Неравенства концентрации для выборок без возвращений *

Толстихин Илья Олегович
Max Planck Institute for Intelligent Systems
ilya@tuebingen.mpg.de

Аннотация

В работе рассматривается явление концентрации значений функций случайных величин, выбранных без возвращений из фиксированного конечного множества, вблизи их математических ожиданий — задача, актуальная в ряде приложений, включая трансдуктивную постановку теории статистического обучения. Помимо обзора известных результатов, активно применяющихся в литературе, в работе изучается два общих подхода, ведущих во многих случаях к достаточно точным неравенствам концентрации. Первый основан на субгауссовском неравенстве С. Г. Бобкова [8] для функций, определенных на срезе Булева куба. Второй подход, предложенный в известной работе В. Хефдинга [17], сводит задачу к рассмотрению выборки независимых случайных величин. На их основе получен ряд неравенств концентрации, включая два новых неравенства для супремумов эмпирических процессов для выборок без возвращений.

1 Введение

В работе рассматривается вопрос концентрации значений функций $f(Z_1, \dots, Z_m)$ вблизи математических ожиданий $\mathsf{E}[f(Z_1, \dots, Z_m)]$, где $\{Z_1, \dots, Z_m\}$ — случайные величины, выбранные равномерно *без возвращений* из некоторого конечного множества действительных чисел мощностью N. В том случае, когда случайные величины независимы (и не обязательно одинаково распределены), поставленный вопрос хорошо изучен [10] и существует множество известных результатов: классические неравенства Хефдинга, Бернштейна и Беннета для сумм случайных величин [17], неравенство МакДиармида и метод мартингалов для более широкого класса функций с *ограниченными разностями* [22], индуктивный метод Талаграна и его неравенство для супремумов эмпирических процессов

*Работа была выполнена, когда автор являлся сотрудником ВЦ РАН. Работа выполнена при поддержке Российского фонда фундаментальных исследований (проект № 14-07-00847) и программы ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики и информационные системы нового поколения».

[24, 11, 10], *энтропийный* метод Леду [10] и другие. Однако, при выборке случайных величин без возвращений они перестают быть независимыми, вследствие чего многие из классических подходов к получению результатов концентрации перестают быть применимыми.

Схема выборки без возвращений применяется для получения оценок надёжности восстановления зависимостей по конечным выборкам данных. Первые оценки равномерного уклонения частот ошибок в двух подвыборках были получены в [1, 2]. Позже эти идеи получили развитие в трансдуктивном подходе теории статистического обучения [25, 12] и в комбинаторной теории переобучения [3, 26, 27, 28].

Трансдуктивный подход предполагает, что *обучающая выборка* X^m — множество m наблюдаемых пар «объект-ответ» из декартового произведения $\mathcal{X} \times \mathcal{Y}$ — получена в результате случайного разбиения конечной и фиксированной *генеральной совокупности* $X^N = \{(x_i, y_i)\}_{i=1}^N$ на два непересекающихся подмножества мощностей m и $N-m$ соответственно, что эквивалентно выбору m пар «объект-ответ» без возвращений из генеральной совокупности. Задача заключается в поиске на основе обучающей выборки отображения (предиктора) $h^*\colon \mathcal{X} \to \mathcal{Y}$ из заранее фиксированного класса отображений $\mathcal{H}$, имеющего минимальное значение *функционала риска*, определенного на второй (скрытой) части генеральной совокупности мощности $N-m$. Поскольку риск предиктора на случайной подвыборке — случайная величина, то большинство утверждений, используемых в анализе подобных задач, несут вероятностный характер.

В анализе классической *индуктивной* постановки теории статистического обучения, где объекты обучающей выборки вместе с ответами на них выбираются независимо из неизвестного распределения $\mathbb{P}$ на $\mathcal{X} \times \mathcal{Y}$, неравенства концентрации [10] играют ключевую роль [19, 9]. Особого внимания заслуживает неравенство Талаграна для супремумов эмпирических процессов [24] (позже усиленное в [11]), на котором основаны наиболее важные результаты серии работ [20, 6, 18], использующие *локальные рademахеровские сложности*. Вопросу же применения неравенств концентрации в трансдуктивном подходе посвящено гораздо меньшее количество работ, среди которых можно отметить [25, 7, 14, 15, 13]. Более того, в литературе до сих пор не изучалась возможность применения локального рademахеровского анализа в рамках трансдуктивного подхода. Подобное направление исследований потребовало бы аналога неравенства Талаграна для выборок без возвращений, так же не рассматривавшегося до сих пор в литературе.

Работа состоит из трех частей. В Разделе 2 будут рассмотрены известные результаты для сумм случайных величин $f(Z_1, \ldots, Z_m) = \sum_{i=1}^m Z_i$, выбранных без возвращений. На этом примере будет наглядно продемонстрировано, что выборки без возвращений ведут к более сильной концентрации по сравнению с выборками независимых случайных величин. В Разделе 3 будет приведено два результата, справедливых для более широкого класса функций f, определенных на случайных разбиениях конечного множества действительных чисел. Первый, полученный Р. Эль-Янивом, Д. Печиони и другими в [15, 13], является аналогом неравенства МакДиармида и усиливает его при росте размера выборки m к N. Второй получен С. Г. Бобковым в [8] и устанавливает субгауссовское поведение таких функций, позволяя, в отличие от неравенства МакДиармида, учитывать дисперсии случайных величин. Наконец, в Разделе 4 на основе описанных подходов будет получено два новых неравенства концентрации для супремумов эмпирических процессов для выборок без возвращений, одно из которых является непосредственным обобщением неравенства Талаграна.

2 Суммы случайных величин

Сумма случайных величин является классическим объектом изучения теории вероятностей и математической статистики. В том случае, когда случайные величины независимы и ограничены, неравенства Хефдинга, Бернштейна и Беннета [17] дают оптимальные неасимптотические верхние оценки отклонения сумм от их математических ожиданий. В этом разделе будут приведены известные результаты для случая, когда слагаемые выбраны без возвращений из конечного множества действительных чисел.

Стандартным методом получения неравенств концентрации для произвольных случайных величин ξ является так называемый *метод Чернова*, заключающийся в применении неравенства Маркова:

$$\mathbb{P}\{\xi - \mathsf{E}[\xi] \geqslant \varepsilon\} = \mathbb{P}\left\{e^{\lambda(\xi-\mathsf{E}[\xi])} \geqslant e^{\lambda\varepsilon}\right\} \leqslant \frac{\mathsf{E}\left[e^{\lambda(\xi-\mathsf{E}[\xi])}\right]}{e^{\lambda\varepsilon}} \tag{1}$$

для неотрицательного $\lambda \geqslant 0$, получении верхней оценки $F(\lambda)$ на производящую функцию моментов случайной величины $\xi - \mathsf{E}[\xi]$:

$$\mathsf{E}\left[e^{\lambda(\xi-\mathsf{E}[\xi])}\right] \leqslant F(\lambda) \tag{2}$$

и последующей минимизации полученной оценки по $\lambda \geqslant 0$:

$$\mathbb{P}\{\xi - \mathsf{E}[\xi] \geqslant \varepsilon\} \leqslant \min_{\lambda\geqslant 0} \frac{F(\lambda)}{e^{\lambda\varepsilon}}.$$

Для суммы $\xi = \sum_{i=1}^m X_i$ независимых случайных величин $\{X_1, \ldots, X_m\}$ описанный метод Чернова особенно удобен в применении, поскольку справедливо следующее:

$$\mathsf{E}\left[e^{\lambda(\xi-\mathsf{E}[\xi])}\right] = \prod_{i=1}^m \mathsf{E}\left[e^{\lambda(X_i-\mathsf{E}[X_i])}\right], \tag{3}$$

благодаря чему задача сводится к построению верхних оценок для отдельных множителей, входящих в правую часть последнего тождества. В зависимости от способа ограничения производящих функций моментов случайных величин $X_i - \mathsf{E}[X_i]$ (см Раздел A Приложения), метод Чернова ведет к неравенствам Хефдинга, Беннетта или Бернштейна. Однако, описанный способ неприменим для суммы *зависимых* случайных величин, поскольку в этом случае тождество (3) перестает выполняться. В частности, это происходит для суммы случайных величин, выбранных без возвращений из конечного множества действительных чисел. В этом случае необходим другой способ построения верхней оценки $F(\lambda)$ на производящую функцию моментов (2).

Рассмотрим конечное множество $\mathcal{C} = \{c_1, \ldots, c_N\}$ ограниченных действительных чисел $c_i \in [0, 1]$, $i = 1 \ldots, N$, с возможными повторениями. Для произвольного натурального числа $m \leqslant N$ всюду далее с помощью $\{Z_1, \ldots, Z_m\}$ и $\{X_1, \ldots, X_m\}$ будем обозначать случайные величины, выбранные равномерно из $\mathcal{C}$ без возвращений и с возвращениями соответственно. Еще раз отметим, что случайные величины $\{X_1, \ldots, X_m\}$ независимы, в то время как $\{Z_1, \ldots, Z_m\}$ таковыми не являются. Введем обозначения $S_m = \frac{1}{m}\sum_{i=1}^m X_i$ и $S'_m = \frac{1}{m}\sum_{i=1}^m Z_i$.

Для получения неравенств концентрации для S'_m в литературе принято использовать два подхода, описанных далее. Первый основан на классическом результате В. Хефдинга, который позволяет свести анализ схемы выборки без возвращений к выборке с возвращениями. Второй, впервые примененный в такой постановке Серфлингом в [23], использует метод мартингалов для непосредственной оценки производящей функции моментов в правой части неравенства (1).

2.1 Метод Хефдинга

Первый подход основан на слеующем результате В. Хефдинга:

Теорема 1 ([17]). [1] *Пусть* $\{U_1, \dots, U_m\}$ *и* $\{W_1, \dots, W_m\}$ *выбраны равномерно из конечного множества* d*-мерных векторов* $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_N\} \subset \mathbb{R}^d$ *с и без возвращений соответственно. Тогда для любой непрерывной и выпуклой функции* $f \colon \mathbb{R}^d \to \mathbb{R}$ *справедливо:*

$$\mathsf{E}\left[f\left(\sum_{i=1}^{m} W_i\right)\right] \leqslant \mathsf{E}\left[f\left(\sum_{i=1}^{m} U_i\right)\right].$$

Как будет показано далее в работе, Теорема 1 является чрезвычайно удобным средством переноса результатов концентрации, справедливых для функций независимых случайных величин, на выборки без возвращений. Положив $f(x) = e^{\lambda x}$, мы немедленно получаем следующее неравенство:

Следствие 1. *Для любого* $\lambda \geqslant 0$ *справедливо:*

$$\mathsf{E}\Big[\exp\big(\lambda\left(S'_m - \mathsf{E}[S'_m]\right)\big)\Big] \leqslant \mathsf{E}\Big[\exp\big(\lambda\left(S_m - \mathsf{E}[S_m]\right)\big)\Big].$$

Доказательство. Положив в Теореме 1 в качестве $f(x) = \exp\left(\frac{\lambda}{m}x\right)$ для произвольной $\lambda > 0$, мы получим:

$$\begin{aligned}
\mathsf{E}\left[\exp\left(\frac{\lambda}{m}\sum_{i=1}^{m}\left(Z_i - \mathsf{E}[Z_i]\right)\right)\right] &= \mathsf{E}\left[\exp\left(\frac{\lambda}{m}\sum_{i=1}^{m} Z_i\right)\right]\exp\left(-\frac{\lambda}{m}\sum_{i=1}^{m}\mathsf{E}[Z_i]\right) \\
&\leqslant \mathsf{E}\left[\exp\left(\frac{\lambda}{m}\sum_{i=1}^{m} X_i\right)\right]\exp\left(-\frac{\lambda}{m}\sum_{i=1}^{m}\mathsf{E}[Z_i]\right) \\
&= \mathsf{E}\left[\exp\left(\frac{\lambda}{m}\sum_{i=1}^{m}\left(X_i - \mathsf{E}[X_i]\right)\right)\right].
\end{aligned}$$

□

Замечание 1. *В последнем доказательстве мы также использовали следующий очевидный факт, который мы приводим без доказательства:*

$$\mathsf{E}[S'_m] = \mathsf{E}[S_m] = \frac{1}{N}\sum_{i=1}^{N} c_i.$$

[1] Хотя в статье В. Хефдинга в явном виде не указано, что результат справедлив для случайных векторов, все доказательства остаются справедливы и для этого случая. См., например, [16] Раздел D.

Применив неравенство (1) для $\xi = S'_m$ и ограничив правую часть с помощью Следствия 1, мы приходим к выводу, что все неравенства концентрации для S_m, полученные с помощью метода Чернова, также справедливы для S'_m. В частности, справедливы следующие неравенства Хефдинга и Бернштейна:

Теорема 2 (Неравенство Хефдинга [17]). *Пусть* $\bar{c} = \frac{1}{N}\sum_{i=1}^N c_i$. *Тогда для любого* $\varepsilon \geqslant 0$ *справедливо:*

$$\begin{aligned}\mathbb{P}\left\{S'_m - \mathsf{E}[S'_m] \geqslant \varepsilon\right\} &\leqslant \exp\bigl(-m\,\mathrm{kl}\,(\bar{c}+\varepsilon\|\bar{c})\bigr)\\ &\leqslant e^{-2m\varepsilon^2},\end{aligned}$$

где мы обозначили дивергенцию Кульбака-Лейблера между двумя распределениями Бернулли с параметрами $0 \leqslant p \leqslant 1$ *и* $0 \leqslant q \leqslant 1$ *с помощью* $\mathrm{kl}\,(p\|q) = p\ln\frac{p}{q} + (1-p)\ln\frac{1-p}{1-q}$. *Аналогичные неравенства справедливы для* $\mathbb{P}\left\{\mathsf{E}[S'_m] - S'_m \geqslant \varepsilon\right\}$.

Теорема 3 (Неравенство Бернштейна [17]). *Пусть* $\bar{c} = \frac{1}{N}\sum_{i=1}^N c_i$ *и*

$$\sigma_N^2 = \frac{1}{N}\sum_{i=1}^N (c_i - \bar{c})^2.$$

Обозначим $h(u) = (1+u)\log(1+u) - u$ *для* $u \geqslant 0$. *Тогда для любого* $\varepsilon \geqslant 0$ *справедливо:*

$$\begin{aligned}\mathbb{P}\left\{S'_m - \mathsf{E}[S'_m] \geqslant \varepsilon\right\} &\leqslant \exp\left(-m\,\sigma_N^2 h\left(\frac{\varepsilon}{\sigma^2}\right)\right)\\ &\leqslant \exp\left(-\frac{m\varepsilon^2}{2(\sigma_N^2 + \varepsilon/3)}\right).\end{aligned} \tag{4}$$

Аналогичные неравенства справедливы для $\mathbb{P}\left\{\mathsf{E}[S'_m] - S'_m \geqslant \varepsilon\right\}$. *Кроме того, для любого* $t > 0$ *с вероятностью не меньше* $1 - e^{-t}$ *выполнено:*

$$\mathsf{E}[S'_m] \leqslant S'_m + \sqrt{\frac{2\sigma_N^2 t}{m}} + \frac{2t}{3m}.$$

Замечание 2. *Неравенство* (4) *следует из первой оценки Теоремы 3 и элементарного неравенства* $h(u) \geqslant \frac{u^2}{2(1+u/3)}$, *справедливого для* $u > 0$ *(см. Упражнение 2.8 [10]). Последнее неравенство теоремы*[2] *следует из* (4) *и неравенства* $\sqrt{a+b} \leqslant \sqrt{a} + \sqrt{b}$.

Отметим важную особенность неравенства Бернштейна (4): при $\varepsilon \ll \sigma_N^2$ оценка ведет себя как $e^{-m\varepsilon^2}$, в то время как для $\varepsilon \gg \sigma_N^2$ — как $e^{-m\varepsilon}$. Мы наблюдаем два «режима» оценки неравенства Бернштейна: малые отклонения S'_m от $\mathsf{E}[S'_m]$ описываются субгауссовским поведением, отвечающим члену порядка $\sqrt{1/m}$ в последнем неравенстве, а большие — более тяжелыми лапласовскими хвостами, которым отвечает член порядка $1/m$.

Обратим внимание, что дисперсии сумм S'_m и S_m, в отличие от их математических ожиданий, не совпадают. Справедливо следующее [17]:

$$\mathsf{D}[S'_m] = \frac{N-m}{N-1}\frac{\sigma_N^2}{m} = \frac{N-m}{N-1}\,\mathsf{D}[S_m],$$

[2] Аккуратный анализ позволит избавиться от множителя 2 перед последним слагаемым [11][Теорема 2.1]

где σ_N^2, определенная в Теореме 3, — дисперсия равномерно распределенной на $\mathcal{C}$ случайной величины. Таким образом, дисперсия S'_m убывает по сравнению с дисперсией S_m по мере роста m, пока не будут выбраны все $m = N$ элементов множества $\mathcal{C}$: в этом случае S'_m вырождается в константу и $\mathsf{D}[S'_m] = 0$. Этот факт (наряду с Теоремой 1) демонстрирует, что S'_m концентрируется сильнее S_m. Авторы [14] дают следующее интуитивное объяснение этого эффекта: последовательное уменьшение размера множества, из которого мы выбираем очередную случайную величину Z_i без возвращения, ведет к уменьшению ее разброса по сравнению со случаем выборки с возвращениями.

2.2 Неравенство Серфлинга

Второй подход основан на непосредственной оценке производящей функции моментов $\mathsf{E}\left[e^{\lambda(\xi-\mathsf{E}[\xi])}\right]$ в правой части неравенста (1) для случайной величины $\xi = S'_m$ без использования Теоремы 1. В работе [23] с помощью метода мартингалов получен следующий результат, который всегда точнее второй верхней оценки Теоремы 2:

Теорема 4 (Серфлинг [23]). *Обозначим $\Delta(m) = \frac{m-1}{N}$. Тогда для любого $\varepsilon \geqslant 0$ справедливо:*

$$\mathbb{P}\left\{S'_m - \mathsf{E}[S'_m] \geqslant \varepsilon\right\} \leqslant \exp\left(-\frac{2m\varepsilon^2}{1-\Delta(m)}\right).$$

Аналогичные неравенства справедливы для $\mathbb{P}\left\{\mathsf{E}[S'_m] - S'_m \geqslant \varepsilon\right\}$.

По мере роста $m \to N$ знаменатель показателя экспоненты убывает к $1/N$, и неравенство существенно уточняет Теорему 2. С другой стороны, при $m = o(N)$ — случай, когда схема выборки без возвращений приближается к схеме с возвращениями, — знаменатель стремится к 1 и неравенство совпадает с Теоремой 2.

Замечание 3. *Как мы отмечали ранее, при $m = N$ случайная величина S'_m вырождается в константу, что влечет за собой тождество $\mathbb{P}\left\{S'_m - \mathsf{E}[S'_m] \geqslant \varepsilon\right\} = 0$ для произвольного $\varepsilon > 0$. Однако, при $m = N$ в правой части Теоремы 4 мы получаем $e^{-2mN\varepsilon^2} > 0$. В [5][Утверждение 4] предложен результат, совпадающий с неравенством Теоремы 4 с точностью до замены $1-(m-1)/N$ в знаменателе показателя экспоненты на $1-m/N$. Кроме того, авторы [5] развивают подход, предложенный Серфлингом, и предлагают неравенства «бернштейновского» типа для выборок без возвращений, которые, подобно тому как Теорема 4 улучшает Теорему 2, уточняют неравенство Теоремы 3 при $m \to N$.*

На примере сумм независимых случайных величин мы показали, что, хотя применение Теоремы 1 является удобным способом получения неравенств концентрации для выборок без возвращений, непосредственная оценка производящей функции моментов $\mathsf{E}\left[e^{\lambda(\xi-\mathsf{E}[\xi])}\right]$ рассматриваемой случайной величины может вести к существенно лучшим результатам. Кроме того было показано, что суммы для выборок без возвращений концентрируются сильнее сумм независимых случайных величин. Результаты следующего раздела показывают, что подобный эффект наблюдается и для более общих функций случайных величин.

3 Функции, определенные на разбиениях

Заметим, что случайную величину S'_m из прошлого раздела можно эквивалентно определить с помощью случайных перестановок. Рассмотрим случайную перестановку, выбранную равномерно из симметрической группы перестановок множества $\{1,\dots,N\}$. Такая перестановка может быть выражена N-мерным вектором $\boldsymbol{\pi}$ с натуральными координатами, полученными перестановкой множества $\{1,\dots,N\}$. Тогда S'_m можно определить как

$$S'_m = S'_m(\boldsymbol{\pi}) = \frac{1}{m}\sum_{i=1}^{m} c_{\pi_i},$$

где π_i — i-ая координата $\boldsymbol{\pi}$, а с помощью $\mathcal{C} = \{c_1,\dots,c_N\}$ мы по-прежнему обозначаем конечное множество ограниченных действительных чисел $c_i \in [0,1]$.

Существует также третий способ определения функции S'_m, основанный на разбиениях. Пусть $(\mathcal{U}^m,\mathcal{U}^u)$ — разбиение множества $\{1,\dots,N\} = \mathcal{U}^m \cup \mathcal{U}^u$ на два непересекающихся подмножества мощностей m и $u = N - m$ соответственно. Мы будем выбирать $(\mathcal{U}^m,\mathcal{U}^u)$ равномерно из множества всех таких разбиений, которых всего $C_N^m = \frac{N!}{m!(N-m)!}$. Тогда

$$S'_m = S'_m(\mathcal{U}^m,\mathcal{U}^u) = \frac{1}{m}\sum_{i\in\mathcal{U}^m} c_i.$$

Очевидно, не все функции $f(\boldsymbol{\pi})$ случайных перестановок $\boldsymbol{\pi}$ могут быть эквивалентно представлены с помощью разбиений: например, подобное представление невозможно для функции $f(\boldsymbol{\pi}) = c_{\pi_1} + c^2_{\pi_2} + c^3_{\pi_3}$. Для суммы случайных величин S'_m это возможно благодаря ее симметричности относительно перестановок слагаемых. Оказывается, для функций, допускающих представление с помощью разбиений, справедлив ряд нетривиальных неравенств концентрации, обзор которых будут приведен в данном разделе. Часть из них первоначально формулировалась в терминах случайных перестановок $\boldsymbol{\pi}$, а часть — в терминах случайных разбиений $\mathcal{U}^m \cup \mathcal{U}^u$. Для удобства и однообразия мы будем формулировать все результаты в терминах перестановок.

Сначала будет рассмотрен аналог классического неравенства МакДиармида (также известного как *неравенство ограниченных разностей*) для выборок без возвращений. Данный результат, подобно неравенству Серфлинга, получен авторами [15] на основе непосредственной работы с производящей функцией моментов с помощью метода мартингалов. Затем будет приведено субгауссовское неравенство С. Г. Бобкова [8], непосредственно связанное с энтропийным подходом М. Леду [10] — мощного и в то же время достаточно простого в использовании подхода к получению неравенств концентрации для выборок независимых случайных величин.

3.1 Неравенство МакДиармида

Одним из наиболее простых и в то же время удобных в применении неравенств концентрации для функций $f(X_1,\dots,X_n)$ независимых случайных величин $\{X_1,\dots,X_n\}$ является неравенство МакДиармида. Для формулировки неравенства МакДиармида нам понадобится понятие *функций с ограниченными разностями*:

Определение 1. *Рассмотрим произвольное множество $\mathcal{Z}$. Функция $f\colon \mathcal{Z}^n \to \mathbb{R}$* имеет ограниченные разности, *если существуют неотрицательные константы $b_1, \dots, b_n \in \mathbb{R}$, такие что*

$$\sup_{\substack{x_1,\dots,x_n,\\ x_i' \in \mathcal{Z}^n}} |f(x_1,\dots,x_n) - f(x_1,\dots,x_{i-1},x_i',x_{i+1},\dots,x_n)| \leqslant b_i, \quad i = 1,\dots,n. \tag{5}$$

Справедлив следующий результат:

Теорема 5 (Неравенство МакДиармида [10])**.** *Пусть функция $f\colon \mathcal{Z}^n \to \mathbb{R}$ имеет ограниченные разности с константами $b_1, \dots, b_n$. Рассмотрим выборку независимых случайных величин $\{X_1, \dots, X_N\}$ из множества $\mathcal{Z}$. Обозначим $\xi = f(X_1, \dots, X_n)$. Тогда для любого $\varepsilon \geqslant 0$ справедливо:*

$$\mathbb{P}\{\xi - \mathsf{E}[\xi] \geqslant \varepsilon\} \leqslant \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n b_i^2}\right).$$

Аналогичное неравенство справедливо для $\mathbb{P}\{\mathsf{E}[\xi] - \xi \geqslant \varepsilon\}$.

Этот результат утверждает, что если функция $f(x_1, \dots, x_n)$ не зависит слишком сильно ни от одного из своих аргументов, то $f(X_1, \dots, X_n)$ концентрируется вокруг математического ожидания $\mathsf{E}[f(X_1, \dots, X_n)]$.

Несложно проверить, что функция $f(X_1, \dots, X_n) = \frac{1}{n}\sum_{i=1}^n X_i$ удовлетворяет условию ограниченных разностей с константами $b_i = \frac{1}{n}$, если $\{X_1, \dots, X_n\}$ — выборка независимых случайных величин, таких что $X_i \in [0, 1]$, $i = 1, \dots, n$. В этом случае неравенство МакДиармида в точности воспроизводит неравенство Хефдинга Теоремы 2. Таким образом, неравенство МакДиармида можно считать обобщением неравенства Хефдинга на более общий (чем сумма) класс функций независимых случайных величин.

Для обобщения неравенства МакДиармида на выборки без возвращений в работе [15] вводится понятие (m, u)-*симметричных относительно перестановок* функций:

Определение 2. *Функция $f\colon \boldsymbol{\pi} \to \mathbb{R}$, заданная на симметрической группе множества $\{1, \dots, N\}$, называется* (m, u)-симметричной относительно перестановок, *если она не меняет своего значения при замене порядка первых m и/или последних $u = N - m$ координат $\boldsymbol{\pi}$. Для краткости такие функции мы будем называть просто* (m, u)-симметричными.

Очевидно, любая (m, u)-симметричная функция $f(\boldsymbol{\pi})$ может быть определена в терминах случайных разбиений. И, наоборот, любая функция, определенная на множестве разбиений $\mathcal{U}^m \cup \mathcal{U}^u$, допускает представление с помощью (m, u)-симметричной функции. Все результаты настоящего раздела будут формулироваться в терминах (m, u)-симметричных функций.

Авторы [15] приводят следующий результат, как и неравенство Серфлинга Теоремы 4 основанный на методе мартингалов:

Теорема 6 (Эль-Янив, Печиони [15])**.** *Пусть $\boldsymbol{\pi}$ — вектор случайной перестановки, выбранной равномерно из симметрической группы перестановок множества $\{1, \dots, N\}$. Пусть $f(\boldsymbol{\pi})$ — (m, u)-симметричная функция, для которой существует константа $\beta > 0$, такая что $|f(\boldsymbol{\pi}) - f(\boldsymbol{\pi}^{i,j})| \leqslant \beta$ для всех $\boldsymbol{\pi}$, $i \in \{1, \dots, m\}$ и $j \in \{m+1, \dots, N\}$, где*

перестановка $\boldsymbol{\pi}^{i,j}$ получена из $\boldsymbol{\pi}$ транспозицией ее i-й и j-й координат. Тогда для любого $\varepsilon \geqslant 0$:

$$\mathbb{P}\left\{f(\boldsymbol{\pi}) - \mathsf{E}[f(\boldsymbol{\pi})] \geqslant \varepsilon\right\} \leqslant \exp\left\{-\frac{2\varepsilon^2}{m\beta^2}\left(\frac{N-1/2}{N-m}\right)\left(1-\frac{1}{2\max(m,N-m)}\right)\right\}.$$

Аналогичное неравенство справедливо и для $\mathbb{P}\left\{\mathsf{E}[f(\boldsymbol{\pi})] - f(\boldsymbol{\pi}) \geqslant \varepsilon\right\}$, поскольку предположения Теоремы инвариантны относительно замены знака функции f.

Теорема 6 является аналогом неравенства МакДиармида для выборок без возвращений, а ее предположения непосредственно связаны с условием ограниченных разностей (5). Грубое сравнение двух неравенств показывает, что они совпадают с точностью до отсутствия выражения $\left(\frac{N-1/2}{N-m}\right)\left(1-\frac{1}{2\max(m,N-m)}\right)$ в показателе экспоненты неравенства МакДиармида. Пренебрегая вторым множителем, который близок к 1 для больших N, можно заключить, что при $m \to N$ оценка Теоремы 6 точнее неравенства МакДиармида.

Заметив, что сумма S'_m из прошлого раздела удовлетворяет условию последней Теоремы с $\beta = \frac{1}{m}$, мы немедленно получаем следствие:

Следствие 2. *Для любого $\varepsilon \geqslant 0$ справедливо:*

$$\mathbb{P}\left\{S'_m - \mathsf{E}[S'_m] \geqslant \varepsilon\right\} \leqslant \exp\left\{-2m\varepsilon^2\left(\frac{N-1/2}{N-m}\right)\left(1-\frac{1}{2\max(m,N-m)}\right)\right\}.$$

Аналогичное неравенство справедливо для $\mathbb{P}\left\{\mathsf{E}[S'_m] - S'_m \geqslant \varepsilon\right\}$.

При больших N последнее неравенство имеет тот же порядок, что неравенство Серфлинга Теоремы 4.

3.2 Неравенство Бобкова

Следующий результат основан на *энтропийном подходе* [10], активно развивавшемся последних 15 лет. Нам понадобится ряд определений.

Рассмотрим случайное разбиение $(\mathcal{U}^m, \mathcal{U}^u)$ множества $\{1,\dots,N\} = \mathcal{U}^m \cup \mathcal{U}^u$ на два непересекающихся подмножества мощностей m и $u = N-m$ соответственно, равномерно распределенное на множестве из всех $\frac{N!}{m!(N-m)!}$ таких разбиений. Далее разбиение будет удобно представлять вектором перестановки $\boldsymbol{\pi}$, подразумевая, что перестановка задает разбиение на подмножества $\{\pi_i\}_{i\in I} \cup \{\pi_j\}_{j\in J}$, где $I = \{1,\dots,m\}$ и $J = \{m+1,\dots,N\}$. *Соседними* естественно считать разбиения, которые *могут быть заданы* перестановками $\boldsymbol{\pi}_1$ и $\boldsymbol{\pi}_2$, отличающимися ровно на одну транспозицию: $\boldsymbol{\pi}_1 = \boldsymbol{\pi}_2^{i,j}$ для некоторых $i \in I$ и $j \in J$ (обратим внимание на то, что каждое разбиение может быть задано $m(N-m)$ разными перестановками). Каждое разбиение, таким образом, имеет ровно $m(N-m)$ соседних разбиений.

Как отмечалось ранее, любая (m,u)-симметричная функция f фактически задается на множестве разбиений. Дискретный градиент $\nabla f(\boldsymbol{\pi})$ такой функции является вещественным вектором размерностью $m(N-m)$ и квадрат его длины выражается следующим образом:

$$V^f(\boldsymbol{\pi}) = |\nabla f(\boldsymbol{\pi})|^2 = \sum_{i\in I}\sum_{j\in J}\left(f(\boldsymbol{\pi}) - f(\boldsymbol{\pi}^{i,j})\right)^2.$$

Следующий результат получен в [8][Теорема 2.1]:

Теорема 7 (С. Г. Бобков, [8]). *Пусть $\boldsymbol{\pi}$ — вектор случайной перестановки, выбранной равномерно из симметрической группы перестановок множества $\{1,\dots,N\}$. Пусть $f(\boldsymbol{\pi})$ — (m,u)-симметричная функция и $\Sigma^2 \geqslant 0$ — действительное число, такое что $V^f(\boldsymbol{\pi}) \leqslant \Sigma^2$ для всех $\boldsymbol{\pi}$. Тогда для любого $\varepsilon \geqslant 0$ справедливо:*

$$\mathbb{P}\left\{f(\boldsymbol{\pi}) - \mathsf{E}[f(\boldsymbol{\pi})] \geqslant \varepsilon\right\} \leqslant \exp\left\{-\frac{(N+2)\varepsilon^2}{4\Sigma^2}\right\}.$$

Аналогичное неравенство справедливо для $\mathbb{P}\left\{\mathsf{E}[f(\boldsymbol{\pi})] - f(\boldsymbol{\pi}) \geqslant \varepsilon\right\}$, поскольку предположения Теоремы инвариантны относительно замены знака функции f.

Замечание 4. *Приведенная здесь формулировка теоремы отличается от первоначальной версии С. Г. Бобкова, формулировавшейся в терминах случайных разбиений $(\mathcal{U}^m,\mathcal{U}^u)$. Она является следствием того, что для (m,u)-симметричных функций f случайные величины $f(\mathcal{U}^m,\mathcal{U}^u)$ и $f(\boldsymbol{\pi})$, как несложно убедиться, одинаково распределены.*

Предположим, что (m,u)-симметричная относительно перестановок функция $f(\boldsymbol{\pi})$ удовлетворяет предположениям Теоремы 6, то есть существует значение β, такое что $|f(\boldsymbol{\pi}) - f(\boldsymbol{\pi}^{i,j})| \leqslant \beta$ для всех $\boldsymbol{\pi}$, $i \in I$ и $j \in J$. Отсюда очевидным образом следует, что она также удовлетворяет предположениям Теоремы 7 с параметром $\sigma^2 = m(N-m)\beta^2$, поскольку:

$$V^f(\boldsymbol{\pi}) = \sum_{i\in I}\sum_{j\in J}\big(f(\boldsymbol{\pi}) - f(\boldsymbol{\pi}^{i,j})\big)^2 \leqslant \sum_{i\in I}\sum_{j\in J}\beta^2 = m(N-m)\beta^2.$$

В этом случае Теорема 7 дает более слабую версию Теоремы 6: показатель экспоненты уменьшается в 8 раз. Тем не менее, предположения Теоремы 7 представляются менее строгими, и известно множество приложений (см. замечания к Теореме 6.7 [10]), для которых Теорема 6 дает лишь тривиальные неравенства, в то время как Теорема 7 позволяет получить достаточно сильные результаты. Один из таких примеров будет приведен в Разделе 4 настоящей работы.

Попробуем применить Теорему 7 для получения неравенств концентрации для (m,u)-симметричной функции $S'_m = \frac{1}{m}\sum_{i=1}^m c_{\pi_i}$. Для этого нам понадобится верхняя оценка на $V^f(\boldsymbol{\pi})$ для $f(\boldsymbol{\pi}) = S'_m(\boldsymbol{\pi})$. Заметим, что для любых $\boldsymbol{\pi}$, $i \in I$ и $j \in J$ справедливо:

$$S'_m(\boldsymbol{\pi}) - S'_m(\boldsymbol{\pi}^{i,j}) = \frac{1}{m}(c_{\pi_i} - c_{\pi_j}),$$

где, как и ранее, перестановка $\boldsymbol{\pi}^{i,j}$ получена из $\boldsymbol{\pi}$ транспозицией ее i-й и j-й координат. Мы получаем

$$V^f(\boldsymbol{\pi}) = \frac{1}{m^2}\sum_{i\in I}\sum_{j\in J}(c_i - c_j)^2.$$

Очевидная оценка $(c_i - c_j)^2 \leqslant 1$ дает нам $V^f(\boldsymbol{\pi}) \leqslant \frac{(N-m)}{m}$, что с учетом Теоремы 7 дает более слабую версию Следствия 2. Мы также можем воспольроваться более точной оценкой, для которой нам понадобится следующий технический результат:

Лемма 8. *Для любой последовательности действительных чисел* $\{x_1, \dots, x_n\}$ *справедливо:*

$$\frac{1}{n(n-1)} \sum_{1 \leqslant i < j \leqslant n} (x_i - x_j)^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \frac{1}{n} \sum_{j=1}^{n} x_j \right)^2.$$

Доказательство.

$$\begin{aligned}
\frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \frac{1}{n} \sum_{j=1}^{n} x_j \right)^2 &= \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i^2 - \frac{2}{n} x_i \sum_{j=1}^{n} x_j + \frac{1}{n^2} \left(\sum_{j=1}^{n} x_j \right)^2 \right) \\
&= \frac{1}{n-1} \left(\sum_{i=1}^{n} x_i^2 - \frac{2}{n} \sum_{i=1}^{n} x_i \sum_{j=1}^{n} x_j + \frac{1}{n^2} \sum_{i=1}^{n} \left(\sum_{j=1}^{n} x_j \right)^2 \right) \\
&= \frac{1}{n-1} \left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left(\sum_{j=1}^{n} x_j \right)^2 \right) \\
&= \frac{1}{n(n-1)} \left((n-1) \sum_{i=1}^{n} x_i^2 - 2 \sum_{1 \leqslant i < j \leqslant n} x_i x_j \right) \\
&= \frac{1}{n(n-1)} \sum_{1 \leqslant i < j \leqslant n} (x_i - x_j)^2.
\end{aligned}$$

□

Лемма 9. *Для* $f(\boldsymbol{\pi}) = S'_m = \frac{1}{m} \sum_{i=1}^{m} c_{\pi_i}$ *справедливо:*

$$\sup_{\boldsymbol{\pi}} V^f(\boldsymbol{\pi}) \leqslant \left(\frac{N}{m} \right)^2 \sigma_N^2, \tag{6}$$

где с помощью σ_N^2 *обозначена дисперсия случайной величины, равномерно распределенной на множестве* $\mathcal{C}$*.*

Доказательство.

$$\begin{aligned}
V^f(\boldsymbol{\pi}) &= \frac{1}{m^2} \sum_{i \in I} \sum_{j \in J} (c_{\pi_i} - c_{\pi_j})^2 \leqslant \frac{1}{m^2} \sum_{1 \leqslant i < j \leqslant N} (c_i - c_j)^2 \\
&= \left(\frac{N(N-1)}{m^2} \right) \frac{1}{N(N-1)} \sum_{1 \leqslant i < j \leqslant N} (c_i - c_j)^2 \qquad (7) \\
&= \left(\frac{N(N-1)}{m^2} \right) \frac{1}{N-1} \sum_{i=1}^{N} \left(c_i - \frac{1}{N} \sum_{i=1}^{N} c_i \right)^2 = \left(\frac{N}{m} \right)^2 \sigma_N^2,
\end{aligned}$$

где в (7) мы воспользовались Леммой 8. □

Теорема 7 вместе с Леммой 9 дают следующее следствие:

Следствие 3. *Для любого $\varepsilon \geqslant 0$ справедливо:*

$$\mathbb{P}\left\{S'_m - \mathsf{E}[S'_m] \geqslant \varepsilon\right\} \leqslant \exp\left\{-\frac{(N+2)m^2\varepsilon^2}{4N^2\sigma_N^2}\right\} \leqslant \exp\left\{-\frac{m^2\varepsilon^2}{4N\sigma_N^2}\right\}. \tag{8}$$

Аналогичные неравенства справедливы и для $\mathbb{P}\left\{\mathsf{E}[S'_m] - S'_m \geqslant \varepsilon\right\}$. Кроме того, для любого $t > 0$ с вероятностью не меньше $1 - e^{-t}$ справедливо:

$$\mathsf{E}[S'_m] \leqslant S'_m + 2\sqrt{\sigma_N^2\left(\frac{N}{m}\right)\frac{t}{m}}.$$

Важно отметить, что силами Теоремы 6 (фактически являющейся аналогом неравенства МакДиармида) невозможно получить подобный результат, учитывающий дисперсию случайной величины. Сравнивая полученный результат с неравенством Бернштейна (а именно — последним неравенством Теоремы 3), мы видим, что член порядка $1/m$ исчез: теперь у оценки лишь один субгауссовский «режим», который описывает и малые и большие отклонения. В то же время у члена порядка $1/\sqrt{m}$ появился дополнительный множитель $\sqrt{2N/m}$, из-за которого оценка Следствия 3 может вырождаться при больших N и $m = o(N)$.

Возникает вопрос: возможно ли силами Теоремы 7, заменив (6) более точной верхней оценкой, улучшить результат Следствия 3? Следующая лемма дает отрицательный ответ на этот вопрос.

Утверждение 1. *Верхняя оценка Леммы 9 является неулучшаемой.*

Доказательство. Мы приведем пример множества $\mathcal{C}$, для которого неравенство Леммы 9 обращается в равенство. Рассмотрим случай, когда $\mathcal{C} = \{c_1, \ldots, c_N\}$, где $c_1, \ldots, c_m = r$, $c_{m+1}, \ldots, c_N = v$ для двух различных действительных чисел $r, v \in [0, 1]$. В этом случае, как легко проверить, супремум $\sup_{\boldsymbol{\pi}} V^f(\boldsymbol{\pi})$ достигается на тождественной перестановке $\boldsymbol{\pi} = (1, 2, \ldots, N)$. Действительно, в этом случае все $m(N-m)$ слагаемых суммы

$$\sum_{i \in I}\sum_{j}(c_{\pi_i} - c_{\pi_j})^2$$

обращаются в $(r-v)^2$, в то время как для любой другой перестановки $\boldsymbol{\pi}$ некоторые из них будут обращаться в ноль. Таким образом, для такого множества $\mathcal{C}$ спарведливо:

$$\sup_{\boldsymbol{\pi}} V^f(\boldsymbol{\pi}) = \frac{m(N-m)}{m^2}(r-v)^2 = \frac{N-m}{m}(r-v)^2.$$

В то же время, легко проверить, что в этом случае

$$\sigma_N^2 = \frac{m(N-m)}{N^2}(r-v)^2.$$

Это завершает доказательство утверждения. □

Замечание 5. *Результат Следствия 3 не является новым. Неравенство (8) без множителя 4 было впервые получено в [21][Лемма 3.1] для случая $N = m \cdot n$, $n \in \mathbb{N}$ на основе совершенно другого подхода. Автор моделирует реализацию выборки без возвращений с помощью двух последовательных шагов: (1) разбиение множества $\mathcal{C}$ на m непересекающихся подмножеств $\mathcal{C}_1, \ldots, \mathcal{C}_m$; (2) случайный выбор одного элемента из каждого из подмножеств $\mathcal{C}_1, \ldots, \mathcal{C}_m$. При фиксированном разбиении, случайные величины, выбираемые на втором шаге, являются независимыми с математическими ожиданиями $\bar{c}_1, \ldots, \bar{c}_m$ соответственно, где $\bar{c}_i$ — среднее значение подмножества $\mathcal{C}_i$. Совмещая эту модель с методом Чернова и пользуясь неравенством Хефдинга для оценок производящих функций, мы приходим к утверждению теоремы.*

Оказывается, для выборок без возвращений концентрируются сильнее не только суммы, но и ряд более общих функций. Это наглядно продемонстрировано на примере неравенства Эль-Янива, равномерно улучшающего неравенство МакДиармида при росте размера выборки m к N. Также мы привели результат Бобкова, позволяющий получать нетривиальные неравенства концентрации, учитывающие дисперсии случайных величин. В следующем разделе мы рассмотрим конкретный класс функций, известный как супремумы эмпирических процессов, и на основе описанных в прошлых разделах результатов получим для него два новых неравенства концентрации.

4 Супремумы эмпирических процессов

В этом разделе будут рассмотрены супремумы эмпирических процессов, с которыми связан один из ключевых результатов теории неравенств концентрации последних лет — неравенство Талаграна, полученное М. Талаграном в [24] и затем усиленное О. Буске и рядом других авторов в [11, 10]. Неравенство Талаграна лежит в основе большинства современных результатов в индуктивной постановке теории статистического обучения [19, 9], включая *локальные радемахеровские сложности*, использованные в серии работ [20, 6, 18]. В отличие от классического подхода В. Н. Вапника и А. Я. Червоненкиса [1], основанного на равномерных по классу функций оценках сходимости выборочных средних к математическим ожиданиям, локальные подходы позволяют сужать подмножества классов, по которым берутся супремумы, что ведет к существенному улучшению оценок. Подобный анализ становится возможным во многом благодаря учету дисперсии неравенством Талаграна.

Поскольку неравенство Талаграна формулируется для выборки независимых случайных величин, оно перестает выполняться для выборок без возвращений. Таким образом, для применения упомянутого выше локального подхода в рамках трансдуктивного обучения необходимо найти замену неравенству Талаграна. В данном разделе будет получено два новых неравенства концентрации для супремумов эмпирических процессов для выборок без возарвщаений. Первое основано на применении неравенства С. Г. Бобкова Теоремы 7. Второе — на методе Хефдинга, описанном в Разделе 2.1.

Введем необходимые обозначения и определения. Пусть $\mathcal{C} = \{c_1, \ldots, c_N\}$ — некоторое конечное множество. Для $m \leqslant N$ рассмотрим последовательности случайных величин $\{Z_1, \ldots, Z_m\}$ и $\{X_1, \ldots, X_m\}$, выбранные равномерно из $\mathcal{C}$ без возвращений и с возвращениями соответственно. Пусть $\mathcal{F}$ — *счетное* множество отображений $f \colon \mathcal{C} \to \mathbb{R}$, таких что

$\mathsf{E}[f(X_1)] = 0$ и $f(x) \in [-1, 1]$ для всех $f \in \mathcal{F}$ и $x \in \mathcal{C}$. Рассмотрим следующие случайные величины:

$$Q_m = \sup_{f \in \mathcal{F}} \sum_{i=1}^{m} f(X_i), \quad Q'_m = \sup_{f \in \mathcal{F}} \sum_{i=1}^{m} f(Z_i).$$

Случайная величина Q_m известна в литературе как *супремум эмпирического процесса*. Для нее существует ряд неравенств концентрации, наиболее важным из которых является неравенство Талаграна. Здесь мы приведем версию О. Буске с оптимальными константами:

Теорема 10 (О. Буске, [11]). *Положим* $\sigma^2 = \sup_{f \in \mathcal{F}} \mathsf{D}[f(X_1)]$, $v = m\sigma^2 + 2\mathsf{E}[Q_m]$ *и для* $u \geqslant -1$ *определим* $\phi(u) = e^u - u - 1$, $h(u) = (1+u)\log(1+u) - u$. *Тогда для* $\lambda \geqslant 0$ *справедливо:*

$$\mathsf{E}\left[e^{\lambda(Q_m - \mathsf{E}[Q_m])}\right] \leqslant e^{v\phi(\lambda)}. \tag{9}$$

Метод Чернова дает для любых $\varepsilon \geqslant 0$ *следующее неравенство:*

$$\mathbb{P}\left\{Q_m - \mathsf{E}[Q_m] \geqslant \varepsilon\right\} \leqslant e^{-vh(\varepsilon/v)}. \tag{10}$$

Кроме того, для любого $t \geqslant 0$ *с вероятностью не меньше* $1 - e^{-t}$ *справедливо:*

$$Q_m \leqslant \mathsf{E}[Q_m] + \sqrt{2vt} + \frac{t}{3}. \tag{11}$$

Снова воспользовавшись неравенством $h(u) \geqslant \frac{u^2}{2(1+u/3)}$ для $u > 0$, мы можем получить следующий более наглядный результат:

$$\mathbb{P}\left\{Q_m - \mathsf{E}[Q_m] \geqslant \varepsilon\right\} \leqslant \exp\left(-\frac{\varepsilon^2}{2(v + \varepsilon/3)}\right). \tag{12}$$

Важно отметить, что если класс $\mathcal{F} = \{f_0\}$ состоит из одного элемента, неравенство (12) в точности воспроизводит неравенство Бернштейна Теоремы 3 для суммы независимых случайных величин. Таким образом, неравенство Талаграна является равномерным по классу функций $\mathcal{F}$ аналогом неравенства Бернштейна.

Заметим, что случайная величина Q'_m может быть эквивалентно определена с помощью случайных перестановок, так же как рассмотренная в Разделе 2 случайная величина S'_m:

$$Q'_m = Q'_m(\boldsymbol{\pi}) = \sup_{f \in \mathcal{F}} \sum_{i=1}^{m} f(c_{\pi_i}). \tag{13}$$

Кроме того, функция $Q'_m(\boldsymbol{\pi})$, очевидно, (m, u)-симметрична относительно перестановок. Следующий результат, полученный в [15], является простым следствием применения Теоремы 6 к функции Q'_m, удовлетворяющей ее условиям для $\beta = 2$:

Следствие 4 (Эль-Янив, Печиони [15]). *Для любого* $\varepsilon \geqslant 0$ *справедливо:*

$$\mathbb{P}\left\{Q'_m - \mathsf{E}[Q'_m] \geqslant \varepsilon\right\} \leqslant \exp\left\{-\frac{\varepsilon^2}{2m}\left(\frac{N - 1/2}{N - m}\right)\left(1 - \frac{1}{2\max(m, N-m)}\right)\right\}.$$

Аналогичное неравенство справедливо и для $\mathbb{P}\left\{\mathsf{E}[Q'_m] - Q'_m \geqslant \varepsilon\right\}$.

Далее будет получено два новых неравенства концентрации для случайной величины Q'_m. Первый результат основан на Теореме 7 и описывает субгауссовское поведение случайной величины Q'_m:

Теорема 11. *Введем обозначение* $\sigma^2 = \sup_{f\in\mathcal{F}} \mathsf{D}[f(X_1)]$. *Тогда для любого* $\varepsilon \geqslant 0$ *справедливо:*

$$\mathbb{P}\left\{Q'_m - \mathsf{E}[Q'_m] \geqslant \varepsilon\right\} \leqslant \exp\left(-\frac{(N+2)\varepsilon^2}{8N^2\sigma^2}\right). \tag{14}$$

Аналогичное неравенство справедливо и для $\mathbb{P}\left\{\mathsf{E}[Q'_m] - Q'_m \geqslant \varepsilon\right\}$. *Кроме того, для любого* $t \geqslant 0$ *с вероятностью не меньше* $1 - e^{-t}$ *справедливо:*

$$Q'_m \leqslant \mathsf{E}[Q'_m] + 2\sqrt{2N\sigma^2 t}.$$

Доказательство Теоремы 11 основано на применении Теоремы 7 к функции $Q'_m(\boldsymbol{\pi})$. Для этого нам необходимо получить верхнюю оценку для $V^{Q'_m}(\boldsymbol{\pi})$, что является непростой задачей. Вместо этого мы получим новую версию Теоремы 7, которая упростит нашу работу с дискретным градиентом. Для любой (m,u)-симметричной функции $f\colon \boldsymbol{\pi} \to \mathbb{R}$ определим величину, связанную с $V^f(\boldsymbol{\pi})$:

$$V_+^f(\boldsymbol{\pi}) \triangleq \sum_{i\in I}\sum_{j\in J}\big(f(\boldsymbol{\pi}) - f(\boldsymbol{\pi}^{i,j})\big)^2 \mathbb{1}\{f(\boldsymbol{\pi}) \geqslant f(\boldsymbol{\pi}^{i,j})\}.$$

Оказывается, справедлива следующая модификация Теоремы 7:

Теорема 12. [3] *Пусть* $\boldsymbol{\pi}$ *— вектор случайной перестановки, выбранной равномерно из симметрической группы перестановок множества* $\{1,\dots,N\}$. *Пусть* $f(\boldsymbol{\pi})$ *—* (m,u)*-симметричная функция и* $\Sigma^2 \geqslant 0$ *— действительное число, такое что* $V_+^f(\boldsymbol{\pi}) \leqslant \Sigma^2$ *для всех* $\boldsymbol{\pi}$. *Тогда справедливо:*

$$\mathbb{P}\left\{f(\boldsymbol{\pi}) - \mathsf{E}[f(\boldsymbol{\pi})] \geqslant \varepsilon\right\} \leqslant \exp\left\{-\frac{(N+2)\varepsilon^2}{8\Sigma^2}\right\}. \tag{15}$$

Аналогичное неравенство справедливо и для $\mathbb{P}\left\{\mathsf{E}[f(\boldsymbol{\pi})] - f(\boldsymbol{\pi}) \geqslant \varepsilon\right\}$.

Доказательство. Мы будем следовать шагам доказательства Теоремы 7, представленным в [8] (см. первое неравенство Теоремы 2.1). Там показано, что для любой (m,u)-симметричной функции $g(\boldsymbol{\pi})$ справедливо:

$$\begin{aligned}(N+2)\big(\mathsf{E}\big[e^{g(\boldsymbol{\pi})}\log e^{g(\boldsymbol{\pi})}\big] - \mathsf{E}\big[e^{g(\boldsymbol{\pi})}\big]\mathsf{E}\big[\log e^{g(\boldsymbol{\pi})}\big]\big)\\ \leqslant \mathsf{E}\left[\sum_{i\in I}\sum_{j\in J}\big(g(\boldsymbol{\pi}) - g(\boldsymbol{\pi}^{i,j})\big)\big(e^{g(\boldsymbol{\pi})} - e^{g(\boldsymbol{\pi}^{i,j})}\big)\right],\end{aligned} \tag{16}$$

где, как и раньше, $I = \{1,\dots,m\}$ и $J = \{m+1,\dots,N\}$. Заметим также, что для любых $a, b \in \mathbb{R}$:

$$(a-b)(e^a - e^b) \leqslant \frac{e^a + e^b}{2}(a-b)^2. \tag{17}$$

[3] Более слабая версия этого результата была представлена в [4].

Обозначим симметрическую группу перестановок множества $\{1, \dots, N\}$ с помощью $\mathcal{S}(N)$. Перепишем правую часть неравенства (16) следующим образом:

$$\begin{aligned}
&\mathsf{E}\left[\sum_{i\in I}\sum_{j\in J}\big(g(\boldsymbol{\pi}) - g(\boldsymbol{\pi}^{i,j})\big)\big(e^{g(\boldsymbol{\pi})} - e^{g(\boldsymbol{\pi}^{i,j})}\big)\right] \\
&= \frac{1}{N!}\sum_{\boldsymbol{\pi}\in\mathcal{S}(N)}\left[\sum_{i\in I}\sum_{j\in J}\big(g(\boldsymbol{\pi}) - g(\boldsymbol{\pi}^{i,j})\big)\big(e^{g(\boldsymbol{\pi})} - e^{g(\boldsymbol{\pi}^{i,j})}\big)\right] \\
&= \frac{2}{N!}\sum_{\boldsymbol{\pi}\in\mathcal{S}(N)}\sum_{i\in I}\sum_{j\in J}\big(g(\boldsymbol{\pi}) - g(\boldsymbol{\pi}^{i,j})\big)\big(e^{g(\boldsymbol{\pi})} - e^{g(\boldsymbol{\pi}^{i,j})}\big)\mathbb{1}\{g(\boldsymbol{\pi}) \geqslant g(\boldsymbol{\pi}^{i,j})\}.
\end{aligned}$$

Воспользовавшись неравенством (17), мы получаем:

$$\begin{aligned}
&\mathsf{E}\left[\sum_{i\in I}\sum_{j\in J}\big(g(\boldsymbol{\pi}) - g(\boldsymbol{\pi}^{i,j})\big)\big(e^{g(\boldsymbol{\pi})} - e^{g(\boldsymbol{\pi}^{i,j})}\big)\right] \\
&\leqslant \frac{2}{N!}\sum_{\boldsymbol{\pi}\in\mathcal{S}(N)}\sum_{i\in I}\sum_{j\in J}\frac{\big(e^{g(\boldsymbol{\pi})} + e^{g(\boldsymbol{\pi}^{i,j})}\big)}{2}\big(g(\boldsymbol{\pi}) - g(\boldsymbol{\pi}^{i,j})\big)^2\mathbb{1}\{g(\boldsymbol{\pi}) \geqslant g(\boldsymbol{\pi}^{i,j})\} \\
&\leqslant \frac{2}{N!}\sum_{\boldsymbol{\pi}\in\mathcal{S}(N)}\sum_{i\in I}\sum_{j\in J}e^{g(\boldsymbol{\pi})}\big(g(\boldsymbol{\pi}) - g(\boldsymbol{\pi}^{i,j})\big)^2\mathbb{1}\{g(\boldsymbol{\pi}) \geqslant g(\boldsymbol{\pi}^{i,j})\} \\
&= 2\mathsf{E}\left[V_+^g(\boldsymbol{\pi})e^{g(\boldsymbol{\pi})}\right].
\end{aligned}$$

Таким образом, справедливо следующее:

$$(N+2)\big(\mathsf{E}\big[e^{g(\boldsymbol{\pi})}\log e^{g(\boldsymbol{\pi})}\big] - \mathsf{E}\big[e^{g(\boldsymbol{\pi})}\big]\mathsf{E}\big[\log e^{g(\boldsymbol{\pi})}\big]\big) \leqslant 2\mathsf{E}\left[V_+^g(\boldsymbol{\pi})e^{g(\boldsymbol{\pi})}\right].$$

Применяя это неравенство к функции λf для произвольного $\lambda \in \mathbb{R}$, мы получаем:

$$(N+2)\big(\mathsf{E}\big[e^{\lambda f(\boldsymbol{\pi})}\log e^{\lambda f(\boldsymbol{\pi})}\big] - \mathsf{E}\big[e^{\lambda f(\boldsymbol{\pi})}\big]\mathsf{E}\big[\log e^{\lambda f(\boldsymbol{\pi})}\big]\big) \leqslant 2\mathsf{E}\left[V_+^{\lambda f}(\boldsymbol{\pi})e^{\lambda f(\boldsymbol{\pi})}\right] \leqslant 2\Sigma^2\lambda^2\mathsf{E}\left[e^{\lambda f(\boldsymbol{\pi})}\right]. \tag{18}$$

В доказательстве Теоремы 7 в [8] отмечено, что неравенство (18) влечет за собой следующую верхнюю оценку на производящую функцию моментов:

$$\mathsf{E}\left[e^{\lambda(f-\mathsf{E}[f])}\right] \leqslant e^{\frac{2\Sigma^2\lambda^2}{N+2}}. \tag{19}$$

Этот факт известен в литературе как «метод Хербста» и лежит в основе энтропийного подхода. Теперь мы применяем метод Чернова, описанный ранее, который дает нам для любого $\lambda, \varepsilon \geqslant 0$:

$$\mathbb{P}\left\{f(\boldsymbol{\pi}) - \mathsf{E}[f(\boldsymbol{\pi})] \geqslant \varepsilon\right\} \leqslant \frac{\mathsf{E}\left[e^{\lambda(f-\mathsf{E}[f])}\right]}{e^{\lambda\varepsilon}} \leqslant e^{\frac{2\Sigma^2\lambda^2}{N+2} - \lambda\varepsilon}.$$

Нам остается минимизировать правую часть последнего неравенства по λ, что достигается при $\lambda = \frac{\varepsilon(N+2)}{4\Sigma^2}$, и мы получаем (15).

Неравенство для левого хвоста распределения может быть получено аналогичным образом, используя (19) при $\lambda < 0$:

$$\mathbb{P}\left\{\mathsf{E}[f(\boldsymbol{\pi})] - f(\boldsymbol{\pi}) \geqslant \varepsilon\right\} = \mathbb{P}\left\{\lambda\big(f(\boldsymbol{\pi}) - \mathsf{E}[f(\boldsymbol{\pi})]\big) \geqslant -\lambda\varepsilon\right\} \leqslant \frac{\mathsf{E}\left[e^{\lambda(f-\mathsf{E}[f])}\right]}{e^{-\lambda\varepsilon}} \leqslant e^{\frac{2\Sigma^2\lambda^2}{N+2}+\lambda\varepsilon}.$$

На этот раз мы полагаем $\lambda = -\frac{\varepsilon(N+2)}{4\Sigma^2}$, что завершает доказательство. □

Замечание 6. *Использованная в доказательстве идея отдельной оценки положительных и отрицательных приращений часто используется в литературе (например в Теореме 6.16 [10]).*

Доказательство Теоремы 11. Рассмотрим две функции $f, g\colon \mathcal{A} \to \mathbb{R}$, определенных на произвольном множестве $\mathcal{A}$, и предположим, что $\sup_{a\in\mathcal{A}} f(a) = f(\bar{a})$ для некоторого $\bar{a} \in \mathcal{A}$. Тогда справедливо:

$$\left(\sup_{a\in\mathcal{A}} f(a) - \sup_{a\in\mathcal{A}} g(a)\right)^2 \mathbb{1}\left\{\sup_{a\in\mathcal{A}} f(a) \geqslant \sup_{a\in\mathcal{A}} g(a)\right\} \leqslant \big(f(\bar{a}) - g(\bar{a})\big)^2. \tag{20}$$

Предположим, что супремум в определении (13) достигается на функции $\bar{f} \in \mathcal{F}$. Тогда из (20) следует:

$$\begin{aligned}
V_+^{Q'_m}(\boldsymbol{\pi}) &= \sum_{i\in I}\sum_{j\in J}\big(Q'_m(\boldsymbol{\pi}) - Q'_m(\boldsymbol{\pi}^{i,j})\big)^2 \mathbb{1}\left\{Q'_m(\boldsymbol{\pi}) \geqslant Q'_m(\boldsymbol{\pi}^{i,j})\right\} \\
&\leqslant \sum_{i\in I}\sum_{j\in J}\left(\sum_{k=1}^{m}\bar{f}(c_{\pi_k}) - \sum_{k=1}^{m}\bar{f}\big(c_{\pi_k^{i,j}}\big)\right)^2 \\
&= \sum_{i\in I}\sum_{j\in J}\Big(\bar{f}(c_{\pi_i}) - \bar{f}(c_{\pi_j})\Big)^2 \\
&\leqslant \sum_{1\leqslant i<j\leqslant N}\Big(\bar{f}(c_i) - \bar{f}(c_j)\Big)^2 = N^2\mathsf{D}[\bar{f}(X_1)],
\end{aligned}$$

где мы воспользовались Леммой 8. Поскольку $\bar{f}$ зависит от выбора $\boldsymbol{\pi}$, мы получаем:

$$V_+^{Q'_m}(\boldsymbol{\pi}) \leqslant N^2 \sup_{f\in\mathcal{F}} \mathsf{D}[f(X_1)].$$

Применение Теоремы 12 завершает доказательство. □

Замечание 7. *В работе [15] авторы пользуются Следствием 4 для воспроизведения в трансдуктивном подходе результатов индуктивного подхода, основанных на* глобальной рademахеровской сдложности. *Однако, важно отметить, что Теоремы 6 не достаточно для анализа, основанного на локальной рademахеровской сложности [20, 6, 18], который требует учета дисперсий случайных величин. Неравенство Теоремы 6, как и неравенство МакДиармида, основано лишь на ограниченности случайных величин и не учитывает их дисперсий.*

Следующий результат основан на непосредственном применении метода Хефдинга, описанного в Разделе 2.1.

Теорема 13. *Положим* $\sigma^2 = \sup_{f\in\mathcal{F}} \mathsf{D}[f(X_1)]$, $v = m\sigma^2 + 2\mathsf{E}[Q_m]$ *и для* $u \geqslant -1$ *определим* $\phi(u) = e^u - u - 1$, $h(u) = (1+u)\log(1+u) - u$. *Тогда для* $\varepsilon \geqslant \mathsf{E}[Q_m] - \mathsf{E}[Q'_m] \geqslant 0$ *справедливо:*

$$\mathbb{P}\left\{Q'_m - \mathsf{E}[Q'_m] \geqslant \varepsilon\right\} \leqslant \exp\left(-vh\left(\frac{\varepsilon - \mathsf{E}[Q_m] + \mathsf{E}[Q'_m]}{v}\right)\right) \tag{21}$$

$$\leqslant \exp\left(-\frac{(\varepsilon - \mathsf{E}[Q_m] + \mathsf{E}[Q'_m])^2}{2v + \frac{2}{3}(\varepsilon - \mathsf{E}[Q_m] + \mathsf{E}[Q'_m])}\right). \tag{22}$$

Кроме того, для любого $t \geqslant 0$ *с вероятностью не менее* $1 - e^{-t}$ *справедливо:*

$$Q'_m \leqslant \mathsf{E}[Q_m] + \sqrt{2vt} + \frac{t}{3}. \tag{23}$$

Мы приведем доказательство для конечного множества $\mathcal{F} = \{f_1, \ldots, f_M\}$. Как отмечено в доказательстве Теоремы 2.11 [8], случай счетного $\mathcal{F}$ доказывается простым переходом к пределу при $M \to \infty$. Для доказательства нам понадобится следующая техническая лемма.

Лемма 14. *Пусть* $\boldsymbol{x} = (x_1, \ldots, x_d)^{\mathrm{T}} \in \mathbb{R}^d$. *Тогда для всех* $\lambda > 0$ *функция*

$$F(\boldsymbol{x}) = \exp\left(\lambda \sup_{i=1,\ldots,d} x_i\right)$$

является выпуклой.

Доказательство. Для начала покажем, что если $g\colon \mathbb{R} \to \mathbb{R}$ — выпуклая и неубывающая функция, а $f\colon \mathbb{R}^d \to \mathbb{R}$ выпукла, то $g\big(f(\boldsymbol{x})\big)$ тоже выпукла. Действительно, для $\alpha \in [0,1]$ и $\boldsymbol{x}', \boldsymbol{x}'' \in \mathbb{R}^d$:

$$g\Big(f\big(\alpha\boldsymbol{x}' + (1-\alpha)\boldsymbol{x}''\big)\Big) \leqslant g\big(\alpha f(\boldsymbol{x}') + (1-\alpha)f(\boldsymbol{x}'')\big) \leqslant \alpha g\big(f(\boldsymbol{x}')\big) + (1-\alpha)\alpha g\big(f(\boldsymbol{x}'')\big).$$

С учетом того, что $g(y) = e^{\lambda y}$ — выпукла и возрастает для $\lambda > 0$, нам остается показать, что $f(\boldsymbol{x}) = \sup_{i=1,\ldots,d}(x_i)$ выпукла. Но для любого $\alpha \in [0,1]$ и $\boldsymbol{x}', \boldsymbol{x}'' \in \mathbb{R}^d$ справедливо:

$$\sup_{i=1,\ldots,d}\big(\alpha x'_i + (1-\alpha)x''_i\big) \leqslant \alpha \sup_{i=1,\ldots,d} x'_i + (1-\alpha)\sup_{i=1,\ldots,d} x''_i.$$

Это завершает доказательство леммы. □

Доказательство Теоремы 13. Пусть последовательности $\{U_1, \ldots, U_m\}$ и $\{W_1, \ldots, W_m\}$ выбраны равномерно с и без возвращений соответственно из конечного множества M-мерных векторов $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N\} \subset \mathbb{R}^M$, где $\boldsymbol{v}_j = \big(f_1(c_j), \ldots, f_M(c_j)\big)^{\mathrm{T}}$. Пользуясь Леммой 14 и Теоремой 1, мы получаем для любого $\lambda > 0$:

$$\mathsf{E}\left[e^{\lambda Q'_m}\right] = \mathsf{E}\left[\exp\left(\lambda \sup_{j=1,\ldots,M}\left(\sum_{i=1}^m W_i\right)_j\right)\right] \leqslant \mathsf{E}\left[\exp\left(\lambda \sup_{j=1,\ldots,M}\left(\sum_{i=1}^m U_i\right)_j\right)\right] = \mathsf{E}\left[e^{\lambda Q_m}\right], \tag{24}$$

где нижним индексом j мы обозначили j-ую координату вектора. Воспользовавшись оценкой (9), мы ограничиваем производящую функцию моментов, стоящую в правой части неравенства (24):

$$\mathsf{E}\left[e^{\lambda Q'_m}\right] \leqslant \mathsf{E}\left[e^{\lambda Q_m}\right] \leqslant e^{\lambda \mathsf{E}[Q_m] + v\phi(\lambda)},$$

или

$$\mathsf{E}\left[e^{\lambda (Q'_m - \mathsf{E}[f'_m])}\right] \leqslant e^{\lambda (\mathsf{E}[Q_m] - \mathsf{E}[Q'_m]) + v\phi(\lambda)}.$$

Воспользовавшись методом Чернова, мы получаем для $\varepsilon \geqslant 0$ и $\lambda > 0$:

$$\mathbb{P}\left\{Q'_m - \mathsf{E}[Q'_m] \geqslant \varepsilon\right\} \leqslant \frac{\mathsf{E}\left[e^{\lambda (Q'_m - \mathsf{E}[Q'_m])}\right]}{e^{\lambda\varepsilon}} \leqslant \exp\bigl(\lambda(\mathsf{E}[Q_m] - \mathsf{E}[Q'_m]) + v\phi(\lambda) - \lambda\varepsilon\bigr). \tag{25}$$

Правая часть последнего неравенства выпукла и достигает своего минимума при

$$\lambda = \log\left(\frac{v + \varepsilon - \mathsf{E}[Q_m] + \mathsf{E}[Q'_m]}{v}\right), \tag{26}$$

откуда и берется техническое условие $\varepsilon \geqslant \mathsf{E}[Q_m] - \mathsf{E}[Q'_m]$. В противном случае мы полагаем $\lambda = 0$ и получаем тривиальную оценку 1. Кроме того, $\mathsf{E}[Q_m] \geqslant \mathsf{E}[Q'_m]$ следует также из Теоремы 1 с учетом Леммы 14. Подстановка (26) в (25) завершает доказательство первого неравенства Теоремы 13. Неравенства (22) и (23) вытекают из (21) так же, как неравенства (12) и (11) из (10). □

Мы видим, что применение метода Хефдинга привело к появлению технического условия на параметр ε. Таким образом, Теорема 13 дает верхнюю оценку отклонения случайной величины Q'_m не от своего математического ожидания $\mathsf{E}[Q'_m]$, а от превосходящего его математического ожидания $\mathsf{E}[Q_m]$. При рассмотрении сумм в Следствии 1 такого эффекта не наблюдалось благодаря совпадению математических ожиданий. В общем случае, очевидно, $\mathsf{E}[Q_m]$ и $\mathsf{E}[Q'_m]$ не равны. Однако, справедлив следующий результат, показывающий, что в ряде случаев условие на ε в Теореме 13 не является строгим:

Лемма 15. *Справедливо следующее:*

$$0 \leqslant \mathsf{E}[Q_m] - \mathsf{E}[Q'_m] \leqslant 2\frac{m^3}{N}.$$

Доказательство. Доказательство первого неравенства было приведено ранее в доказательстве Теоремы 13. Перейдем к доказательству второго. Воспользовавшись определением, запишем:

$$\mathsf{E}[Q_m] - \mathsf{E}[Q'_m] = \frac{1}{N^m} \sum_{x_1,\ldots,x_m} \sup_{f\in\mathcal{F}} \sum_{i=1}^{m} f(x_i) + \left(\frac{1}{N^m} - \frac{(N-m)!}{N!}\right) \sum_{z_1,\ldots,z_m} \sup_{f\in\mathcal{F}} \sum_{i=1}^{m} f(z_i),$$

где первая сумма берется по всем упорядоченным последовательностям $(x_1, \ldots, x_m)$, в которых есть повторяющиеся члены, а вторая — по упорядоченным последовательностям $(z_1, \ldots, z_m)$ без повторяющихся членов. Несложно убедиться, что во второй сумме всего

$m! \cdot C_N^m$ слагаемых, а в первой, следовательно, $N^m - m! \cdot C_N^m$ слагаемых. С учетом того, что $\frac{1}{N^m} \leqslant \frac{(N-m)!}{N!}$ и $f(x) \in [-1, 1]$ для всех x, мы получаем:

$$\begin{aligned}
\mathsf{E}[Q_m] - \mathsf{E}[Q'_m] &\leqslant m\left(\frac{N^m - m! \cdot C_N^m}{N^m}\right) + m\left(\frac{(N-m)!}{N!} - \frac{1}{N^m}\right) m! \cdot C_N^m \\
&= 2m\left(\frac{N^m - m! \cdot C_N^m}{N^m}\right) \\
&= 2m - 2m\left(1 \cdot \left(1 - \frac{1}{N}\right) \cdots \left(1 - \frac{m-1}{N}\right)\right).
\end{aligned}$$

Оценивая все множители второго слагаемого снизу, мы получаем:

$$\begin{aligned}
\mathsf{E}[Q_m] - \mathsf{E}[Q'_m] &\leqslant 2m - 2m\left(1 - \frac{m-1}{N}\right)^m \\
&= 2m\left(\frac{m-1}{N}\right)\left(1 + \left(1 - \frac{m-1}{N}\right) + \cdots + \left(1 - \frac{m-1}{N}\right)^{m-1}\right) \\
&\leqslant 2m\left(\frac{m-1}{N}\right) m \\
&\leqslant 2\frac{m^3}{N}.
\end{aligned}$$

□

Приведем короткий анализ и сравнение трех неравенств, представленных в настоящем разделе.

Неравенство Следствия 4, в отличие от результатов Теорем 11 и 13, не учитывает дисперсий случайных величин и основано лишь на предположении об ограниченности функций из $\mathcal{F}$. Тем не менее, применение неравенства МакДиармида Теоремы 5 для случайной величины Q_m (для выборок *с возвращениями*) дает более слабый результат, в особенности для размеров выборки m, близких к N. Этот эффект, уже обсуждавшийся в прошлом разделе, ведет к более сильной концентрации случайной величины Q'_m по сравнению с Q_m. С точки зрения теории статистического обучения это означает, что свойства задачи трансдуктивного обучения выгодно отличаются от свойств индуктивной постановки. В частности, этот факт служил одной из мотиваций работы [15]. Тем не менее, возможность использования этого преимущества на практике для получения более точных процедур обучения остается по-прежнему открытым вопросом.

Неравенства Теорем 11 и 13 учитывают дисперсию случайных величин. Первое является субгауссовским неравенством и дает совпадающие оценки для обоих хвостов распределения случайной величины Q'_m. Грубое сравнение неравенств Следствия 4 и Теоремы 11 показывает, что для больших N и $m = o(N)$ неравенство Теоремы 11 вырождается, в то время как неравенство Следствия 4 может давать нетривиальный результат. Однако, в случае $m = O(N)$, более актуальном в трансдуктивном обучении, ситуация меняется. Например, при $m = N/2$, результат сравнения зависит от соотношения величин $-\varepsilon^2/m$ и $-\varepsilon^2/(16\, m\sigma^2)$, и неравенство Теоремы 11 оказывается точнее уже при $\sigma^2 < 1/16$.

Сравнение неравенств Теорем 11 и 13 является менее очевидным. Для больших N и $m = o(N)$ неравенство (14) в отличие от (22) может вырождаться. С другой стороны, Теорема 13 накладывает техническое ограничение на величину отклонения ε. Последовательно воспользовавшись элементарными неравенствами $\sqrt{a+b} \leqslant \sqrt{a} + \sqrt{b}$ и $\sqrt{ab} \leqslant \frac{a+b}{2}$, правую часть неравенства (23) можно оценить сверху с помощью выражения $2\mathsf{E}[Q_m] + \sqrt{2m\sigma^2 t} + Ct$ для некоторой константы C. Поскольку типичным порядком величины $\mathsf{E}[Q_m]$ является $\sqrt{m}$ (например, для случая конечного множества $\mathcal{F} = \{f_1, \ldots, f_M\}$ это вытекает из Теорем 2.1 и 3.5 работы [19]), оценка неравенства (23) также имеет порядок $\sqrt{m}$. С учетом неравенства $\mathsf{E}[Q_m] \geqslant \mathsf{E}[Q'_m]$ эти размышления свидетельствуют о том, что условие на ε ограничивает область применимости Теоремы 13 не слишком сильно.

В заключение данного раздела отметим, что, к сожалению, неравенства для Q'_m Теорем 11 и 13, в отличие от Следствия 4, не ведут к более точным оценкам по сравнению с известными результатами для Q_m при росте $m \to N$. Существование подобных неравенств концентрации позволило бы надеяться на улучшение (хоть и незначительное) локального радемахеровского анализа в трансдуктивном подходе и является интересным вопросом для дальнейших исследований.

5 Заключение

В данной работе представлен подробный обзор известных в литературе результатов о концентрации значений функций, зависящих от выборок без возвращений, вблизи их математических ожиданий. На основе описанных подходов получено два новых неравенства концентрации для супремумов эмпирических процессов для выборок без возвращений. В отличие от всех предыдущих результатов оба неравенства учитывают дисперсию $\sup_{f \in \mathcal{F}} \mathsf{D}[f(X)]$ случайных величин, что часто ведет к существенным улучшениям.

Неравенства концентрации для супремумов эмпирических процессов, учитывающие дисперсию, (среди которых неравенство Талаграна играет особую роль) лежат в основе современных подходов в индуктивной постановке теории статистического обучения, основанных на локальном радемахеровском анализе. Подобные подходы, в частности, позволяют получать *быстрые* (порядка $o(1/n^{-1/2})$) скорости сходимости среднего риска предиктора, обученного с помощью метода минимизации эмпирического риска, к минимальному по всему классу среднему риску [19, 9]. Результаты настоящей работы позволяют впервые применить подобный анализ в рамках трансдуктивного подхода, что является главным направлением дальнейших исследований.

Кроме того, интересным направлением является возможность получения неравенства концентрации для выборок без возвращений, равномерно улучшающего неравенство Талаграна при росте размера выборки $m \to N$, подобно тому как неравенство Эль-Янива и Печиони улучшает классическое неравенство МакДиармида. Этот вопрос, по-видимому, является непростым и ведет к другому более общему вопросу о возможности применения энтропийного метода М. Леду для выборок без возвращений.

Автор выражает благодарности С. Г. Бобкову и С. Н. Минскеру за ценные советы и плодотворные обсуждения.

A Леммы Хефдинга и Бернштейна

Лемма 16 (Хефдинг [10]). *Пусть для случайной величины ξ выполнено $\mathsf{E}[\xi] = 0$ и $a \leqslant \xi \leqslant b$ для некоторых $a, b \in \mathbb{R}$. Тогда для любого $\lambda > 0$ справедливо:*

$$\mathsf{E}\left[e^{\lambda\xi}\right] \leqslant e^{\lambda^2(b-a)^2/8}.$$

Лемма 17 (Бернштейн [10]). *Пусть для случайной величины ξ выполнено $\mathsf{E}[\xi] = 0$ и $|\xi| \leqslant c$ для некоторой $c \in \mathbb{R}$. Тогда для любого $\lambda > 0$ справедливо:*

$$\mathsf{E}\left[e^{\lambda\xi}\right] \leqslant \exp\left\{\mathsf{E}[\xi^2]\left(\frac{e^{\lambda c} - 1 - \lambda c}{c^2}\right)\right\}.$$

Список литературы

[1] *Вапник В. Н., Червоненкис А. Я.* О равномерной сходимости частот появления событий к их вероятностям // *Теория вероятностей и ее применения.* — 1971. — Т. 16, № 2. — С. 264–280.

[2] *Вапник В. Н., Червоненкис А. Я.* Теория распознавания образов. — М.: Наука, 1974.

[3] *Воронцов К. В.* Комбинаторная теория переобучения: результаты, приложения и открытые проблемы // Математические методы распознавания образов: 15-ая Всеросс. конф.: Докл. — М.: МАКС Пресс, 2011. — С. 40-43.

[4] *Толстихин И. О.* Локализация оценок избыточного риска в комбинаторной теории переобучения // Междунар. конф. ИОИ-9. — 2012. — С. 54–57.

[5] *R. Bardenet, O. A. Maillard.* Concentration inequalities for sampling without replacement // http://arxiv.org/abs/1309.4029. — 2013.

[6] *P. Bartlett, O. Bousquet, S. Mendelson.* Local Rademacher Complexities // The Annals of Statistics. — 2005. — V. 33, No. 4. — Pp. 1497–1537.

[7] *A. Blum, J. Langford.* PAC-MDL Bounds // Proceedings of the International Conference on Computational Learning Theory (COLT), 2003.

[8] *S. Bobkov.* Concentration of normalized sums and a central limit theorem for noncorrelated random variables // Annals of Probability. — 2004. — V. 32.

[9] *S. Boucheron, G. Lugosi, O. Bousquet.* Theory of Classification: a Survey of Recent Advances // ESAIM: Probability and Statistics. — 2005. — No. 9. — Pp. 323–375.

[10] *S. Boucheron, G. Lugosi, P. Massart.* Concentration Inequalities: A Nonasymptotic Theory of Independence. — Oxford University Press, 2013.

[11] *O. Bousquet.* Concentration Inequalities and Empirical Processes Theory Applied to the Analysis of Learning Algorithms. — PhD thesis, Ecole Polytechnique, 2002.

[12] *O. Bousquet, S. Boucheron, G. Lugosi.* Introduction to statistical learning theory // Lecture Notes in Artificial Intelligence, 2004.

[13] *C. Cortes, M. Mohri, D. Pechyony, A. Rastogi.* Stability Analysis and Learning Bounds for Transductive Regression Algorithms // http://arxiv.org/abs/0904.0814. — 2009.

[14] *P. Derbeko, R. El-Yaniv, R. Meir.* Explicit Learning Curves for Transduction and Application to Clustering and Compression Algorithms // Journal of Artificial Intelligence Research. — 2004. — V. 22.

[15] *R. El-Yaniv, D. Pechyony.* Transductive Rademacher Complexity and its Applications // Journal of Artificial Intelligence Research. — 2009.

[16] *D. Gross, V. Nesme.* Note on sampling without replacing from a fnite collection of matrices // http://arxiv.org/abs/1001.2738v2. — 2010.

[17] *W. Hoeffding.* Probability inequalities for sums of bounded random variables // Journal of the American Statistical Association. — 1963. — V. 58, No. 301. — Pp. 13–30.

[18] *V. Koltchinskii.* Local Rademacher Complexities and Oracle Inequalities in Risk Minimization // The Annals of Statistics. — 2006. — V. 34, No. 6. — Pp. 2593–2656.

[19] *V. Koltchinskii.* Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: École D'Été de Probabilités de Saint-Flour XXXVIII-2008 // Ecole d'été de probabilités de Saint-Flour. Springer, 2011.

[20] *V. Koltchinskii, D. Panchenko.* Rademacher processes and bounding the risk of function learning // High Dimensional Probability, II / Ed. by D. E. Gine, J. Wellner. — Birkhauser, 1999. — Pp. 443–457.

[21] *M. Pascal.* Rates of convergence in the central limit theorem for empirical processes // Annales de l'institut Henri Poincaré (B) Probabiliés et Statistiques. — 1986. — V. 22, No. 4. — Pp. 381–423.

[22] *C. McDiarmid.* On the method of bounded differences // Surveys in Combinatorics. — 1989. — Pp. 148–188.

[23] *R. J. Serfling.* Probability inequalities for the sum in sampling without replacement // The Annals of Statistics. — 1974. — V. 1, No. 1. — Pp. 39–48.

[24] *M. Talagrand.* New concentration inequalities in product spaces // Inventiones Mathematicae. — 1996. — V. 126.

[25] *V. Vapnik.* Statistical Learning Theory. — New York: Wiley, 1998.

[26] *K. V. Vorontsov.* Exact combinatorial bounds on the probability of overfitting for empirical risk minimization // Pattern Recognition and Image Analysis. — 2010. — Vol. 20, No. 3. — Pp. 269–285.

[27] *K. V. Vorontsov, A. A. Ivahnenko.* Tight combinatorial generalization bounds for threshold conjunction rules // 4-th Int'l Conf. on Pattern Recognition and Machine Intelligence (PReMI'11). Lecture Notes in Computer Science, Springer-Verlag, pp. 66–73, 2011.

[28] *K. V. Vorontsov, A. I. Frey, E. A. Sokolov.*, Computable Combinatorial Overitting Bounds // Machine Learning and Data Analysis — 2013. — Vol. 1, No. 6. — Pp. 734-743.