}) = \min_{\mathbf{h}\in\mathcal{H}}\Big\{\|(\mathbf{1}-\mathbf{h})\cdot\boldsymbol{\mu}\|_2^2 + \sigma^2\,\|\mathbf{h}\|_2^2\Big\}.$$

Этот риск часто называется риском оракула, потому что он равен риску «оценки»

$$\hat{\boldsymbol{\mu}}^{\circ}(\mathbf{Y}) = \mathbf{h}^{\circ}(\boldsymbol{\mu})\cdot\mathbf{Y}, \quad \text{где} \quad \mathbf{h}^{\circ}(\boldsymbol{\mu}) = \operatorname*{argmin}_{\mathbf{h}\in\mathcal{H}}\Big\{\|(\mathbf{1}-\mathbf{h})\cdot\boldsymbol{\mu}\|_2^2 + \sigma^2\,\|\mathbf{h}\|_2^2\Big\}, \tag{10}$$

которая в обычном смысле оценкой, конечно, не является, поскольку зависит от неизвестного вектора $\boldsymbol{\mu}$. Статистику этот вектор неизвестен, но оракулу он доступен. Понятно, что если исходное семейство оценок достаточно хорошее (то есть соответствующая оракульная оценка имеет низкий риск), а метод комбинации позволяет с большой точностью приблизиться к оракульной оценке (10), то финальная оценка будет иметь низкий риск. В то же время, для заданного метода комбинации оценок отдельный интерес представляет то, насколько комбинированная финальная оценка проигрывает оракульной. Подобные результаты обычно формулируются в форме так называемых оракульных неравенств для риска агрегированной оценки:

$$\mathbf{E}_{\boldsymbol{\mu}}\,\|\bar{\boldsymbol{\mu}}(\mathbf{Y}) - \boldsymbol{\mu}\|_2^2 \le r^{\mathcal{H}}(\boldsymbol{\mu}) + \delta\left(r^{\mathcal{H}}(\boldsymbol{\mu})\right),$$

где поправка $\delta\left(r^{\mathcal{H}}(\boldsymbol{\mu})\right)$ мала по сравнению с риском оракула $r^{\mathcal{H}}(\boldsymbol{\mu})$.

Интересно получать оракульные неравенства, равномерные по параметру $\boldsymbol{\mu}$. Соответствующие методы комбинации будут адаптивными в следующем смысле: независимо от того, какая оценка оракульная для данного $\boldsymbol{\mu}$, комбинированная оценка будет не сильно уступать оракульной. При этом величина поправки δ в оракульном неравенстве характеризует степень адаптивности метода комбинирования для заданного семейства оценок.

Достаточно естественным методом, позволяющим получить хорошую итоговую оценку на основе уже имеющихся, является *выпуклая агрегация* оценок:

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

где $w^{\mathbf{h}}$ – некоторые числа, называемые весами, которые неотрицательны и в сумме равны единице. Одним из таких методов, получившим в последние годы значительное внимание, является метод экспоненциального взвешивания. Этот метод заключается в следующем.

1. Для каждой оценки $\hat{\boldsymbol{\mu}}^{\mathbf{h}}(\mathbf{Y})$ подсчитывается величина $r(\mathbf{Y},\mathbf{h})$ – несмещенная оценка риска $R(\hat{\boldsymbol{\mu}}^{\mathbf{h}},\boldsymbol{\mu})$ оценки $\hat{\boldsymbol{\mu}}^{\mathbf{h}}(\mathbf{Y})$ с точностью до аддитивной постоянной, одинаковой для всех оценок [8].

2. В качестве финальной оценки берется выпуклая комбинация

$$\hat{\boldsymbol{\mu}}^{\beta}(\mathbf{Y}) = \sum_{\mathbf{h}\in\mathcal{H}} w^{\mathbf{h}}(\mathbf{Y})\hat{\boldsymbol{\mu}}^{\mathbf{h}}(\mathbf{Y})$$

   с весами

$$w^{\mathbf{h}}(\mathbf{Y}) = \frac{\pi^{\mathbf{h}}}{Z(\mathbf{Y})}\exp\left(-\frac{r(\mathbf{Y},\mathbf{h})}{2\beta\sigma^2}\right).$$

   Здесь

   – $\boldsymbol{\pi} = \left\{\pi^{\mathbf{h}}\right\}_{\mathbf{h}\in\mathcal{H}}$ – некоторое фиксированное распределение на оценках семейства, отражающее априорное знание о них. Например, оно может присваивать больший вес более «простым» оценкам, штрафуя более «сложные».

   – $\beta > 0$ – параметр метода, управляющий степенью концентрации весов на оценке с наименьшей величиной $r(\mathbf{Y},\mathbf{h})$.

   – $Z(\mathbf{Y})$ – константа, вычисляемая из условия нормировки весов.

Данная работа посвящена исследованию метода экспоненциального взвешивания для семейства проекционных оценок. **Целью** исследования является ответ на следующий естественный вопрос.

– Как и с какой скоростью потери метода $\left\|\bar{\boldsymbol{\mu}}^{\beta}(\mathbf{Y}) - \boldsymbol{\mu}\right\|_2^2$ концентрируются вблизи риска оракула?

Работа имеет следующую структуру. В Главе 1 дается обзор известных результатов для методов комбинации оценок; особый акцент делается на методе экспоненциального взвешивания. В Главе 2 обсуждается основной теоретический результат данной работы, дающий ответ на поставленные вопросы. В ней же приведены результаты численного эксперимента, иллюстрирующего полученную теорему. Доказательству

теоретического результата посвящена Глава 3; в ее разделах последовательно вводятся необходимые понятия, формулируются вспомогательные леммы, после чего проводится само доказательство.

На защиту выносятся следующие положения.

– Доказано оракульное неравенство концентрации для метода экспоненциального взвешивания в упорядоченном семействе проекционных оценок. Это неравенство покрывает, в частности, не исследованный в литературе диапазон значений параметра метода.

– Проведен численный эксперимент для иллюстрации полученного теоретического результата.

Глава 1

Обзор литературы

В этой главе вначале мы сделаем краткий обзор известных результатов для различных методов комбинации оценок. В первом разделе речь пойдет о классическом методе минимизации эмпирического риска, выбирающем, в зависимости от наблюдений, единственную оценку из семейства. Во втором разделе мы перейдем к так называемым методам выпуклой агрегации, в которых выбор «хеджируется»: в качестве финальной выбирается выпуклая комбинация оценок. Из этих методов мы подробно остановимся на методе экспоненциального взвешивания, исследуемом в данной работе: приведем его эвристическое обоснование и перечислим результаты, известные для этого метода.

1.1. Метод минимизации эмпирического риска

По-видимому, исторически первым, а также самым простым выражением идеи комбинирования статистических оценок является выбор единственной оценки с помощью метода несмещенного оценивания риска. Этот давно известный в статистике метод восходит к классическим работам H. Akaike [9] и C. Mallows [10] и состоит в следующем.

Вначале выписываются несмещенные оценки $r(\mathbf{Y},\mathbf{h})$ рисков $R(\hat{\boldsymbol{\mu}}^{\mathbf{h}},\boldsymbol{\mu})$, то есть такие функции наблюдений $\mathbf{Y}$, что

$$\mathbf{E}_{\boldsymbol{\mu}}\, r(\mathbf{Y},\mathbf{h}) = R(\hat{\boldsymbol{\mu}}^{\mathbf{h}},\boldsymbol{\mu}).$$

Это оказывается несложно сделать, например, с помощью аппарата несмещенного оценивания риска Штейна (Stein Unbiased Risk Estimate) [8]. Нетрудно проверить, что для каждой оценки $\hat{\boldsymbol{\mu}}^{\mathbf{h}}$ вида (7) несмещенная оценка риска имеет, с точностью до не зависящей от $\mathbf{h}$ аддитивной постоянной, вид

$$r(\mathbf{Y},\mathbf{h}) = \left\|\hat{\boldsymbol{\mu}}^{\mathbf{h}}(\mathbf{Y})\right\|_2^2 - 2\langle \mathbf{Y}, \hat{\boldsymbol{\mu}}^{\mathbf{h}}(\mathbf{Y})\rangle + 2\sigma^2 \sum_{k=1}^{\infty} h_k,$$

где $\langle\cdot,\cdot\rangle$ – стандартное скалярное произведение в ℓ^2. Далее, в качестве финальной берется оценка с минимальным $r(\mathbf{Y},\mathbf{h})$, то есть оценка

$$\bar{\boldsymbol{\mu}}^0(\mathbf{Y}) = \hat{\mathbf{h}}^\circ \cdot \mathbf{Y}, \quad \text{где} \quad \hat{\mathbf{h}}^\circ = \operatorname*{argmin}_{\mathbf{h}\in\mathcal{H}} r(\mathbf{Y},\mathbf{h}). \tag{1.1}$$

Оракульное неравенство для метода минимизации эмпирического риска было получено в работе [11] А. Кнайпом. Этот результат справедлив для семейств оценок вида (9) с дополнительными *требованиями упорядоченности*:

(R1) Для любого $\mathbf{h} \in \mathcal{H}$ выполнено $1 \geq h_1 \geq h_2 \geq \ldots \geq 0$.

(R2) Для любых двух $\mathbf{h} \neq \mathbf{g} \in \mathcal{H}$ выполнено либо $h_i \geq g_i \quad \forall i \in \mathbb{N}$, либо $h_i \leq g_i \quad \forall i \in \mathbb{N}$.

Первое из этих требований по сути означает, что компоненты оцениваемого параметра $\boldsymbol{\mu}$ упорядочены в соответствии с представлением статистика об их важности. Вспомнив про исходную задачу (1), это можно трактовать как предположение о том, что ортонормированный базис подобран «правильно» для априорной информации о f (к примеру, выбран тригонометрический базис, а f гладкая). Смысл второго требования в том, что упорядочено само семейство оценок. Отметим, что это требование позволяет отказаться от дискретности множества $\mathcal{H}$; в результате, данным требованиям, после выбора правильного базиса, удовлетворяют многие семейства статистических оценок. В качестве таких примеров можно привести:

– ядерные оценки, параметризованные переменной шириной полосы;

– полиномиальную регрессию с переменным числом ортогональных базисных полиномов;

– оценки линейной регрессии с регуляризацией по методу Тихонова.

В последний класс попадает, например, оценивание на основе минимаксных сглаживающих сплайнов [12].

Итак, приведем основной результат, известный для метода минимизации эмпирического риска.

Теорема 1 (Kneip, 1994) *Пусть задано семейство оценок* (9)*, удовлетворяющее требованиям упорядоченности (R1), (R2). Для оценки* (1.1) *при всех* $\boldsymbol{\mu} \in \ell^2$ *справедливо неравенство*

$$\left\|\bar{\boldsymbol{\mu}}^0(\mathbf{Y}) - \boldsymbol{\mu}\right\|_2^2 \leq r^{\mathcal{H}}(\boldsymbol{\mu}) + \mathrm{K}\sigma^2\sqrt{\frac{r^{\mathcal{H}}(\boldsymbol{\mu})}{\sigma^2}}, \tag{1.2}$$

где $r^{\mathcal{H}}(\boldsymbol{\mu})$ *– риск оракула, а* K *– универсальная константа.*

Данное оракульное неравенство показывает адаптивность финальной оценки (1.1): в случае если $r^{\mathcal{H}}(\boldsymbol{\mu})/\sigma^2 \gg 1$, риск финальной оценки будет примерно равен риску оракула, поскольку в этом случае $\sqrt{r^{\mathcal{H}}(\boldsymbol{\mu})/\sigma^2} \ll r^{\mathcal{H}}(\boldsymbol{\mu})/\sigma^2$. Если же $r^{\mathcal{H}}(\boldsymbol{\mu})$ по порядку равно σ^2, то же самое будет справедливо и для риска финальной оценки.

1.2. Выпуклая агрегация

Обратимся теперь к более общим методам *выпуклой агрегации* оценок. В этих методах финальная оценка имеет вид

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

где веса $\mathbf{w} = \left\{w^{\mathbf{h}}\right\}_{\mathbf{h}\in\mathcal{H}}$ неотрицательны и в сумме равны единице и $\mathcal{H}$ дискретно. Отметим, что метод минимизации эмпирического риска формально является одним из методов выпуклой агрегации.

По-видимому, впервые методы выпуклой агрегации были математически исследованы А. С. Немировским и А. Б. Юдитским ([13], [14]) и независимо О. Катони [15]; впоследствии эти методы были адаптированы для широкого класса статистический моделей (см., например, [16], [17], [18], [19]).

Основным вопросом при выпуклой агрегации оценок является, очевидно, правильный выбор весов $\mathbf{w}$. Подход к выбору весов, использованный в работах [13], [14], [15], сводится к следующему. Пусть, во-первых, агрегируется всего M оценок, то есть $|\mathcal{H}| = M$. Во-вторых, пусть кроме наблюдений (3) нам доступны еще дополнительные наблюдения того же $\boldsymbol{\mu}$:

$$Y'_k = \mu_k + \epsilon\xi'_k, \quad k \in \mathbb{N}, \tag{1.3}$$

где $\boldsymbol{\xi}'$ — независимый от $\boldsymbol{\xi}$ из (3) стандартный белый гауссовский шум с известной дисперсией ϵ^2. Эти наблюдения в некотором смысле являются стохастическим аналогом оракула: если бы $\epsilon = 0$, то они играли бы его роль. Построим оценки исходного семейства по первой выборке (3), а веса $\mathbf{w}$ подберем по второй выборке (1.3). Конкретнее, возьмем набор весов, минимизирующий эмпирический риск $\|\mathbf{Y}' - \bar{\boldsymbol{\mu}}^{\mathbf{w}}\|_2^2$, то есть

решающий оптимизационную задачу

$$\min_{\mathbf{w}} \|\mathbf{Y}' - \bar{\boldsymbol{\mu}}^{\mathbf{w}}\|_2^2 \tag{1.4}$$
$$\text{s.t.} \quad \mathbf{w} \in \mathcal{S}_M = \left\{ \mathbf{w} \geq 0, \quad \sum_{\mathbf{h} \in \mathcal{H}} w^{\mathbf{h}} = 1 \right\}.$$

Это конечномерная выпуклая задача минимизации квадратичной функции на вероятностном симплексе, для решения которой известны эффективные алгоритмы [20].

Основной результат [14], справедливый для выбора весов $\mathbf{w}^*$, решающих задачу (1.4), заключается в том, что, при некоторых дополнительных требованиях, для квадратичного риска агрегированной оценки справедливо[1]

$$\mathbf{E}_{\boldsymbol{\mu}} \left\| \bar{\boldsymbol{\mu}}^{\mathbf{w}^*}(\mathbf{Y}, \mathbf{Y}') - \boldsymbol{\mu} \right\|_2^2 - r^{\mathcal{H}}(\boldsymbol{\mu}) = O\left(\sqrt{\ln M}\right).$$

Этот результат базируется на соображениях стохастической оптимизации. Действительно, фиксируем мысленно набор весов $\mathbf{w}$. В силу взаимной независимости $\mathbf{Y}$ и $\mathbf{Y}'$ математическое ожидание *эмпирического риска*

$$\|\mathbf{Y}' - \bar{\boldsymbol{\mu}}^{\mathbf{w}}\|_2^2$$

с точностью до не зависящей от $\mathbf{w}$ постоянной равно настоящему квадратичному риску

$$\mathbf{E}_{\boldsymbol{\mu}} \|\bar{\boldsymbol{\mu}}^{\mathbf{w}} - \boldsymbol{\mu}\|_2^2. \tag{1.5}$$

Таким образом, минимизируются по $\mathbf{w}$ случайные зашумленные наблюдения, в среднем дающие целевую функцию $\mathbf{w}$, которую мы бы хотели минимизировать на самом деле. Если бы мы и строили, и агрегировали оценки по одной и той же выборке, то есть минимизировали бы $\|\mathbf{Y} - \bar{\boldsymbol{\mu}}^{\mathbf{w}}\|_2^2$ вместо $\|\mathbf{Y}' - \bar{\boldsymbol{\mu}}^{\mathbf{w}}\|_2^2$, это было бы несправедливо, так как величина $\|\mathbf{Y} - \bar{\boldsymbol{\mu}}^{\mathbf{w}}\|_2^2$ уже не является несмещенной оценкой (1.5).

Самый очевидный недостаток приведенного результата, по-видимому, в том, что он справедлив только для конечного числа оценок M и потому неприменим уже в случае счетного множества $\mathcal{H}$. Это происходит в результате того, что в доказательстве не используется важный факт: $\hat{\boldsymbol{\mu}}^{\mathbf{h}}$ являются статистическими оценками $\boldsymbol{\mu}$ и потому должны концентрироваться настоящего значения этого параметра; результат останется справедливым, если оценки $\hat{\boldsymbol{\mu}}^{\mathbf{h}}$ заменить на фиксированные последовательности, то

[1] На самом деле этот результат выполнен для любого семейства из M оценок, не обязательно имеющего вид (9). В этом разделе мы решили сформулировать его в таком виде для единства обозначений.

есть вместо задачи агрегирования статистических оценок решать так называемую задачу функциональной агрегации.

Второй недостаток состоит в том, что в оптимизационной задаче (1.4) не учитывается возможное априорное знание об оценках в исходном семействе. Например, она могла бы быть заданной в виде распределения $\boldsymbol{\pi} = \left\{\pi^{\mathbf{h}}\right\}_{\mathbf{h}\in\mathcal{H}}$, где более «сложным» в некотором смысле оценкам присваивался бы меньший вес.

Наконец, очевидно, что метод, в котором оценки строятся по одной выборке, а агрегируются с помощью другой, приводит к потере имеющейся статистической информации, заключенной в обеих выборках. Действительно, дополнительные наблюдения можно было бы использовать для уточнения самих оценок, что привело бы к уменьшению оракульного риска.

От последних двух недостатков, впрочем, удалось избавиться в методе экспоненциального взвешивания, о котором пойдет речь в следующем разделе.

1.2.1. Метод экспоненциального взвешивания

Важным частным случаем метода выпуклой агрегации служит метод экспоненциального взвешивания, о котором уже шла речь ранее. Именно этот метод является объектом исследования в данной работе. Нам будет удобно придти к этому методу на основе следующих эвристических соображений.

Пусть вначале нам доступна, как и ранее, дополнительная выборка $\mathbf{Y}'$ (1.3), статистически независимая от исходной. Обратимся сначала к оптимизационной задаче (1.4). Как уже отмечалось ранее, при таком выборе весов не учитывается априорная информация об оценках $\hat{\boldsymbol{\mu}}^{\mathbf{h}}$, которая может быть задана в виде фиксированного распределения $\boldsymbol{\pi} = \{\pi^{\mathbf{h}}\}_{\mathbf{h}\in\mathcal{H}}$. Эта априорная информация может быть учтена, если надлежащим образом регуляризовать оптимизационную задачу. Будем «штрафовать» веса $\mathbf{w}$ за отличие от априорных весов $\boldsymbol{\pi}$ с помощью дивергенции Кульбака–Лейблера $\mathcal{K}(\cdot,\cdot)$, то есть перейдем к оптимизационной задаче

$$\begin{aligned} &\min_{\mathbf{w}} \quad \|\mathbf{Y}' - \bar{\boldsymbol{\mu}}^{\mathbf{w}}\|_2^2 + 2\beta\sigma^2\mathcal{K}(\mathbf{w},\boldsymbol{\pi}) \\ &\text{s.t.} \quad \mathbf{w} \in \mathcal{S} = \left\{\mathbf{w} \geq 0, \quad \sum_{\mathbf{h}\in\mathcal{H}} w^{\mathbf{h}} = 1\right\}, \end{aligned}$$

где параметр регуляризации $\beta > 0$ отвечает за относительную важность априорной информации; как и ранее, $\mathcal{S}$ – симплекс всех вероятностных распределений на множестве

$\mathcal{H}$.

Следующий шаг – верхняя апроксимация эмпирического риска. В силу выпуклости квадратичной функции для любых фиксированных весов $\mathbf{w}$ справедливо следующее неравенство:

$$\|\mathbf{Y}' - \bar{\boldsymbol{\mu}}^{\mathbf{w}}\|_2^2 = \left\|\mathbf{Y}' - \sum_{\mathbf{h}\in\mathcal{H}} w^{\mathbf{h}} \hat{\mu}^{\mathbf{h}}(\mathbf{Y})\right\|_2^2 \leq \sum_{\mathbf{h}\in\mathcal{H}} w^{\mathbf{h}} \left\|\mathbf{Y}' - \hat{\boldsymbol{\mu}}^{\mathbf{h}}(\mathbf{Y})\right\|_2^2.$$

Мы, таким образом, можем релаксировать целевую функцию последней оптимизационной задачи, перейдя от нее к задаче

$$\begin{aligned} \min_{\mathbf{w}} \quad & \sum_{\mathbf{h}\in\mathcal{H}} w^{\mathbf{h}} \left\|\mathbf{Y}' - \hat{\boldsymbol{\mu}}^{\mathbf{h}}(\mathbf{Y})\right\|_2^2 + 2\beta\mathcal{K}(\mathbf{w},\boldsymbol{\pi}) \\ \text{s.t.} \quad & \mathbf{w} \in \mathcal{S}. \end{aligned}$$

Применив условия Каруша–Куна–Таккера и проделав несложные вычисления, нетрудно убедиться, что эта задача уже имеет явное решение:

$$\begin{aligned} w^{\mathbf{h}} &= \frac{\pi^{\mathbf{h}}}{Z'} \exp\left(-\frac{\left\|\mathbf{Y}' - \hat{\boldsymbol{\mu}}^{\mathbf{h}}(\mathbf{Y})\right\|_2^2}{2\beta\sigma^2}\right) \\ &= \frac{\pi^{\mathbf{h}}}{Z} \exp\left(-\frac{\left\|\hat{\boldsymbol{\mu}}^{\mathbf{h}}(\mathbf{Y})\right\|_2^2 - 2\langle\mathbf{Y}', \hat{\boldsymbol{\mu}}^{\mathbf{h}}(\mathbf{Y})\rangle}{2\beta\sigma^2}\right), \end{aligned} \tag{1.6}$$

где Z, Z' – нормировочные константы.

Наконец, попытаемся обойтись единственной выборкой $\mathbf{Y}$. Основная сложность при этом связана со скалярным произведением $\langle\mathbf{Y}', \hat{\boldsymbol{\mu}}^{\mathbf{h}}(\mathbf{Y})\rangle$ в (1.6). Нам нужно заменить эту величину на некоторую другую, близкую к ней, но построенную только по выборке $\mathbf{Y}$. Для того чтобы понять, как это сделать, посмотрим, как устроено это скалярное произведение в вероятностном смысле. Очевидно, мы имеем

$$\begin{aligned} \langle\mathbf{Y}', \hat{\boldsymbol{\mu}}^{\mathbf{h}}(\mathbf{Y})\rangle = \sum_{k=1}^{\infty} h_k \left(\mu_k + \epsilon\xi_k'\right)(\mu_k + \sigma\xi_k) = \sum_{k=1}^{\infty} h_k\mu_k^2 \\ +\sigma\sum_{k=1}^{\infty} h_k\mu_k\xi_k + \epsilon\sum_{k=1}^{\infty} h_k\mu_k\xi_k' + \epsilon\sigma\sum_{k=1}^{\infty} h_k\xi_k\xi_k'. \end{aligned} \tag{1.7}$$

Заметим, что во второй строке в последней формуле находятся случайные величины, имеющие при фиксированном векторе $\mathbf{h}$ нулевое среднее. Имеем

$$\begin{aligned} \langle\mathbf{Y}, \hat{\boldsymbol{\mu}}^{\mathbf{h}}(\mathbf{Y})\rangle = \sum_{k=1}^{\infty} h_k (\mu_k + \sigma\xi_k)^2 = \sum_{k=1}^{\infty} h_k\mu_k^2 + \sigma^2\sum_{k=1}^{\infty} h_k \\ +2\sigma\sum_{k=1}^{\infty} h_k\mu_k\xi_k + \sigma^2\sum_{k=1}^{\infty} h_k\left(\xi_k^2 - 1\right). \end{aligned} \tag{1.8}$$

В этом тождестве в последней строке также находятся случайные величины с нулевым средним. Поэтому, сравнивая (1.7) и (1.8), можно сказать, что

$$\mathbf{E}_{\boldsymbol{\mu}}\langle \mathbf{Y}', \hat{\boldsymbol{\mu}}^{\mathbf{h}}(\mathbf{Y})\rangle = \mathbf{E}_{\boldsymbol{\mu}}\langle \mathbf{Y}, \hat{\boldsymbol{\mu}}^{\mathbf{h}}(\mathbf{Y})\rangle - \sigma^2 \sum_{k=1}^{\infty} h_k.$$

Эти соображения обосновывают метод выпуклой агрегации оценок, который называется *методом экспоненциального взвешивания* и имеет вид

$$\bar{\boldsymbol{\mu}}^{\beta}(\mathbf{Y}) = \sum_{\mathbf{h}\in\mathcal{H}} w^{\mathbf{h}}(\mathbf{Y})\hat{\boldsymbol{\mu}}^{\mathbf{h}}(\mathbf{Y}), \quad w^{\mathbf{h}}(\mathbf{Y}) = \frac{\pi^{\mathbf{h}}}{Z(\mathbf{Y})} \exp\left(-\frac{r(\mathbf{Y},\mathbf{h})}{2\beta\sigma^2}\right),$$

где $\beta > 0$ – параметр метода;

$$r(\mathbf{Y},\mathbf{h}) = \left\|\hat{\boldsymbol{\mu}}^{\mathbf{h}}(\mathbf{Y})\right\|_2^2 - 2\langle \mathbf{Y}, \hat{\boldsymbol{\mu}}^{\mathbf{h}}(\mathbf{Y})\rangle + 2\sigma^2 \sum_{k=1}^{\infty} h_k,$$

$$Z(\mathbf{Y}) = \sum_{\mathbf{h}\in\mathcal{H}} \pi^{\mathbf{h}} \exp\left(-\frac{r(\mathbf{Y},\mathbf{h})}{2\beta\sigma^2}\right).$$

Как несложно сообразить, в последней части приведенного анализа мы убедились в том, что $r(\mathbf{Y},\mathbf{h})$ является несмещенной оценкой риска $R(\hat{\boldsymbol{\mu}}^{\mathbf{h}},\boldsymbol{\mu})$ оценки $\hat{\boldsymbol{\mu}}^{\mathbf{h}}$ с точностью до аддитивной поправки, не зависящей от $\mathbf{h}$.

Отметим, что параметр β управляет степенью концентрации весов на оценке с наименьшей величиной $r(\mathbf{Y},\mathbf{h})$. Действительно, положив $\beta = \infty$, мы получим $w^{\mathbf{h}} = \pi^{\mathbf{h}}$, то есть полный отказ от информации, заключенной в наблюдениях. В то же время предел $\beta \to 0$ даст метод классический метод минимизации эмпирического риска: весь вес сосредотачивается на оценке с наименьшим значением $r(\mathbf{Y},\mathbf{h})$ (очевидно, такая оценка почти наверное единственна).[2]

1.2.2. Результаты для экспоненциального взвешивания

Первый известный нам результат для метода экспоненциального взвешивания со строго положительными β был получен в [21]. В этой работе рассматривалось произвольное семейство из M проекционных оценок вида (9); для такого семейства все $\mathbf{h} \in \mathcal{H}$ – последовательности из единиц и нулей.

Теорема 2 (Leung & Barron, 2006) *Пусть M проекционных оценок агрегируются по методу экспоненциального взвешивания с $\beta \geq 2$. Для всех $\boldsymbol{\mu} \in \ell^2$ справедливо неравен-*

[2] Из этих соображений мы обозначили оценку по методу минимизации эмпирического риска как $\bar{\boldsymbol{\mu}}^0$, то есть $\bar{\boldsymbol{\mu}}^{\beta}$ при $\beta = 0$.

ство

$$\mathbf{E}_{\boldsymbol{\mu}} \left\| \bar{\boldsymbol{\mu}}^{\beta}(\mathbf{Y}) - \boldsymbol{\mu} \right\|_2^2 \leq \min_{\mathbf{w} \in \mathcal{S}_M} \left\{ \mathbf{E}_{\boldsymbol{\mu}} \left\| \bar{\boldsymbol{\mu}}^{\mathbf{w}}(\mathbf{Y}) - \boldsymbol{\mu} \right\|_2^2 + 2\beta\sigma^2 \mathcal{K}(\mathbf{w}, \boldsymbol{\pi}) \right\}.$$

Для равномерного априорного распределения $\boldsymbol{\pi}^{\mathbf{h}} \equiv 1/M$ *правую часть данного неравенства можно оценить сверху как* [3]

$$\min_{\mathbf{w} \in \mathcal{S}_M} \left\{ \mathbf{E}_{\boldsymbol{\mu}} \left\| \bar{\boldsymbol{\mu}}^{\mathbf{w}}(\mathbf{Y}) - \boldsymbol{\mu} \right\|_2^2 + 2\beta\sigma^2 \mathcal{K}(\mathbf{w}, \boldsymbol{\pi}) \right\} \leq r^{\mathcal{H}}(\boldsymbol{\mu}) + 2\beta\sigma^2 \ln M.$$

Как видим, авторам этой теоремы удалось обойтись без разделения выборки и связанной с ней потерей статистической информации. К сожалению, данный результат, как и результат из [14], можно использовать только для конечного семейства оценок (что, как и для результата из [14] является на самом деле следствием того факта, что в доказательстве не учитывается концентрация статистических оценок около оцениваемого параметра).

Ограничение $M < \infty$ удалось снять в работе [23], рассматривая, как и в [11], семейства оценок (9) с требованиями упорядоченности (R1), (R2).

Теорема 3 (Golubev, 2013) *Пусть задано не более чем счетное семейство оценок вида* (9)*, для которого выполнены требования (R1), (R2). При* $\beta \geq 2$*, априорных весах* $\boldsymbol{\pi}^{\mathbf{h}} \equiv 1$ *и всех* $\boldsymbol{\mu} \in \ell^2$ *для оценки по методу экспоненциального взвешивания справедливо неравенство*

$$\mathbf{E}_{\boldsymbol{\mu}} \left\| \bar{\boldsymbol{\mu}}^{\beta}(\mathbf{Y}) - \boldsymbol{\mu} \right\|_2^2 \leq r^{\mathcal{H}}(\boldsymbol{\mu}) + 2\beta\sigma^2 \ln \left\{ \frac{r^{\mathcal{H}}(\boldsymbol{\mu})}{\sigma^2} \left[1 + \Psi\left(\frac{r^{\mathcal{H}}(\boldsymbol{\mu})}{\sigma^2} \right) \right] \right\}, \tag{1.9}$$

где $\Psi(r)$ *– ограниченная функция* $r \geq 1$*, и* $\Psi(r) \to 0$ *при* $r \to +\infty$*.*

Стоит также упомянуть, что для риска оценивания с помощью метода экспоненциального взвешивания известны достаточно хорошие верхние границы и в некоторых других статистических моделях [22], [24], [25].

Два результата для метода экспоненциального взвешивания, теорему 1, покрывающую случай $\beta = 0$, то есть метод минимизации эмпирического риска, и теорему 3, работающую для $\beta \geq 2$, можно сравнить между собой, сделав следующие выводы.

[3] Мы не затрагиваем здесь вопрос выбора априорного распределения и все результаты для простоты формулируем для равномерного $\boldsymbol{\pi}$. Заинтересованному в данном вопросе читателю следует обратиться к работам [21], [19], [22].

Во-первых, при формальном сравнении последних слагаемых в неравенствах (1.2) и (1.9) на первый взгляд может показаться, что экспоненциальное взвешивание должно быть существенно лучше классического метода минимизации эмпирического риска. В действительности это не совсем так, и чтобы разобраться в том, что происходит на самом деле, нужно понимать, какие методы и результаты лежат в основе доказательств теорем 1 и 3.

Теорема 1 является следствием следующего концентрационного неравенства Кнайпа [11]

$$\mathbf{P}\Big\{\big\|\bar{\boldsymbol{\mu}}^0(\mathbf{Y}) - \boldsymbol{\mu}\big\|_2 \geq \sqrt{r^{\mathcal{H}}(\boldsymbol{\mu})} + x\Big\} \leq \exp\Big\{-\mathrm{C}_1\,[x - \mathrm{C}_2]_+^2\Big\}, \tag{1.10}$$

где $\mathrm{C}_{1,2}$ – универсальные константы.

В тоже время теорема 3 – результат совершенно иного класса. В основе ее доказательства [23] лежит метод, предложенный в [21] и основанный на формуле Стейна [8] для несмещенной оценки риска. По существу это означает, что в отличии от (1.10) концентрация ошибки оценивания $\big\|\bar{\boldsymbol{\mu}}^\beta(\mathbf{Y}) - \boldsymbol{\mu}\big\|_2^2$ вблизи риска оракула неравенством (1.9) никак не контролируется.

Во-вторых, в теореме 3 присутствует условие $\beta \geq 2$. К сожалению, невозможно сказать, действительно ли это условие принципиально важно для реальной применимости экспоненциального взвешивания, или же это чисто техническое ограничение, связанное с методом доказательства. Отметим, что в литературе неизвестно никаких оракульных неравенств для случая $0 < \beta < 2$.

Таким образом, относительно метода экспоненциального взвешивания остаются неясными два естественных вопроса:

– как и с какой скоростью концентрируется $\big\|\bar{\boldsymbol{\mu}}^\beta(\mathbf{Y}) - \boldsymbol{\mu}\big\|_2^2$ вблизи риска оракула (при любых $\beta \geq 0$)?

– при каких значениях β возможно и целесообразно применение экспоненциального взвешивания?

В следующей главе мы сформулируем теорему, дающую ответ на поставленные вопросы. При этом теоремы 1 и 3 становятся нашими основными ориентирами, с которыми следует сравнивать новый результат.

Глава 2

Полученные результаты

Чтобы сформулировать основной теоретический результат данной работы, нам понадобиться вначале ввести некоторые дополнительные объекты.

Избыточный риск Для ответа на вопрос о концентрации потерь

$$\left\|\bar{\boldsymbol{\mu}}^{\beta}(\mathbf{Y}) - \boldsymbol{\mu}\right\|_2^2$$

агрегированной по методу экспоненциального взвешивания оценки около риска оракула нам надо ввести величину, которая, в отличие от разности

$$\mathbf{E}_{\boldsymbol{\mu}} \left\|\bar{\boldsymbol{\mu}}^{\beta}(\mathbf{Y}) - \boldsymbol{\mu}\right\|_2^2 - r^{\mathcal{H}}(\boldsymbol{\mu}), \tag{2.1}$$

будет отслеживать эту концентрацию. В качестве такой величины мы возьмем избыточный риск

$$\Delta^{\beta}(\boldsymbol{\mu}) \stackrel{\text{def}}{=} \mathbf{E}_{\boldsymbol{\mu}} \left[\left\|\bar{\boldsymbol{\mu}}^{\beta}(\mathbf{Y}) - \boldsymbol{\mu}\right\|_2^2 - r^{\mathcal{H}}(\boldsymbol{\mu})\right]_+ .$$

Семейство упорядоченных проекционных оценок Мы будем работать со следующим упорядоченным семейством $\mathcal{H}_{\Pi}$ проекционных оценок:

$$\mathcal{H}_{\Pi} = \left\{\mathbf{h}^m \,\middle|\, \quad h_k = \mathbf{1}\{k \le m\} \quad \right\}_{m\in\mathbb{N}}, \tag{2.2}$$

где $\mathbf{1}\{k \le m\}$ – индикатор того, что $k \le m$. Данное семейство имеет вид (9) и удовлетворяет условиям упорядоченности (R1), (R2). Поэтому для него остаются справедливыми наши «ориентиры», с которыми можно будет сравнивать полученный результат, – теоремы 1 и 3.

Заметим, что из концентрационного неравенства Кнайпа (1.10), вообще говоря, следует более точное неравенство, чем в теореме 1. Именно, утверждение этой теоремы останется верной, если заменить разность рисков (2.1) на избыточный риск $\Delta^{\beta}(\boldsymbol{\mu})$. Мы покажем, что сходное утверждение выполнено при любых $\beta \ge 0$.

2.1. Неравенство концентрации

Основным результатом данной работы является следующая

Теорема 4. *При всех* $\beta \geq 0$*, априорных весах* $\boldsymbol{\pi}^{\mathbf{h}} \equiv 1$ *и при всех* $\boldsymbol{\mu} \in \ell^2$ *выполнено неравенство:*

$$\Delta^{\beta}(\boldsymbol{\mu}) \leq \mathrm{K}\sigma^2 \left[\frac{r^{\mathcal{H}_{\Pi}}(\boldsymbol{\mu})}{\sigma^2} + 4\beta L_{\beta} \left(\frac{r^{\mathcal{H}_{\Pi}}(\boldsymbol{\mu})}{\sigma^2} \right) \right]^{1/2} + 4\beta\sigma^2 L_{\beta} \left(\frac{r^{\mathcal{H}_{\Pi}}(\boldsymbol{\mu})}{\sigma^2} \right). \tag{2.3}$$

В нем

- $L_{\beta}(r) = \ln\left(\sqrt{r} + \mathrm{C} + 2\sqrt{\beta} + \Psi_{\beta}(r)\right)$*;*
- $\mathrm{K}, \mathrm{C} > 0$ – *универсальные константы;*
- $\Psi_{\beta}(r)$ – *такая функция, для* $\mathrm{C}_* = \mathrm{e}^{(\mathrm{e}+1)/\mathrm{e}}$

$$\Psi_{\beta}(r) \leq \begin{cases} 6\sqrt{2r}\beta \Big/ \ln\left(\dfrac{2\sqrt{2r}\beta}{\mathrm{C}_*}\right), & \text{если} \quad \ln\left(\dfrac{2\sqrt{2r}\beta}{\mathrm{C}_*}\right) \geq 1 \vee 8\beta, \\ 3\mathrm{e}^{1\vee 8\beta}\mathrm{C}_*, & \text{если} \quad \ln\left(\dfrac{2\sqrt{2r}\beta}{\mathrm{C}_*}\right) < 1 \vee 8\beta. \end{cases}$$

Важность и новизна полученных результатов Насколько нам известно, теорема 4 – первый результат, отвечающий на вопрос о скорости концентрации потерь около риска оракула, для оценки, агрегированной методом экспоненциального взвешивания. Из полученного неравенства видно, что при $\beta \asymp 1$ избыточный риск имеет, как и в теореме 1, порядок корня из оракульного риска, то есть относительно мал, если этот риск велик, и имеет тот же порядок, если оракульный риск $\asymp 1$. Нами продемонстрирован следующий феномен: разница риска агрегированной оценки и оракульного риска имеет гораздо меньший порядок – логарифма из риска оракула – чем характерные потери этой оценки! Можно сформулировать это по-другому: потери оценки, полученной с помощью метода экспоненциального взвешивания, концентрируются около риска оракула гораздо медленнее, чем можно было бы ожидать после теоремы 3.

Кроме того, из данной в теореме оценки функции $\Psi_{\beta}(r)$ нетрудно получить, что при фиксированном риске оракула в пределе $\beta \to 0$ зависящие от β слагаемые в правой части неравенства (2.3) обнуляются; таким образом, мы альтернативным способом (без использования неравенства (1.10)) доказываем «концентрационную» версию теоремы 1, в которой разность рисков (2.1) заменена на избыточный риск $\Delta^{\beta}(\boldsymbol{\mu})$.

Наконец, важно отметить, что в теореме 4 впервые в литературе характеризуется случай $0 < \beta < 2$, что позволяет ответить на второй из поставленных вопросов. Из результатов теоремы 4 видно, что при $\beta \lesssim 10$ метод не сильно чувствителен к выбору β, и классический метод минимизации эмпирического риска, по сути, не уступает методу экспоненциального взвешивания.

2.2. Численный эксперимент

Чтобы проиллюстрировать теорему 4, был проведен численный эксперимент. Для значений A, выбранных достаточно близко в диапазоне $[0,50]$, было сгенерировано по $N = 8000$ статистически независимых реплик наблюдений

$$\mathbf{Y} = \boldsymbol{\mu}(A) + \boldsymbol{\xi} \in \mathbb{R}^{300}, \quad \xi \sim \mathcal{N}(0, I).$$

$\boldsymbol{\mu}(A)$ выбирался независимо от $\boldsymbol{\xi}$ из гауссовского распределения с независимыми компонентами, $\mathbf{E}\mu_k(A) = 0$ и $\mathbf{E}\mu_k^2(A) = A\exp\left(-k^2/5000\right)$. После этого оракульный риск

$$r^{\mathcal{H}_\Pi}(\boldsymbol{\mu}(A)) = \min_{1\le m\le 300}\left\{m + \sum_{k=m+1}^{300}\mu_k^2(A)\right\}$$

усреднялся по методу Монте-Карло:

$$r^{\mathcal{H}_\Pi}(A) = \mathbf{E}_N\left[r^{\mathcal{H}_\Pi}(\boldsymbol{\mu}(A))\right].$$

Кроме того, подсчитывался усредненный избыточный риск

$$\Delta^{\mathcal{H}_\Pi}(A) = \mathbf{E}_N\left[\left\|\bar{\boldsymbol{\mu}}^{\beta}(\mathbf{Y}) - \boldsymbol{\mu}(A)\right\|_2^2 - r_A^{\mathcal{H}_\Pi}\right]_+.$$

для $\beta \in \{0,1,4\}$.

Результаты симуляций показаны на рис. 2.1. Из них можно сделать следующие выводы.

– Отчетливо видна зависимость $\Delta^{\mathcal{H}_\Pi} \sim \mathrm{K}\sqrt{r^{\mathcal{H}_\Pi}}$ для больших значений риска оракула.

– Ни одно значение β не дает равномерно лучший избыточный риск, однако неплохим выбором на практике может стать $\beta = 1$.

– Для небольших значений риска оракула экспоненциальное взвешивание для $\beta \in [1,4]$ оказывается существенно лучше классического метода минимизации эмпирического риска.

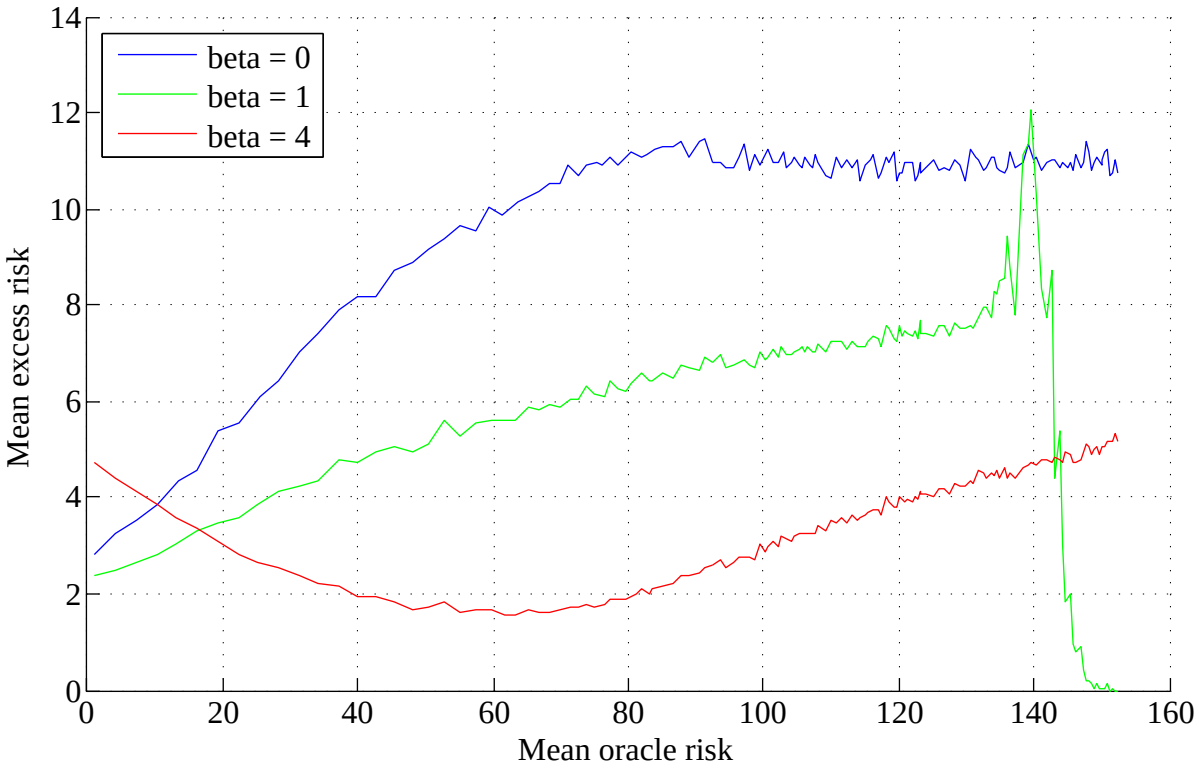

Рис. 2.1. Избыточный риск для различных значений риска оракула.

Глава 3

Доказательства

3.1. Предварительные замечания

Перед тем как приступать к доказательству теоремы 4, приведем некоторые простые факты о семействе упорядоченных проекционных оценок $\mathcal{H}_\Pi$. Введем краткое обозначение $\hat{\boldsymbol{\mu}}^m$ для оценок $\hat{\boldsymbol{\mu}}^{\mathbf{h}^m}(\mathbf{Y}) = \mathbf{h}^m \cdot \mathbf{Y}$ из этого семейства, где компоненты вектора $\mathbf{h}^m$ равны $h_k^m = \mathbf{1}\{k \le m\}$. Кроме того, обозначим

$$r_m(\boldsymbol{\mu}) = R(\hat{\boldsymbol{\mu}}^{\mathbf{h}^m}, \boldsymbol{\mu}), \qquad w_m(\mathbf{Y}) = w^{\mathbf{h}^m}(\mathbf{Y}).$$

– Риск оценки $\hat{\boldsymbol{\mu}}^m$ дается выражением

$$r_m(\boldsymbol{\mu}) = \|(\mathbf{1} - \mathbf{h}^m) \cdot \boldsymbol{\mu}\|_2^2 + \sigma^2 \|\mathbf{h}^m\|_2^2 = \sum_{k=1}^{\infty} \left(1 - h_k^m\right) \mu_k^2 + \sigma^2 h_k^m. \tag{3.1}$$

Как несложно видеть, $\forall\, m \in \mathbb{N} \quad r_m(\boldsymbol{\mu}) \ge \sigma^2$. Отметим, что для фиксированного $\boldsymbol{\mu}$ число

$$m^\circ(\boldsymbol{\mu}) \stackrel{\text{def}}{=} \operatorname*{argmin}_{m \in \mathbb{N}} r_m(\boldsymbol{\mu}) = \operatorname*{argmin}_{m \in \mathbb{N}} \left\{ \sum_{k=m+1}^{\infty} \mu_k^2 + \sigma^2 m \right\},$$

отвечающее выбору оракула $\mathbf{h}^\circ(\boldsymbol{\mu}) = \mathbf{h}^{m^\circ(\boldsymbol{\mu})}$, можно эвристически проинтерпретировать как эффективную размерность задачи. Чтобы продемонстрировать это, перепишем $m^\circ(\boldsymbol{\mu})$ в виде

$$m^\circ(\boldsymbol{\mu}) = \operatorname*{argmin}_{m \in \mathbb{N}} \left\{ \sum_{k=1}^{m} \left(\sigma^2 - \mu_k^2\right) \right\} = \operatorname*{argmin}_{m \in \mathbb{N}} \left\{ \sum_{k=1}^{m} \left(1 - \frac{\mu_k^2}{\sigma^2}\right) \right\} \tag{3.2}$$

и усилим на секунду требование $\boldsymbol{\mu} \in \ell^2$, предположив, что последовательность $\left\{\mu_m^2\right\}$ не возрастает. Тогда из (3.2) следует, что

$$m^\circ(\boldsymbol{\mu}) = \max \left\{ m \,\middle|\, \forall k \le m \quad \frac{\mu_k^2}{\sigma^2} \ge 1 \right\},$$

то есть $m^\circ(\boldsymbol{\mu})$ – число первых компонент сигнала $\boldsymbol{\mu}$, для которых квадрат отношения сигнал-шум больше единицы.

– Из (1.8) и (3.1) легко получить, что величина

$$\begin{aligned} r(\mathbf{Y}, \mathbf{h}^m) &\stackrel{\text{def}}{=} \|\hat{\boldsymbol{\mu}}^m\|_2^2 - 2\langle \mathbf{Y}, \hat{\boldsymbol{\mu}}^m \rangle + 2\sigma^2 \sum_{k \in \mathbb{N}} h_k^m \\ &= \sum_{k=1}^{\infty} -Y_k^2 h_k^m + 2\sigma^2 h_k^m \end{aligned} \tag{3.3}$$

является несмещенной оценкой риска оценки $\hat{\boldsymbol{\mu}}^m$ с точностью до постоянной $\|\boldsymbol{\mu}\|_2^2$, одинаковой для всех оценок:

$$\mathbf{E}\, r(\mathbf{Y}, \mathbf{h}^m) = r_m(\boldsymbol{\mu}) - \|\boldsymbol{\mu}\|_2^2 .$$

При этом случайная ошибка оценивания риска равна

$$r(\mathbf{Y}, \mathbf{h}^m) - r_m(\boldsymbol{\mu}) = -\|\boldsymbol{\mu}\|_2^2 - 2\sigma \sum_{k=1}^{\infty} h_k^m \mu_k \xi_k + \sigma^2 \sum_{k=1}^{\infty} h_k^m (1 - \xi_k^2). \tag{3.4}$$

3.2. Вспомогательные утверждения

Доказательство теоремы 4 будет опираться на несколько лемм. Первая из них широко известна, и потому ее доказательство мы опускаем; доказательство одного из неравенств леммы дано в [27].

Лемма 1. *Пусть $\{\xi_k\}_{k=1}^{\infty}$ – последовательность независимых случайных величин, распределенных по закону $\mathcal{N}(0,1)$, а $\boldsymbol{\mu}$ – произвольный вектор в l^2. Тогда*

$$\mathbf{E}\left[\max_{m\in\mathbb{N}}\left\{\sum_{k=1}^{m}(\xi_k^2 - 1) - U(\alpha)m\right\}\right]_+ \le \frac{1}{\alpha},$$
$$\mathbf{E}\left[\max_{m\in\mathbb{N}}\left\{\sum_{k=1}^{m}(1 - \xi_k^2) - U_*(\alpha)m\right\}\right]_+ \le \frac{1}{\alpha},$$
$$\mathbf{E}\left[\max_{m\in\mathbb{N}}\left\{\sum_{k=m}^{\infty}\xi_k\mu_k - \frac{\alpha}{2}\sum_{k=m}^{\infty}\mu_k^2\right\}\right]_+ \le \frac{1}{\alpha},$$

где функции $U(\alpha)$ и $U_*(\alpha)$

$$U(\alpha) = -\frac{\alpha + \ln(1-2\alpha)/2}{\alpha}, \quad U_*(\alpha) = \frac{\alpha - \ln(1+2\alpha)/2}{\alpha}.$$

Следующий факт – простое аналитическое следствие из вида функций $U(\alpha)$ и $U_*(\alpha)$; его доказательство также можно найти в [27].

Лемма 2. *Для функций $U^{-1}(\nu)$, $U_*^{-1}(\nu)$, обратных, соответственно, к функциям $U(\alpha)$, $U_*(\alpha)$, справедливы оценки:*

$$U^{-1}(\nu) \ge \frac{\nu}{1+2\nu}, \quad U_*^{-1}(\nu) \ge \nu.$$

Нам также понадобится техническое утверждение, доказанное в [27].

Лемма 3. *Пусть* $\{p_k\}_{k=1}^K$ *и* $\{q_k\}_{k=1}^\infty$ *– неотрицательные числовые последовательности, причем*

$$q_1 = 1,\ q_k \le \exp\left[-\rho(k-2)-1\right],\ k = 2,\ldots,\quad \rho > 0.$$

Обозначим

$$w_k = \frac{p_k}{P+Q}\mathbf{1}\{k \le K\} + \frac{q_{k-K}}{P+Q}\mathbf{1}\{k > K\},$$

где

$$P = \sum_{k=1}^{K} p_k, \quad Q = \sum_{k=1}^{\infty} q_k \ge 1.$$

Тогда

$$H(\mathbf{w}) \stackrel{\text{def}}{=} \sum_{k=1}^{\infty} w_k \ln\frac{1}{w_k} \le \ln\Big[K + \mathrm{e}^{R(\rho)}\Big],$$

где

$$\begin{aligned} R(\rho) =& \frac{2(1+\rho)}{\mathrm{e}\rho}\mathbf{1}\Big\{\rho \le \frac{1}{\mathrm{e}-1}\Big\} \\ &+ \frac{1+\rho(\mathrm{e}+1)}{\mathrm{e}\rho}\exp\Big[\frac{1-\rho(\mathrm{e}-1)}{1+\rho(\mathrm{e}+1)}\Big]\mathbf{1}\Big\{\rho \ge \frac{1}{\mathrm{e}-1}\Big\}. \end{aligned} \tag{3.5}$$

Лемма 3 контролирует изменение энтропии при переходе от конечного распределения вероятностей $\{p_k/P\}_1^K$, энтропия которого ограничена $\log_2 K$, к смеси этого распределения и другого, $\{q_k/Q\}_1^\infty$, экспоненциально быстро убывающего. За добавление $\{q_k/Q\}$ мы платим штраф в виде $\exp R(\rho)$ под знаком логарифма, где ρ – скорость убывания $\{q_k\}$.

Наконец, нам понадобится следующее несложное утверждение.

Лемма 4. *Определим для* $x, \beta, \mathrm{C}_* > 0$ *функцию*

$$\Phi_\beta(x) = \min_{0<\epsilon<1/(4\beta)} \left\{x\epsilon + \mathrm{C}_*\mathrm{e}^{1/\epsilon}\right\}.$$

Для $\Phi_\beta(x)$ *выполнены неравенства*

$$\Phi_\beta(x) \le \begin{cases} \dfrac{3x}{\ln(x/\mathrm{C}_*)}, & \text{если} \quad \ln(x/\mathrm{C}_*) \ge 1 \vee 8\beta; \\ 3e^{1\vee 8\beta}\mathrm{C}_*, & \text{если} \quad \ln(x/\mathrm{C}_*) < 1 \vee 8\beta. \end{cases}$$

Доказательство Рассмотрим вначале случай $\ln(x/\mathrm{C}_*) \ge 1 \vee 8\beta$. Выберем

$$\epsilon(x) = \frac{2}{\ln(x/\mathrm{C}_*)} \le \frac{1}{4\beta}$$

и ограничим

$$\Phi_\beta(x) \le x\left(\frac{2}{\ln(x/\mathrm{C}_*)} + \frac{1}{\sqrt{x/\mathrm{C}_*}}\right).$$

Используя неравенство $\sqrt{t} > \ln t$ при $t > 0$, получаем, как и требовалось, $\Phi_\beta(x) \le 3x/\ln(x/\mathrm{C}_*)$.

В случае $\ln(x/\mathrm{C}_*) < 1 \vee 8\beta$ выберем

$$\epsilon(x) \equiv \frac{2}{1 \vee 8\beta} \le \frac{1}{4\beta},$$

учтем, что $x < \mathrm{C}_* \mathrm{e}^{1 \vee 8\beta}$, и получим требуемое неравенство. ∎

3.3. Доказательство неравенства концентрации

Приступим к доказательству теоремы 4. Будем для краткости писать $\mathbf{E}$ вместо $\mathbf{E}_{\boldsymbol{\mu}}$, имея в виду всегда математическое ожидание при фиксированном $\boldsymbol{\mu}$.

Итак, наша конечная цель – оценить сверху избыточный риск $\Delta^\beta(\boldsymbol{\mu})$. Начнем с верхнего ограничения на потери $\left\|\bar{\boldsymbol{\mu}}^\beta(\mathbf{Y}) - \boldsymbol{\mu}\right\|^2$, справедливого в силу выпуклости квадратичной функции:

$$\begin{aligned} \left\|\bar{\boldsymbol{\mu}}^\beta(\mathbf{Y}) - \boldsymbol{\mu}\right\|_2^2 &= \left\|\sum_{m\in\mathbb{N}} w_m(\mathbf{Y})\left[\hat{\boldsymbol{\mu}}^m(\mathbf{Y}) - \boldsymbol{\mu}\right]\right\|_2^2 \\ &\le \sum_{m\in\mathbb{N}} w_m(\mathbf{Y})\left\|\hat{\boldsymbol{\mu}}^m(\mathbf{Y}) - \boldsymbol{\mu}\right\|_2^2. \end{aligned} \tag{3.6}$$

Очевидно, что если для двух случайных величин ξ, η справедливо $\xi \le \eta$, то $\mathbf{E}\left[\xi\right]_+ \le \mathbf{E}\left[\eta\right]_+$, поэтому

$$\Delta^\beta(\boldsymbol{\mu}) \le \mathbf{E}\left[\bar{\mathcal{L}}(\mathbf{Y}) - r^{\mathcal{H}_\Pi}(\boldsymbol{\mu})\right]_+, \tag{3.7}$$

где

$$\bar{\mathcal{L}}(\mathbf{Y}) \stackrel{\text{def}}{=} \sum_{m\in\mathbb{N}} w_m(\mathbf{Y})\left\|\hat{\boldsymbol{\mu}}^{\boldsymbol{m}}(\mathbf{Y}) - \boldsymbol{\mu}\right\|_2^2$$

– экспоненциально агрегированные потери оценок $\hat{\mu}^m$.

Исследуем далее случайную величину $\bar{\mathcal{L}}(\mathbf{Y}) - r^{\mathcal{H}_\Pi}(\boldsymbol{\mu})$. Начнем с преобразования потерь $\|\hat{\boldsymbol{\mu}}^m(\mathbf{Y}) - \boldsymbol{\mu}\|_2^2$ фиксированной оценки:

$$\begin{aligned} \left\|\hat{\boldsymbol{\mu}}^m(\mathbf{Y}) - \boldsymbol{\mu}\right\|_2^2 &= \sum_{k=1}^\infty \left[Y_k h_k^m - \mu_k\right]^2 = \sum_{k=1}^\infty \left[\left(\mu_k + \sigma\xi_k\right) h_k^m - \mu_k\right]^2 \\ &= \sum_{k=1}^\infty \left[\mu_k\left(h_k^m - 1\right) + \sigma\xi_k h_k^m\right]^2 \\ &= \sum_{k=1}^\infty \mu_k^2\left(1 - h_k^m\right) + \sigma^2\xi_k^2 h_k^m. \end{aligned} \tag{3.8}$$

При последнем переходе мы учли $h_k^m\left(1-h_k^m\right)=0$. Перегруппируем слагаемые в последней части (3.8):

$$\begin{aligned}\|\hat{\boldsymbol{\mu}}^m(\mathbf{Y})-\boldsymbol{\mu}\|_2^2 &= \sum_{k=1}^{\infty}(\mu_k^2+2\sigma\mu_k\xi_k)(1-h_k^m)+\sigma^2\xi_k^2h_k^m-2\sigma\mu_k\xi_k(1-h_k^m)\\ &= \sum_{k=1}^{\infty}\left[Y_k^2(1-h_k^m)-\sigma^2\xi_k^2\right]+2\sigma^2\sum_{k=1}^{m}\xi_k^2-2\sigma\sum_{k=m+1}^{\infty}\mu_k\xi_k.\end{aligned}$$

Сумма слагаемых в квадратных скобках в последней строке равна

$$\begin{aligned}\sum_{k=1}^{\infty}Y_k^2(1-h_k^m)-\sigma^2\xi_k^2 &= \sum_{k=1}^{\infty}-Y_k^2h_k^m+\mu_k^2+2\sigma\mu_k\xi_k\\ &= r(\mathbf{Y},\mathbf{h}^m)-2\sigma^2\sum_{k=1}^{\infty}h_k^m+\|\boldsymbol{\mu}\|_2^2+2\sigma\langle\boldsymbol{\mu},\boldsymbol{\xi}\rangle,\end{aligned}$$

откуда

$$\begin{aligned}\|\hat{\boldsymbol{\mu}}^m(\mathbf{Y})-\boldsymbol{\mu}\|_2^2 &= r(\mathbf{Y},\mathbf{h}^m)+\|\boldsymbol{\mu}\|_2^2+2\sigma\langle\boldsymbol{\mu},\boldsymbol{\xi}\rangle\\ &\quad+2\sigma^2\sum_{k=1}^{m}\left(\xi_k^2-1\right)-2\sigma\sum_{k=m+1}^{\infty}\xi_k\mu_k.\end{aligned}\tag{3.9}$$

Заметим, что слагаемое

$$\tilde{r}(\mathbf{Y},\mathbf{h}^m)\stackrel{\text{def}}{=}r(\mathbf{Y},\mathbf{h}^m)+\|\boldsymbol{\mu}\|_2^2+2\sigma\langle\boldsymbol{\mu},\boldsymbol{\xi}\rangle\tag{3.10}$$

в (3.9) совпадает с $r(\mathbf{Y},\mathbf{h}^m)$ точностью до случайной добавки, одинаковой для всех оценок. Поэтому для любого $m\in\mathbb{N}$ выполнено:

$$\tilde{r}(\mathbf{Y},\mathbf{h}^m)-\tilde{r}(\mathbf{Y},\mathbf{h}^{\hat{m}(\mathbf{Y})})=r(\mathbf{Y},\mathbf{h}^m)-r(\mathbf{Y},\mathbf{h}^{\hat{m}(\mathbf{Y})}),$$

где число

$$\hat{m}(\mathbf{Y})\stackrel{\text{def}}{=}\operatorname*{argmin}_{m\in\mathbb{N}}r(\mathbf{Y},\mathbf{h}^m)$$

соответствует минимальной несмещенной оценке риска. Используя это соображение вместе с (3.9), мы можем оценить $\bar{\mathcal{L}}(\mathbf{Y})$ как

$$\begin{aligned}\bar{\mathcal{L}}(\mathbf{Y}) &\le \tilde{r}(\mathbf{Y},\mathbf{h}^{\hat{m}(\mathbf{Y})})+\sum_{m\in\mathbb{N}}w_m(\mathbf{Y})\left[r(\mathbf{Y},\mathbf{h}^m)-r(\mathbf{Y},\mathbf{h}^{\hat{m}(\mathbf{Y})})\right]\\ &\quad+\sum_{m\in\mathbb{N}}w_m(\mathbf{Y})\left[2\sigma^2\sum_{k=1}^{m}\left(\xi_k^2-1\right)-2\sigma\sum_{k=m+1}^{\infty}\xi_k\mu_k\right].\end{aligned}\tag{3.11}$$

Заметим, что в силу определения $\hat{m}(\mathbf{Y})$ выполнено

$$\tilde{r}(\mathbf{Y},\mathbf{h}^{\hat{m}(\mathbf{Y})})\le\tilde{r}(\mathbf{Y},\mathbf{h}^{m^\circ(\boldsymbol{\mu})}).\tag{3.12}$$

Рассмотрим теперь величину $\tilde{r}(\mathbf{Y}, \mathbf{h}^m)$ для произвольного $m \in \mathbb{N}$. Из (3.4) и (3.10) сразу видно, что

$$\tilde{r}(\mathbf{Y}, \mathbf{h}^m) = r_m(\boldsymbol{\mu}) + 2\sigma \sum_{k=m+1}^{\infty} \xi_k \mu_k + \sigma^2 \sum_{k=1}^{m} (1 - \xi_k^2);$$

в частности, для $m = m^\circ(\boldsymbol{\mu})$ имеем

$$\tilde{r}(\mathbf{Y}, \mathbf{h}^{m^\circ(\boldsymbol{\mu})}) = r^{\mathcal{H}_\Pi}(\boldsymbol{\mu}) + 2\sigma \sum_{k=m^\circ(\boldsymbol{\mu})+1}^{\infty} \xi_k \mu_k + \sigma^2 \sum_{k=1}^{m^\circ(\boldsymbol{\mu})} (1 - \xi_k^2). \tag{3.13}$$

3.3.1. Разбиение на три слагаемых

Объединяя (3.11), (3.12) и (3.13), получаем:

$$\begin{aligned} \bar{\mathcal{L}}(\mathbf{Y}) - r^{\mathcal{H}_\Pi}(\boldsymbol{\mu}) \leq{}& 2\sigma \sum_{k=m^\circ(\boldsymbol{\mu})+1}^{\infty} \xi_k \mu_k + \sigma^2 \sum_{k=1}^{m^\circ(\boldsymbol{\mu})} (1 - \xi_k^2) \\ &+ \sum_{m \in \mathbb{N}} w_m(\mathbf{Y}) \left\{ 2\sigma^2 \sum_{k=1}^{m} \left(\xi_k^2 - 1 \right) - 2\sigma \sum_{k=m+1}^{\infty} \xi_k \mu_k \right\} \\ &+ \sum_{m \in \mathbb{N}} w_m(\mathbf{Y}) \left[r(\mathbf{Y}, \mathbf{h}^m) - r(\mathbf{Y}, \mathbf{h}^{\hat{m}(\mathbf{Y})}) \right]. \end{aligned} \tag{3.14}$$

Поскольку для любой пары случайных величин ξ, η справедливо $\mathbf{E}[\xi + \eta]_+ \leq \mathbf{E}[\xi]_+ + \mathbf{E}[\eta]_+$, то, используя последнее неравенство, мы можем ограничить $\mathbf{E}[\bar{\mathcal{L}}(\mathbf{Y}) - r^{\mathcal{H}_\Pi}(\boldsymbol{\mu})]_+$, а значит и $\Delta^\beta(\boldsymbol{\mu})$, если оценим функционал $\mathbf{E}[\cdot]_+$ для каждого из слагаемых в правой части (3.14).

3.3.2. Оценка первого слагаемого

Функционал $\mathbf{E}[\cdot]_+$ для первого слагаемого можно очень просто оценить сверху с помощью неравенства Коши – Буняковского:

$$\begin{aligned} &\mathbf{E} \left[2\sigma \sum_{k=m^\circ(\boldsymbol{\mu})+1}^{\infty} \xi_k \mu_k + \sigma^2 \sum_{k=1}^{m^\circ(\boldsymbol{\mu})} (1 - \xi_k^2) \right]_+ \\ &\leq \mathbf{E}^{1/2} \left[2\sigma \sum_{k=m^\circ(\boldsymbol{\mu})+1}^{\infty} \xi_k \mu_k + \sigma^2 \sum_{k=1}^{m^\circ(\boldsymbol{\mu})} (1 - \xi_k^2) \right]^2 \\ &= \sigma \left[\sum_{k=1}^{\infty} 4 \left(1 - h_k^{m^\circ(\boldsymbol{\mu})} \right) \mu_k^2 + 2h_k^{m^\circ(\boldsymbol{\mu})} \sigma^2 \right]^{1/2} \leq 2\sigma^2 \sqrt{\frac{r^{\mathcal{H}_\Pi}(\boldsymbol{\mu})}{\sigma^2}}. \end{aligned} \tag{3.15}$$

3.3.3. Оценка второго слагаемого

Второе слагаемое в правой части (3.14) ограничим, используя технику, опирающуюся на лемму 1. Для этого вначале «возмутим» его с помощью произвольного числа $\gamma \geq 0$, сделав следующее преобразование:

$$\begin{aligned}&\sum_{m\in\mathbb{N}} w_m(\mathbf{Y})\left\{2\sigma^2\sum_{k=1}^{m}(\xi_k^2-1)-2\sigma\sum_{k=m+1}^{\infty}\xi_k\mu_k\right\}\\&=\sum_{m\in\mathbb{N}} w_m(\mathbf{Y})\left\{2\sigma^2\left[\sum_{k=1}^{m}(\xi_k^2-1)-\gamma m\right]+2\sigma^2\gamma m-2\sigma\sum_{k=m+1}^{\infty}\xi_k\mu_k\right\}.\end{aligned} \tag{3.16}$$

Будем теперь работать со вторым слагаемым в квадратных скобках в последней строке (3.16):

$$\sum_{m\in\mathbb{N}} w_m(\mathbf{Y})\sigma^2 m=\bar{\mathcal{L}}(\mathbf{Y})-\sum_{m\in\mathbb{N}} w_m(\mathbf{Y})\left[\sigma^2\sum_{k=1}^{m}(\xi_k^2-1)+\sum_{k=m+1}^{\infty}\mu_k^2\right]. \tag{3.17}$$

Здесь мы воспользовались (3.8), дополнив $\sigma^2 m$ до $\|\hat{\boldsymbol{\mu}}^m(\mathbf{Y})-\boldsymbol{\mu}\|_2^2$. Объединив (3.16) и (3.17), а затем оценив сверху взвешенное среднее максимальным значением, получаем:

$$\begin{aligned}&\sum_{m\in\mathbb{N}} w_m(\mathbf{Y})\left\{2\sigma^2\sum_{k=1}^{m}\left(\xi_k^2-1\right)-2\sigma\sum_{k=m+1}^{\infty}\xi_k\mu_k\right\}\\&\leq 2\gamma\bar{\mathcal{L}}(\mathbf{Y})+2\sigma^2(1-\gamma)\max_{u\in\mathbb{N}}\left\{\sum_{k=1}^{u}(\xi_k^2-1)-\frac{\gamma}{1-\gamma}u\right\}\\&\qquad+2\sigma\max_{v\in\mathbb{N}}\left\{\sum_{k=v+1}^{\infty}-\xi_k\mu_k-\frac{\gamma}{\sigma}\mu_k^2\right\}\end{aligned} \tag{3.18}$$

Прибавим и вычтем $2\gamma r^{\mathcal{H}_\Pi}(\boldsymbol{\mu})$ из обеих частей (3.18). После этого, проконтролировав максимумы в правой части (3.18) с помощью лемм 1 и 2, получаем неравенство:

$$\begin{aligned}&\mathbf{E}\left[\sum_{m\in\mathbb{N}} w_m(\mathbf{Y})\left\{2\sigma^2\sum_{k=1}^{m}\left(\xi_k^2-1\right)-2\sigma\sum_{k=m+1}^{\infty}\xi_k\mu_k\right\}\right]_+\\&\leq 2\gamma\mathbf{E}\left[\bar{\mathcal{L}}(\mathbf{Y})-r^{\mathcal{H}_\Pi}(\boldsymbol{\mu})\right]_+ + 2\gamma r^{\mathcal{H}_\Pi}(\boldsymbol{\mu})+\sigma^2\left[\frac{2(1-\gamma^2)}{\gamma}+\frac{1}{\gamma}\right].\end{aligned}$$

Минимизируя правую часть этого неравенства по $\gamma \geq 0$, мы приходим к неравенству

$$\begin{aligned}&\mathbf{E}\left[\sum_{m\in\mathbb{N}} w_m(\mathbf{Y})\left\{2\sigma^2\sum_{k=1}^{m}\left(\xi_k^2-1\right)-2\sigma\sum_{k=m+1}^{\infty}\xi_k\mu_k\right\}\right]_+\\&\leq 2\sqrt{6}\sigma^2\left[\frac{1}{\sigma^2}\mathbf{E}\left[\bar{\mathcal{L}}(\mathbf{Y})-r^{\mathcal{H}_\Pi}(\boldsymbol{\mu})\right]_+ + \frac{r^{\mathcal{H}_\Pi}(\boldsymbol{\mu})}{\sigma^2}-1\right]^{1/2}.\end{aligned} \tag{3.19}$$

3.3.4. Оценка третьего слагаемого

Перейдем к оценке третьего слагаемого в (3.14). Докажем, что для $\beta > 0$ в обозначениях теоремы справедливо неравенство

$$\mathbf{E} \sum_{m \in \mathbb{N}} w_m(\mathbf{Y}) \left[r(\mathbf{Y}, \mathbf{h}^m) - r(\mathbf{Y}, \mathbf{h}^{\hat{m}(\mathbf{Y})}) \right] \leq 4\beta\sigma^2 L_\beta \left(\frac{r^{\mathcal{H}_\Pi}(\boldsymbol{\mu})}{\sigma^2} \right), \tag{3.20}$$

где функция L_β дана в формулировке доказываемой теоремы, а оператор $[\cdot]_+$ опущен в силу того, что аргумент математического ожидания неотрицателен). Для краткости будем впредь писать $\hat{m}$ вместо $\hat{m}(\mathbf{Y})$.

В основе дальнейшего доказательства будут лежать приемы, использованные в [27] для доказательства аналогичного неравенства в частном случае $\beta \geq 2$. Ключевую роль в нем играет следующий замечательный факт: оказывается, что с полной вероятностью после конечного случайного момента, в среднем имеющего порядок отношения сигнал-шум $r^{\mathcal{H}_\Pi}(\boldsymbol{\mu})/\sigma^2$, несмещенные оценки рисков начинают линейно расти с m.

Наша идея состоит в том, чтобы свести доказательство к нестохастической лемме 3. Для этого фиксируем произвольное число $\epsilon > 0$ и рассмотрим случайный момент

$$M_\epsilon = \max \left\{ m \,\middle|\, r(\mathbf{Y}, \mathbf{h}^m) - r(\mathbf{Y}, \mathbf{h}^{\hat{m}}) \leq 2\beta\epsilon\sigma^2 \left[m - \hat{m} \right] + 2\beta\sigma^2 \right\}, \tag{3.21}$$

делящий множество $\{r(\mathbf{Y}, \mathbf{h}^m)\}_{m \in \mathbb{N}}$ на два подмножества. На первом из них, для которого $m \leq M_\epsilon$, поведение несмещенных оценок $r(\mathbf{Y}, \mathbf{h}^m)$ в целом случайно, однако на втором, для которого $m > M_\epsilon$, случайность наблюдается в существенно меньшей мере, поскольку

$$r(\mathbf{Y}, \mathbf{h}^m) > r(\mathbf{Y}, \mathbf{h}^{\hat{m}}) + 2\beta\epsilon\sigma^2 [m - \hat{m}] + 2\beta\sigma^2, \quad m > M_\epsilon.$$

Теперь, логарифмируя $w_m(\mathbf{Y})$, находим

$$2\beta\sigma^2 \ln \frac{1}{w_m(\mathbf{Y})} = r(\mathbf{Y}, \mathbf{h}^m) - r(\mathbf{Y}, \mathbf{h}^{\hat{m}})$$
$$+2\beta\sigma^2 \ln \left\{ \sum_{l \in \mathbb{N}} \exp \left[- \frac{r(\mathbf{Y}, \mathbf{h}^l) - r(\mathbf{Y}, \mathbf{h}^{\hat{m}})}{2\beta\sigma^2} \right] \right\},$$

где слагаемое во второй строке положительно в силу того, что по крайней мере одна из экспонент в сумме под логарифмом имеет нулевой показатель. Отсюда имеем неравенство

$$\sum_{m \in \mathbb{N}} w_m(\mathbf{Y}) \left[r(\mathbf{Y}, \mathbf{h}^m) - r(\mathbf{Y}, \mathbf{h}^{\hat{m}}) \right] \leq 2\beta\sigma^2 \sum_{m \in \mathbb{N}} w_m(\mathbf{Y}) \ln \frac{1}{w_m(\mathbf{Y})}.$$

Ограничив энтропию весов $\{w_m(\mathbf{Y})\}_{m\in\mathbb{N}}$ с помощью определения (3.21), леммы 3 и вогнутости логарифма, получаем

$$\mathbf{E}\sum_{m\in\mathbb{N}} w_m(\mathbf{Y})\left[r(\mathbf{Y},\mathbf{h}^m)-r(\mathbf{Y},\mathbf{h}^{\hat{m}})\right]\le 2\beta\sigma^2\ln\left[\mathbf{E}M_\epsilon-1+\mathrm{e}^{R(\epsilon)}\right]. \tag{3.22}$$

Чтобы завершить доказательство (3.20), надо ограничить величину $\mathbf{E}M_\epsilon$. Воспользовавшись (3.21), (3.1) и (3.4), получаем

$$M_\epsilon=\max\Bigg\{m\Bigg|\quad\sum_{k=m+1}^{\infty}\mu_k^2+\sigma^2(1-2\beta\epsilon)m+2\sigma\sum_{k=m+1}^{\infty}\mu_k\xi_k-\sigma^2\sum_{k=1}^{m}(\xi_k^2-1)\le$$
$$\le\sum_{k=\hat{m}+1}^{\infty}\mu_k^2+\sigma^2(1-2\beta\epsilon)\hat{m}+2\sigma\sum_{k=\hat{m}+1}^{\infty}\mu_k\xi_k-\sigma^2\sum_{k=1}^{\hat{m}}(\xi_k^2-1)+2\beta\sigma^2\Bigg\}.$$

Перепишем последнее определение, прибавив и отняв из левой части неравенства $\gamma r_m(\boldsymbol{\mu})=\gamma\sum_{k=m+1}^{\infty}\mu_k^2+\gamma\sigma^2 m$, а из правой $\gamma r_{\hat{m}}(\boldsymbol{\mu})=\gamma\sum_{k=\hat{m}+1}^{\infty}\mu_k^2+\gamma\sigma^2\hat{m}$, где $\gamma\in(0,1-2\beta\epsilon)$ фиксировано:

$$\begin{aligned}M_\epsilon=\max\Bigg\{m\Bigg|\quad&(1-\gamma)\sum_{k=m+1}^{\infty}\mu_k^2+\sigma^2(1-2\beta\epsilon-\gamma)m\\&+\gamma\sum_{k=m+1}^{\infty}\mu_k^2+2\sigma\sum_{k=m+1}^{\infty}\mu_k\xi_k-\sigma^2\sum_{k=1}^{m}(\xi_k^2-1)+\gamma\sigma^2 m\\&\le(1+\gamma)\sum_{k=\hat{m}+1}^{\infty}\mu_k^2+\sigma^2(1+\gamma)\hat{m}\\&\quad-\gamma\sum_{k=\hat{m}+1}^{\infty}\mu_k^2+2\sigma\sum_{k=\hat{m}+1}^{\infty}\mu_k\xi_k\\&\quad-\sigma^2\sum_{k=1}^{\hat{m}}(\xi_k^2-1)-(\gamma+2\beta\epsilon)\sigma^2\hat{m}+2\beta\sigma^2\Bigg\}.\end{aligned}$$

Неравенство под максимумом можно ослабить, получив

$$\begin{aligned}M_\epsilon\le\max\Bigg\{m\Bigg|\quad&\sigma^2(1-2\beta\epsilon-\gamma)m\\&+\min_{s\ge1}\left[\gamma\sum_{k=s+1}^{\infty}\mu_k^2+2\sigma\sum_{k=s+1}^{\infty}\mu_k\xi_k\right]+\sigma^2\min_{s\ge1}\left[\sum_{k=1}^{s}(1-\xi_k^2)+\gamma s\right]\\&\le(1+\gamma)\left[\sum_{k=\hat{m}+1}^{\infty}\mu_k^2+\sigma^2\hat{m}\right]\\&\quad+\max_{s\ge1}\left[-\gamma\sum_{k=s+1}^{\infty}\mu_k^2+2\sigma\sum_{k=s+1}^{\infty}\mu_k\xi_k\right]\\&\quad+\sigma^2\max_{s\ge1}\left[\sum_{k=1}^{s}(1-\xi_k^2)-(\gamma+2\beta\epsilon)s\right]+2\beta\sigma^2\Bigg\}.\end{aligned}$$

Максимумы и минимумы в этом выражении проконтролируем с помощью лемм 1 и 2, в результате чего после несложных преобразований получим

$$(1-2\beta\epsilon-\gamma)\sigma^2\mathbf{E}M_\epsilon \le (1+\gamma)\mathbf{E}\left[\sum_{k=\hat{m}+1}^{\infty}\mu_k^2+\sigma^2\hat{m}\right] + \frac{6\sigma^2}{\gamma}+2(\beta+1)\sigma^2. \tag{3.23}$$

Нетрудно проверить с помощью простых вычислений, что при любых $G, Z > 0$

$$\min_{\gamma\in[0,G]}\left\{\frac{\sigma^2}{\gamma}+\gamma Z\right\} \le \frac{\sigma^2}{G}+2\sigma\sqrt{Z}. \tag{3.24}$$

Переписав неравенство (3.23) в ввиде

$$\begin{aligned}(1-2\beta\epsilon)\sigma^2\mathbf{E}M_\epsilon &\le \mathbf{E}\left[\sum_{k=\hat{m}+1}^{\infty}\mu_k^2+\sigma^2\hat{m}\right]\\ &+\frac{6\sigma^2}{\gamma}+\gamma\left\{\sum_{k=\hat{m}+1}^{\infty}\mu_k^2+\sigma^2\hat{m}+\sigma^2\mathbf{E}M_\epsilon\right\}+2(\beta+1)\sigma^2,\end{aligned} \tag{3.25}$$

и применив (3.24), найдем

$$\begin{aligned}(1-2\beta\epsilon)\sigma^2\mathbf{E}M_\epsilon &\le \mathbf{E}\left[\sum_{k=\hat{m}+1}^{\infty}\mu_k^2+\sigma^2\hat{m}\right]\\ &+2\sqrt{6}\sigma\left[\sum_{k=\hat{m}+1}^{\infty}\mu_k^2+\sigma^2\hat{m}+\sigma^2\mathbf{E}M_\epsilon\right]^{1/2}+\left[2(\beta+1)+\frac{6}{1-2\beta\epsilon}\right]\sigma^2\\ &\le \mathbf{E}\left[\sum_{k=\hat{m}+1}^{\infty}\mu_k^2+\sigma^2\hat{m}\right]+2\sqrt{6}\sigma\left[\sum_{k=\hat{m}+1}^{\infty}\mu_k^2+\sigma^2\hat{m}\right]^{1/2}\\ &+2\sqrt{6}\sigma^2\mathbf{E}^{1/2}M_\epsilon+\left[2(\beta+1)+\frac{6}{1-2\beta\epsilon}\right]\sigma^2.\end{aligned}$$

Последний переход справедлив в силу того, что для любых двух неотрицательных чисел

$$\sqrt{a+b} \le \sqrt{a}+\sqrt{b}. \tag{3.26}$$

Выделяя в первой и последней частях цепного неравенства полные квадраты, получаем

$$\sigma^2\left\{\sqrt{1-2\beta\epsilon}\sqrt{\mathbf{E}M_\epsilon}-\sqrt{\frac{6}{1-2\beta\epsilon}}\right\}^2 \le \left\{\mathbf{E}^{1/2}\left[\sum_{k=\hat{m}+1}^{\infty}\mu_k^2+\sigma^2\hat{m}\right]+\sigma\sqrt{6}\right\}^2 + \left[2\beta+8\frac{1+\beta\epsilon}{1-2\beta\epsilon}\right]\sigma^2$$

и, воспользовавшись неравенством (3.26), а также ограничением $2\beta\epsilon < 1$, после простых алгебраических преобразований приходим к следующему неравенству:

$$\sqrt{1-2\beta\epsilon}\sigma\sqrt{\mathbf{E}M_\epsilon} \le \mathbf{E}^{1/2}\left[\sum_{k=\hat{m}+1}^{\infty}\mu_k^2+\sigma^2\hat{m}\right]+\left[\sqrt{2\beta}+\frac{\mathrm{C}'}{\sqrt{1-2\beta\epsilon}}\right]\sigma, \tag{3.27}$$

где C' – универсальная константа, которую при желании легко можно восстановить из выкладок.

Оценим теперь сверху математическое ожидание в правой части этого неравенства. Воспользовавшись (3.3) и определением случаной величины $\hat{m}$, находим

$$-\sum_{k=1}^{\hat{m}} Y_k^2 + 2\sigma^2\hat{m} \le -\sum_{k=1}^{m} Y_k^2 + 2\sigma^2 m$$

или, что эквивалентно,

$$\begin{aligned}\sum_{k=\hat{m}+1}^{\infty} \mu_k^2 + \sigma^2\hat{m} + 2\sigma \sum_{k=\hat{m}+1}^{\infty} \mu_k\xi_k - \sigma^2\sum_{k=1}^{\hat{m}}(\xi_k^2-1)\\ \le \sum_{k=m+1}^{\infty} \mu_k^2 + \sigma^2 m + 2\sigma \sum_{k=m+1}^{\infty} \mu_k\xi_k - \sigma^2\sum_{k=1}^{m}(\xi_k^2-1).\end{aligned}$$

Как и ранее, «возмутим» последнее неравенство с помощью произвольного числа $\delta \in (0,1)$:

$$\begin{aligned}(1-\delta)\left[\sum_{k=\hat{m}+1}^{\infty} \mu_k^2 + \sigma^2\hat{m}\right] \le \sum_{k=m+1}^{\infty} \mu_k^2 + \sigma^2 m\\ +2\sigma\sum_{k=m+1}^{\infty}\mu_k\xi_k + \sigma^2\sum_{k=1}^{m}(\xi_k^2-1)\\ +\max_{s\ge 1}\left[-2\sigma\sum_{k=s+1}^{\infty}\mu_k\xi_k - \delta\sum_{k=s+1}^{\infty}\mu_k^2\right]\\ +\sigma^2\max_{s\ge 1}\left[\sum_{k=1}^{s}(\xi_k^2-1) - \delta s\right],\end{aligned}$$

откуда с помощью лемм 1 и 2 найдем

$$(1-\delta)\mathbf{E}\left[\sum_{k=\hat{m}+1}^{\infty} \mu_k^2 + \sigma^2\hat{m}\right] \le \sum_{k=m+1}^{\infty} \mu_k^2 + \sigma^2 m + \sigma^2\left[2+\frac{3}{\delta}\right].$$

Минимизируя правую часть этого неравенства по m, получаем

$$\mathbf{E}\left[\sum_{k=\hat{m}+1}^{\infty} \mu_k^2 + \sigma^2\hat{m}\right] \le r^{\mathcal{H}_\Pi}(\boldsymbol{\mu}) + 2\sigma^2 + \frac{3\sigma^2}{\delta} + \delta\mathbf{E}\left[\sum_{k=\hat{m}+1}^{\infty} \mu_k^2 + \sigma^2\hat{m}\right]. \tag{3.28}$$

С помощью (3.24) минимизируем правую часть (3.28) по δ. Имеем

$$\mathbf{E}\left[\sum_{k=\hat{m}+1}^{\infty} \mu_k^2 + \sigma^2\hat{m}\right] \le r^{\mathcal{H}_\Pi}(\boldsymbol{\mu}) + 5\sigma^2 + 2\sqrt{3}\sigma\mathbf{E}^{1/2}\left[\sum_{k=\hat{m}+1}^{\infty} \mu_k^2 + \sigma^2\hat{m}\right],$$

и отсюда, выделив полный квадрат, сразу получаем

$$\begin{aligned}\mathbf{E}^{1/2}\left[\sum_{k=\hat{m}+1}^{\infty} \mu_k^2 + \sigma^2\hat{m}\right] &\le \sqrt{r^{\mathcal{H}_\Pi}(\boldsymbol{\mu}) + 8\sigma^2} + \sqrt{3}\sigma\\ &\le \sqrt{r^{\mathcal{H}_\Pi}(\boldsymbol{\mu})} + 5\sigma.\end{aligned} \tag{3.29}$$

Подставив теперь (3.29) в (3.27), мы придем к

$$\sqrt{1-2\beta\epsilon}\sigma\sqrt{\mathbf{E}M_\epsilon} \le \sqrt{r^{\mathcal{H}_\Pi}(\boldsymbol{\mu})} + \left[\sqrt{2\beta} + \frac{\mathrm{C}'}{\sqrt{1-2\beta\epsilon}}\right]\sigma, \tag{3.30}$$

Чтобы упростить вычисления, далее будем считать, что

$$0 < \epsilon < \frac{1}{4\beta}.$$

При выполнении этого условия (3.30) можно записать в следующем виде:

$$\sigma\sqrt{\mathbf{E}M_\epsilon} \le \sqrt{r^{\mathcal{H}_\Pi}(\boldsymbol{\mu})} + 2\sqrt{2}\beta\epsilon\sqrt{r^{\mathcal{H}_\Pi}(\boldsymbol{\mu})} + \left(\mathrm{C} + 2\sqrt{\beta}\right)\sigma,$$

где C – универсальная константа. Здесь мы также воспользовались тем, что при $x \in (0,1/2)$

$$\frac{1}{\sqrt{1-x}} \le 1 + \sqrt{2}x.$$

Подставив последнее из полученных неравенств в (3.22) и опять воспользовавшись (3.26), найдем

$$\begin{aligned}\mathbf{E}\sum_{m\in\mathbb{N}} w_m(\mathbf{Y})\left[r(\mathbf{Y},\mathbf{h}^m) - r(\mathbf{Y},\mathbf{h}^{\hat{m}})\right] \le 4\beta\sigma^2 \ln\Bigg[\sqrt{\frac{r^{\mathcal{H}_\Pi}(\boldsymbol{\mu})}{\sigma^2}} + \mathrm{C} + 2\sqrt{\beta} \\ +2\sqrt{2}\beta\epsilon\sqrt{\frac{r^{\mathcal{H}_\Pi}(\boldsymbol{\mu})}{\sigma^2}} + \mathrm{e}^{R(\epsilon)/2}\Bigg].\end{aligned} \tag{3.31}$$

Обозначим

$$\mathrm{C}_* = \mathrm{e}^{(\mathrm{e}+1)/\mathrm{e}}$$
$$r = r^{\mathcal{H}_\Pi}(\boldsymbol{\mu})/\sigma^2.$$

Из (3.5) очевидна оценка

$$\frac{R(\epsilon)}{2} \le \frac{2+\epsilon(\mathrm{e}+1)}{\epsilon\mathrm{e}} < \frac{1}{\epsilon} + \ln \mathrm{C}_*$$

Воспользовавшись этой оценкой и обозначив

$$x = 2\sqrt{2}\beta\sqrt{r},$$

применим лемму 4, чтобы оценить минимум по $0 < \epsilon < 1/(4\beta)$ слагаемого во второй строке (3.31). В результате этого получим, что (3.20) выполнено для $L_\beta(r)$ из формулировки теоремы 4; при этом

$$\Psi_\beta(r) = \min_{0<\epsilon<1/(4\beta)}\left\{2\sqrt{2}\beta\epsilon\sqrt{r} + \mathrm{e}^{R(\epsilon)/2}\right\}.$$

3.3.5. Объединение границ

Наконец, объединим (3.14), (3.15), (3.19) и (3.20), что даст нам следующее неравенство относительно $t = \mathbf{E}\left[\bar{\mathcal{L}}(\mathbf{Y}) - r^{\mathcal{H}_\Pi}(\boldsymbol{\mu})\right]_+ / \sigma^2$:

$$t \leq 2\sqrt{r} + 2\sqrt{6}\sqrt{t + r - 1} + 4\beta L_\beta(r).$$

Выделяя в нем полные квадраты и используя (3.7), мы завершаем доказательство теоремы. ∎

Заключение

В данной работе исследовался важный метод агрегации статистических оценок – метод экспоненциального взвешивания – на примере семейства проекционных оценок с переменной шириной полосы. Известные в литературе оракульные неравенства, связывающие риск агрегированной по этому методу оценки с риском лучшей проекционной оценки, обладают двумя существенными недостатками:

– они не отслеживают концентрацию потерь оценивания;

– они не покрывают важный диапазон значений параметра метода β.

В настоящей работе было сформулировано и доказано оракульное неравенство, лишенное указанных недостатков. Для иллюстрации полученного результата был проведен численный эксперимент, показавший, кроме того, что при умеренных значениях риска оракула экспоненциальное взвешивание превосходит классический метод минимизации эмпирического риска.

Отметим, что семейство проекционных оценок с переменной шириной полосы является важным частным случаем общего класса семейств упорядоченных оценок. В качестве направления будущей работы можно было бы выбрать получение результатов, аналогичных теореме 4, для общего случая упорядоченного семейства. Из доказательства этой теоремы хорошо видно принципиальное отличие семейства проекционных оценок от прочих семейств упорядоченных оценок: исследование данного семейства сводится, по сути, к контролю случайного процесса с независимыми приращениями и потому может быть осуществлено сравнительно легко. В общем случае процесс, который надо проконтролировать, обладает зависимыми приращениями, и для его контроля надо использовать более продвинутый математический аппарат.

Литература

1. Tsybakov A. Introduction to Nonparametric Estimation. Berlin: Springer, 2009. Vol. 11 of Springer Series in Statistics.

2. S. Efromovich. Nonparametric Curve Estimation: Methods, Theory and Applications. Springer Series in Statistics. New York: Springer, 1999.

3. Ибрагимов И. А., Хасьминский Р. З. Об оценке бесконечномерного параметра в гауссовском белом шуме // Докл. АН СССР. 1977. Т. 236. № 5. С. 1053–1055.

4. Ибрагимов И. А., Хасьминский Р. З. Асимптотическая теория оценивания. М.: Наука, 1979.

5. Пинскер М. С. Оптимальная фильтрация квадратично интегрируемого сигнала, наблюдаемого на фоне белого гауссовского шума // Проблемы передачи информации. 1980. Т. 16. С. 120–133.

6. Nussbaum M. Minimax risk: Pinsker bound // Encyclopedia of Statistical Sciences. New York: Wiley, 1999. URL: http://dx.doi.org/10.1002/0471667196.ess1098.

7. Rasmussen C. E., Williams C. K. I. Gaussian Processes for Machine Learning. Cambridge, Massachusetts: MIT Press, 2006. URL: http://www.GaussianProcess.org/gpml.

8. Stein C. Estimation of the mean of a multivariate normal distribution // Proc. Prague Symp. on Asymptotic Statistics. Prague, Czechoslovakia: 1973. Pp. 345–381.

9. Akaike H. Information theory and an extension of the maximum likelihood principle // Proc. 2nd Int. Sympos. on Information Theory. 1971. Pp. 267–281.

10. Mallows C. Some comments on C_p // Technometrics. 1973. Vol. 15. Pp. 661–675.

11. Kneip A. Ordered linear smoothers // Annals of Statistics. 1994. Vol. 22. Pp. 835–866.

12. Speckman P. Spline smoothing and optimal rates of convergence in nonparametric regression models // Annals of Statistics. 1985. — 09. Vol. 13, no. 3. Pp. 970–983. URL: http://dx.doi.org/10.1214/aos/1176349650.

13. Nemirovski A. Topics in Non-Parametric Statistics. Berlin: Springer-Verlag, 2000. Vol. 1738 of Lecture Notes in Mathematics. URL: http://www2.isye.gatech.edu/~nemirovs/Lect_SaintFlour.pdf.

14. Juditsky A., Nemirovski A. Functional aggregation for nonparametric regression // Annals of Statistics. 2000. Vol. 28. Pp. 681–712. URL: http://www2.isye.gatech.edu/~nemirovs/AnnStat_FunctAggr_2000.pdf.

15. Catoni O. Statistical Learning Theory and Stochastic Optimization. Berlin: Springer-Verlag, 2004. Vol. 1851 of Lecture Notes in Mathematics. URL: http://www.math.ens.fr/~catoni/homepage/saintFlourDraft.ps.

16. Yang Y. Combining different procedures for adaptive regression // Journal of Multivariate Analysis. 2000. Vol. 74. Pp. 135–161.

17. Rigollet P., Tsybakov A. Linear and convex aggregation of density estimators // Mathematical Methods of Statistics. 2007. Vol. 16. Pp. 260–280. URL: http://www.ams.org/mathscinet-getitem?mr=2356821.

18. Lecué G. Simultaneous adaptation to the margin and to complexity in classification // Annals of Statistics. 2007. Vol. 35. Pp. 1698–1721. URL: http://www.ams.org/mathscinet-getitem?mr=2351102.

19. Rigollet P., Tsybakov A. Sparse estimation by exponential weighting // Statistical Science. 2012. Vol. 27. Pp. 558–575. URL: http://arxiv.org/pdf/1108.5116.pdf.

20. Нестеров Ю. Е. Введение в выпуклую оптимизацию. М.: Изд. МЦНМО, 2010.

21. Leung G., Barron A. Information theory and mixing least-squares regressions // IEEE Transactions on Information Theory. 2006. Vol. 52. Pp. 3396–3410.

22. Dalalyan A., Salmon J. Sharp oracle inequalities for aggregation of affine estimators // Annals of Statistics. 2012. Vol. 40. Pp. 2327–2355. URL: http://arxiv.org/abs/1104.3969v2.

23. Chernousova E., Golubev Yu., Krymova E. Ordered smoothers with exponential weighting // Electronic Journal of Statistics. 2013. Vol. 7. Pp. 2395–2419. URL: http://dx.doi.org/10.1214/13-EJS849.

24. Alquier P., Lounici K. PAC-bayesian bounds for sparse regression estimation with exponential weights // Electronic Journal of Statistics. 2011. Vol. 5. Pp. 127–145. URL: http://arxiv.org/pdf/1009.2707.pdf.

25. Arias-Castro E., Lounici K. Variable selection with exponential weights and l_0-penalization // arXiv preprint. 2012. URL: http://arxiv.org/abs/1208.2635v2.

26. Golubev Yu., Ostrovski D. Concentration inequalities for the exponential weighting method // Mathematical Methods of Statistics. 2014. Vol. 23, no. 1. Pp. 20–37. URL: http://dx.doi.org/10.3103/S1066530714010025.

27. Голубев Г. К. Экспоненциальное взвешивание и оракульные неравенства для проекционных оценок // Проблемы передачи информации. 2012. Т. 48. С. 269–280. URL: http://arxiv.org/pdf/1206.4285.pdf.