ПРОБЛЕМЫ ПЕРЕДАЧИ ИНФОРМАЦИИ

Том 56 **2020** **Вып. 2**

УДК 621.391.1 : 519.72

© 2020 г. **А. Макур, Л. Чжэн**

СРАВНЕНИЕ КОЭФФИЦИЕНТОВ СЖАТИЯ ДЛЯ f-ДИВЕРГЕНЦИЙ[1,2]

Коэффициенты сжатия – это зависящие от распределений константы, использующиеся для улучшения стандартных неравенств об обработке данных для f-дивергенций (или относительных f-энтропий) и приводящие к так называемым "сильным" неравенствам об обработке данных. Для любого двумерного совместного распределения, т.е. пары, состоящей из вектора вероятностей и стохастической матрицы, известно, что коэффициенты сжатия для f-дивергенций ограничены сверху единицей, а снизу – коэффициентом сжатия для χ^2-дивергенции. Мы показываем, что верхняя граница достигается, когда совместное распределение разложимо, а нижней можно достичь, устремляя f-дивергенции на входе для коэффициентов сжатия к нулю. Затем устанавливается линейная верхняя граница на коэффициенты сжатия совместных распределений для некоторого класса f-дивергенций через коэффициенты сжатия для χ^2-дивергенции, причем эта граница уточняется для выделенного специального случая дивергенции Кульбака – Лейблера (КЛ-дивергенции). Далее дается альтернативное доказательство того факта, что коэффициенты сжатия для КЛ- и χ^2-дивергенций совпадают для двумерных гауссовских распределений (где для первого коэффициента может налагаться ограничение на второй момент). Наконец, обобщается известный результат о том, что коэффициенты сжатия для стохастических матриц (после вычисления экстремума по всевозможным векторам вероятностей) для всех нелинейных операторно выпуклых f-дивергенций равны. В частности, доказывается, что так называемый предпорядок "меньшего искажения" на стохастических матрицах эквивалентным образом характеризуется любой нелинейной операторно выпуклой f-дивергенцией. Как приложение этой характеризации, выводится также обобщение сильного неравенства Самородницкого об обработке данных.

Ключевые слова: коэффициент сжатия, f-дивергенция/относительная f-энтропия, сильное неравенство об обработке данных, предпорядок меньшего искажения, максимальная корреляция.

DOI: 10.31857/S0555292320020011

§ 1. Введение

Коэффициенты сжатия для f-дивергенций (или относительных f-энтропий) широко изучались в теории информации [1–8], статистике [9–13], анализе цепей Маркова в теории вероятностей [14–16], а также в теории матриц [17–19], причем особенно важными частными случаями являются расстояние по вариации, дивергенция Кульбака – Лейблера (КЛ-дивергенция) и взаимная информация, а также χ^2-дивергенция. Как будет показано на примерах, эти коэффициенты – зависящие от распре-

[1] Работа выполнена при финансовой поддержке фонда NSF (грант 1216476) и стипендии компании Хьюлетт-Паккард.

[2] Весьма предварительная версия настоящей статьи была представлена в докладе [1].

делений константы, использующиеся для улучшения обычных неравенств об обработке данных для f-дивергенций и приводящие к так называемым "сильным" неравенствам об обработке данных. Коэффициенты сжатия для f-дивергенций бывают двух типов: относящиеся к парам из вектора вероятностей и стохастической матрицы, т.е. совместным распределениям, и относящиеся только к стохастическим матрицам, т.е. условным распределениям. Цель настоящей статьи в широком смысле – изучить различные неравенства и равенства между коэффициентами сжатия в обеих постановках. В постановке, связанной с совместными распределениями, мы главным образом устанавливаем общие границы на коэффициенты сжатия для определенных классов f-дивергенций, а также выводим специфические границы на коэффициенты сжатия для КЛ-дивергенции через коэффициенты сжатия для χ^2-дивергенции (или квадрата максимальной корреляции Хиршфельда – Гебелейна – Реньи). С другой стороны, в постановке, связанной со стохастическими матрицами, мы обобщаем известную эквивалентность некоторых определенных коэффициентов сжатия, доказывая эквивалентность характеризации предпорядка "меньшего искажения" (less noisy preorder) на стохастических матрицах [20].

Более точно, настоящая статья включают в себя следующее:

1. Полностью независимый обзор результатов о коэффициентах сжатия в § 2 (в котором мы в основном рассматриваем свойства коэффициентов сжатия, а не их применения).
2. Обобщения известных свойств коэффициентов сжатия для совместных распределений, такие как результат о разложимости в теореме 1 (или в п. 3 предложения 3), мета-сильное неравенство об обработке данных в п. 6 предложения 3 и характеризация коэффициента сжатия для χ^2-дивергенции через коэффициенты сжатия для f-дивергенций при стремлении f-дивергенции на входе к нулю в теореме 2 (что явно объясняет интуитивное понимание нижней границы на максимальную корреляцию в п. 7 предложения 3).
3. Зависящие от распределений нижние границы на КЛ-дивергенцию и некоторые классы f-дивергенций через χ^2-дивергенцию в леммах 2 и 5 соответственно (эти границы используются для доказательства дальнейших линейных верхних границ на коэффициенты сжатия).
4. Линейные верхние границы на коэффициенты сжатия совместных распределений для КЛ-дивергенции и некоторые классы f-дивергенций через коэффициент сжатия для χ^2-дивергенции в теоремах 3, 4 и следствии 1.
5. Альтернативное доказательство известной эквивалентности между коэффициентами сжатия для КЛ- и χ^2-дивергенций двумерных гауссовских распределений в § 5, которое также устанавливает эквивалентность при дополнительном ограничении на среднюю мощность (или второй момент) – см. теорему 5.
6. Эквивалентные характеризации предпорядка меньшего искажения на стохастических матрицах через нелинейные операторно выпуклые f-дивергенции в теореме 6. Эта теорема обобщает известный из литературы важный результат о том, что коэффициенты сжатия для всех нелинейных операторно выпуклых f-дивергенций для заданной стохастической матрицы равны (см. предложение 6).
7. Применения наших основных результатов, такие как обобщение сильного неравенства Самородницкого об обработке данных на нелинейную операторно выпуклую f-информацию в теореме 7, а также простое доказательство в предложении 7 того факта, что скорость сходимости к стационарному состоянию для КЛ-дивергенции определяется коэффициентом сжатия для χ^2-дивергенции для эргодической обратимой цепи Маркова.

Структура статьи. Кратко обрисуем дальнейшее изложение. Вначале в § 2 приведен обзор все разрастающейся литературы по коэффициентам сжатия. В этом параграфе собраны формальные определения и важнейшие свойства обоих вышеупомянутых вариантов коэффициентов сжатия, а также вкратце описано их появление

в изучении эргодичности. Затем в § 3 сформулированы и пояснены наши основные результаты и рассмотрена относящаяся к этому литература. В § 4 представлены некоторые полезные границы, связывающие f-дивергенции и χ^2-дивергенцию, которые будут использованы для вывода линейных верхних границ на коэффициенты сжатия совместных распределений для некоторого класса f-дивергенций и КЛ-дивергенции. В продолжение этого сюжета в § 5 доказана эквивалентность некоторых коэффициентов сжатия двумерных гауссовских распределений. В § 6 доказывается эквивалентность характеризаций предпорядка меньшего искажения на стохастических матрицах через нелинейные операторно выпуклые f-дивергенции, а затем выводится обобщение сильного неравенства Самородницкого об обработке данных. Наконец, в § 7 даются заключительные замечания и намечаются направления дальнейших исследований.

§ 2. Обзор результатов по коэффициентам сжатия

Здесь мы определим коэффициенты сжатия для f-дивергенций и изложим некоторые известные факты о них. Вначале в п. 2.1 вводятся предварительные определения и обозначения, относящиеся к f-дивергенциям, а в последующих пунктах вкратце изложим первые необходимые сведения о коэффициентах сжатия и сильных неравенствах об обработке данных.

2.1. *f*-дивергенция. Рассмотрим дискретное пространство элементарных событий $\mathcal{X} \triangleq \{1, \dots, |\mathcal{X}|\}$ мощности $2 \leqslant |\mathcal{X}| < +\infty$, элементы которого будем без ограничения общности считать натуральными числами. Через $\mathcal{P}_{\mathcal{X}} \subseteq \left(\mathbb{R}^{|\mathcal{X}|}\right)^*$ обозначим вероятностный симплекс в $\left(\mathbb{R}^{|\mathcal{X}|}\right)^*$, состоящий из всех вероятностных мер на $\mathcal{X}$, где $\left(\mathbb{R}^{|\mathcal{X}|}\right)^*$ – двойственное к $\mathbb{R}^{|\mathcal{X}|}$ векторное пространство, состоящее из всех векторов-строк длины $|\mathcal{X}|$. Мы понимаем $\mathcal{P}_{\mathcal{X}}$ как множество всевозможных распределений случайной величины X со значениями $\mathcal{X}$, и будем представлять каждую вероятностную меру $P_X \in \mathcal{P}_{\mathcal{X}}$ в виде вектора-строки

$$P_X = (P_X(1), \dots, P_X(|\mathcal{X}|)) \in \left(\mathbb{R}^{|\mathcal{X}|}\right)^*.$$

Через

$$\mathcal{P}_{\mathcal{X}}^{\circ} \triangleq \{P_X \in \mathcal{P}_{\mathcal{X}} : \forall x \in \mathcal{X},\ P_X(x) > 0\}$$

обозначим относительную внутренность множества $\mathcal{P}_{\mathcal{X}}$. Популярным "расстоянием" между вероятностными мерами в теории информации и статистике является понятие f-дивергенции, независимо введенное в [21, 22] и в [23]. (Это понятие также независимо появлялось в [24], [25] и [26, 27].)

Определение 1 (f-дивергенция [21–23]). Для заданной выпуклой функции $f\colon (0, \infty) \to \mathbb{R}$, такой что $f(1) = 0$, *f-дивергенцией* вероятностной меры $P_X \in \mathcal{P}_{\mathcal{X}}$ относительно вероятностной меры $R_X \in \mathcal{P}_{\mathcal{X}}$ называется величина

$$D_f(R_X \,\|\, P_X) \triangleq \sum_{x \in \mathcal{X}} P_X(x) f\left(\frac{R_X(x)}{P_X(x)}\right) = \mathbf{E}_{P_X}\left[f\left(\frac{R_X(X)}{P_X(X)}\right)\right], \tag{1}$$

где через $\mathbf{E}_{P_X}[\,\cdot\,]$ обозначено математическое ожидание относительно P_X и в силу непрерывности и некоторых других соображений (подробнее см. в [28, § 3]) приняты соглашения $f(0) = \lim\limits_{t \to 0^+} f(t)$ (в том числе, возможно, равный бесконечности), $0f(0/0) = 0$ и $0f(r/0) = \lim\limits_{p \to 0^+} pf(r/p) = r \lim\limits_{p \to 0^+} pf(1/p)$ для всех $r > 0$ (этот предел также может быть равен бесконечности).

Понятие f-дивергенции обобщает несколько известных мер расхождения (дивергенции), используемых в теории информации, статистике и теории вероятностей. Приведем несколько примеров:

1. *Расстояние по вариации*: Для $f(t) = \frac{1}{2}|t-1|$ соответствующая f-дивергенция является расстоянием по вариации (total variation (TV) distance):

$$\|R_X - P_X\|_{\mathrm{TV}} \triangleq \max_{A\subseteq\mathcal{X}} |R_X(A) - P_X(A)| = \frac{1}{2}\|R_X - P_X\|_1, \tag{2}$$

где $P_X(A) = \sum\limits_{x\in A} P_X(x)$ для любой $A \subseteq \mathcal{X}$, через $\|\cdot\|_p$ обозначена ℓ^p-норма, $p \in [1,\infty]$, причем второе равенство, как и некоторые другие характеризации расстояния по вариации, доказано в [29, гл. 4].

2. *Дивергенция Кульбака – Лейблера* (*КЛ-дивергенция*) [30, § 2]: Для функции $f(t) = t\log(t)$, где $\log(\cdot)$ здесь и далее – натуральный логарифм (по основанию e), соответствующая f-дивергенция является КЛ-дивергенцией (или *относительной энтропией*):

$$D(R_X\,\|\,P_X) \triangleq \sum_{x\in\mathcal{X}} R_X(x)\log\left(\frac{R_X(x)}{P_X(x)}\right). \tag{3}$$

3. *χ^2-дивергенция* (*Неймана*) [31]: Для $f(t) = (t-1)^2$ или $f(t) = t^2 - 1$ соответствующая f-дивергенция является χ^2-дивергенцией:

$$\chi^2(R_X\,\|\,P_X) \triangleq \sum_{x\in\mathcal{X}} \frac{(R_X(x) - P_X(x))^2}{P_X(x)}. \tag{4}$$

Существуют и другие варианты χ^2-дивергенции – см., например, [32], где описаны варианты Пирсона и Вайды, а также их связь с f-дивергенциями.

4. *Дивергенция Хеллингера порядка* $\alpha \in (0,\infty)\setminus\{1\}$ [33, определение 2.10]: Для $f(t) = \frac{t^\alpha - 1}{\alpha - 1}$ соответствующая f-дивергенция является дивергенцией Хеллингера (или *дивергенцией Цаллиса*) порядка α:

$$\mathcal{H}_\alpha(R_X\,\|\,P_X) \triangleq \frac{1}{\alpha-1}\left(\sum_{x\in\mathcal{X}} R_X(x)^\alpha P_X(x)^{1-\alpha} - 1\right), \tag{5}$$

где $\frac{1}{2}\mathcal{H}_{\frac{1}{2}}(R_X\,\|\,P_X)$ – *квадрат расстояния Хеллингера*, $\mathcal{H}_2(R_X\,\|\,P_X) = \chi^2(R_X\,\|\,P_X)$ – это χ^2-дивергенция, а случай $\alpha = 1$ в смысле аналитического продолжения соответствует КЛ-дивергенции $\mathcal{H}_1(R_X\,\|\,P_X) = D(R_X\,\|\,P_X)$ (см. [34, § II]).

5. *Дивергенция Винце – Ле Кама порядка* $\lambda \in (0,1)$ [35–37]: для функции $f(t) = \lambda(1-\lambda)\frac{(t-1)^2}{\lambda t + (1-\lambda)}$ соответствующая f-дивергенция – это дивергенция Винце – Ле Кама порядка λ:

$$\mathrm{LC}_\lambda(R_X\,\|\,P_X) \triangleq \lambda(1-\lambda)\sum_{x\in\mathcal{X}} \frac{(R_X(x) - P_X(x))^2}{\lambda R_X(x) + (1-\lambda)P_X(x)} =$$
$$= \left(\frac{\lambda}{1-\lambda}\right)\chi^2(R_X\,\|\,\lambda R_X + (1-\lambda)P_X), \tag{6}$$

где частный случай при $\lambda = \frac{1}{2}$ известен как *расстояние Винце – Ле Кама* (или *triangular discrimination*).

Хотя в общем случае f-дивергенции не являются метриками[3], они обладают несколькими полезными свойствами. Для изложения некоторых из эти свойств обозначим через $\mathcal{Y} \triangleq \{1, \ldots, |\mathcal{Y}|\}$ другой дискретный алфавит, где $2 \leqslant |\mathcal{Y}| < +\infty$, и введем соответствующий вероятностный симплекс $\mathcal{P}_{\mathcal{Y}}$ возможных вероятностных мер для случайной величины Y со значениями в $\mathcal{Y}$. Кроме того, пусть $\mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ – множество всех $|\mathcal{X}| \times |\mathcal{Y}|$ стохастических по строкам матриц в пространстве $\mathbb{R}^{|\mathcal{X}|\times|\mathcal{Y}|}$. Всюду далее мы отождествляем условное распределение Y при заданном X (или "дискретный канал") $P_{Y|X}$ со стохастической по строкам матрицей вероятностей перехода $W \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ (т.е. $P_{Y|X} = W$), где x-й строкой матрицы W для любого $x \in \mathcal{X}$ является условная вероятностная мера $P_{Y|X=x} \in \mathcal{P}_{\mathcal{Y}}$. Мы интерпретируем $W\colon \mathcal{P}_{\mathcal{X}} \to \mathcal{P}_{\mathcal{Y}}$ как отображение, переводящее вероятностные меры $P_X \in \mathcal{P}_{\mathcal{X}}$ на входе в вероятностные меры $P_Y = P_X W \in \mathcal{P}_{\mathcal{Y}}$ на выходе путем умножения слева. Ниже перечислены некоторые хорошо известные свойства f-дивергенций (см. [21, 22]):

1. *Неотрицательность и рефлексивность*: Для любых $R_X, P_X \in \mathcal{P}_{\mathcal{X}}$ справедливо $D_f(R_X \,\|\, P_X) \geqslant 0$ (согласно неравенству Йенсена), причем при $R_X = P_X$ имеет место равенство. Кроме того, если f строго выпукла в единице, т.е. $\lambda f(x) + (1-\lambda)f(y) > f(1)$ для всех $x, y \in (0,\infty)$ и $\lambda \in (0,1)$, таких что $\lambda x + (1-\lambda)y = 1$, то равенство имеет место тогда и только тогда, когда $R_X = P_X$. (Заметим, что во всех приведенных выше примерах f-дивергенции функция f строго выпукла в единице.)
2. *Аффинная инвариантность*: Рассмотрим любую аффинную функцию $\alpha(t) = a(t-1)$, где $a \in \mathbb{R}$. Ясно, что $D_\alpha(R_X \,\|\, P_X) = 0$ для любых $R_X, P_X \in \mathcal{P}_{\mathcal{X}}$. Следовательно, f и $f + \alpha$ задают одну и ту же f-дивергенцию, т.е. $D_{f+\alpha}(R_X \,\|\, P_X) = D_f(R_X \,\|\, P_X)$ для любых $R_X, P_X \in \mathcal{P}_{\mathcal{X}}$ (где $f + \alpha$ определяется поточечным сложением).
3. *Двойственность Чисара*: Для функции f рассмотрим ее *сопряженную по Чисару* функцию

   $$f^*\colon (0,\infty) \to \mathbb{R}, \quad f^*(t) = tf\Big(\frac{1}{t}\Big),$$

   которая также выпукла, строго выпукла в единице тогда и только тогда, когда f строго выпукла в единице, и удовлетворяет равенству $f^{**} = f$. Тогда $D_{f^*}(P_X \,\|\, R_X) = D_f(R_X \,\|\, P_X)$ для любых $R_X, P_X \in \mathcal{P}_{\mathcal{X}}$.
4. *Совместная выпуклость*: Отображение $(R_X, P_X) \mapsto D_f(R_X \,\|\, P_X)$ выпукло по паре вероятностных мер на входе.
5. *Неравенство об обработке данных* (см. [23, 24]): Для любой $W \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ и любых $R_X, P_X \in \mathcal{P}_{\mathcal{X}}$ справедливо (в силу выпуклости соответствующих функций)

   $$D_f(R_X W \,\|\, P_X W) \leqslant D_f(R_X \,\|\, P_X), \tag{7}$$

   где равенство имеет место, если Y является достаточной статистикой для X, позволяющей делать выводы о паре (R_X, P_X) (см. [28, определение 5]). Кроме того, если f строго выпукла и $D_f(R_X \,\|\, P_X) < \infty$, то равенство имеет место тогда и только тогда, когда Y – достаточная статистика для X, позволяющая делать выводы о паре (R_X, P_X) (см., например, [28, теорема 14; 38, п. 3.1]).

Хотя в [22] и [39, § 2] представлено оригинальное изложение этих свойств, большей дидактикой обладает изложение в [38, § 6]. Заметим, что в силу свойства 2 мы рассматриваем f-дивергенции только для нелинейных функций f.

Теперь определим понятие "информации" между случайными величинами, соответствующее произвольной f-дивергенции, для которой также справедливо нера-

[3] Мы зачастую выделяем f-дивергенции, являющиеся метриками, называя их "расстояниями" (например, расстояние по вариации, расстояние Хеллингера) и оставляя термин "дивергенция" для тех f-дивергенций, которые метриками не являются (например, КЛ-дивергенция, χ^2-дивергенция).

венство об обработке данных. Для случайных величин X и Y с совместной вероятностной мерой $P_{X,Y}$, состоящей из (P_X, W), *взаимная f-информация* между X и Y определяется [19] (см. также [8, формула (V.8); 40, формула (11)]) как

$$I_f(X;Y) \triangleq D_f(P_{X,Y} \,\|\, P_X P_Y) = \sum_{x\in\mathcal{X}} P_X(x) D_f(P_{Y|X=x} \,\|\, P_Y), \tag{8}$$

где через $P_X P_Y$ обозначено произведение распределений, определяемых маргинальными вероятностными мерами P_X и P_Y, и при этом используется соглашение, что $P_X(x)D_f(P_{Y|X=x} \,\|\, P_Y) = 0$, если $P_X(x) = 0$. Для $f(t) = t\log(t)$ взаимная f-информация соответствует стандартной *взаимной информации* (в смысле определения Фано, см. [41, п. 2.3; 38, п. 2.3]):

$$I(X;Y) \triangleq D(P_{X,Y} \,\|\, P_X P_Y).$$

Более того, взаимная f-информация обладает некоторыми естественными свойствами информационных мер. Например, если X и Y независимы, то $I_f(X;Y) = 0$, и верно обратное, если f строго выпукла в единице.

Пусть теперь U – еще одна случайная величина с дискретным алфавитом $\mathcal{U} \triangleq \{1,\dots,|\mathcal{U}|\}$, такая что $2 \leqslant |\mathcal{U}| < +\infty$. Если (U,X,Y) имеют совместное распределение и образуют цепь Маркова $U \to X \to Y$, т.е. U условно независима от Y при заданном X, то они удовлетворяют неравенству об обработке данных [39]

$$I_f(U;Y) \leqslant I_f(U;X), \tag{9}$$

где равенство имеет место, если Y – достаточная статистика для X, позволяющая делать выводы об U (т.е. $U \to Y \to X$ также образуют цепь Маркова). Более того, если f строго выпукла и $I_f(U;X) < \infty$, то равенство имеет место тогда и только тогда, когда Y – достаточная статистика для X, позволяющая делать выводы об U. Заметим, что хотя Чисар в [39] изучал несколько другое понятие, известное как *f-информативность*, соотношение (9) можно извлечь из доказательства предложения 2.1 в [39].

Не стоит и говорить, что неравенства об обработке данных (7) и (9) являются обобщениями более известных неравенств об обработке данных для КЛ-дивергенции и взаимной информации (см., например, [30, теорема 4.1; 41]). Наконец, заметим, что хотя по поводу неравенств об обработке данных (7) и (9) мы ссылаемся на работы [21–24] и [39], соответственно, оба этих неравенства были независимо доказаны в [26,27]. В частности, в [27] изучались *обобщенные информационные функционалы*, и частный случай из [27, теорема 5.1] дает $D_f(P_U P_Y \,\|\, P_{U,Y}) \leqslant D_f(P_U P_X \,\|\, P_{U,X})$ для любой цепи Маркова $U \to X \to Y$. По двойственности Чисара для f-дивергенций отсюда вытекает (9).

В заключение этого пункта вкратце опишем "локально квадратичное поведение" f-дивергенций. Локальные аппроксимации f-дивергенций важны с геометрической точки зрения, поскольку они преобразуют окрестности стохастических многообразий с определенными f-дивергенциями в качестве метрики в пространства со скалярным произведением, где метрика задается информацией Фишера – Рао [42–44]. Рассмотрим некоторую выделенную вероятностную меру $P_X \in \mathcal{P}_\mathcal{X}^\circ$ (задающую "центр локальной окрестности" рассматриваемых вероятностных мер) и любую другую вероятностную меру $R_X \in \mathcal{P}_\mathcal{X}$. Определим вектор *сферического возмущения* меры R_X относительно P_X:

$$K_X \triangleq (R_X - P_X)\,\mathrm{diag}\big(\sqrt{P_X}\,\big)^{-1}, \tag{10}$$

где $\sqrt{\cdot}$ означает поэлементное извлечение квадратного корня из компонент вектора, а через $\mathrm{diag}(\cdot)$ обозначена диагональная матрица с соответствующими компо-

нентами вектора на главной диагонали. С помощью вектора K_X можно построить траекторию сферически возмущенных вероятностных мер:

$$R_X^{(\varepsilon)} = P_X + \varepsilon K_X \operatorname{diag}\big(\sqrt{P_X}\,\big) = \tag{11}$$

$$= (1-\varepsilon)P_X + \varepsilon R_X, \tag{12}$$

параметризованную параметром $\varepsilon \in (0,1)$, которая соответствует выпуклым комбинациям R_X и P_X. Заметим, что K_X задает направление траектории (11), а параметр ε контролирует близость между $R_X^{(\varepsilon)}$ и P_X. Равенство (11) объясняет, почему K_X называется вектором "сферического возмущения"; вектор K_X пропорционален члену первого порядка по $\varepsilon \to 0$ возмущения между векторами $\sqrt{R_X^{(\varepsilon)}}$ и $\sqrt{P_X}$, являющимися вложениями вероятностных мер $R_X^{(\varepsilon)}$ и P_X в единичную сферу в пространстве $\big(\mathbb{R}^{|\mathcal{X}|}\big)^*$.

Теперь предположим, что функция $f\colon (0,\infty) \to \mathbb{R}$, задающая рассматриваемую f-дивергенцию, дважды дифференцируема в единице и $f''(1) > 0$. Тогда с помощью формулы Тэйлора можно показать, что эта f-дивергенция локально пропорциональна χ^2-дивергенции (см. [45, § 4], или [41] для случая КЛ-дивергенции):

$$D_f\big(R_X^{(\varepsilon)} \,\|\, P_X\big) = \frac{f''(1)}{2}\varepsilon^2 \chi^2(R_X \,\|\, P_X) + o(\varepsilon^2) = \tag{13}$$

$$= \frac{f''(1)}{2}\varepsilon^2 \,\|K_X\|_2^2 + o(\varepsilon^2), \tag{14}$$

где используется стандартная o-символика Бахмана – Ландау. Локальная аппроксимация в (14) несколько более удобна, чем в (13). Действительно, можно построить траекторию (11) с помощью сферического вектора возмущения $K_X \in \big(\mathbb{R}^{|\mathcal{X}|}\big)^*$, который удовлетворяет условию ортогональности $\sqrt{P_X}K_X^T = 0$, но не имеет вид (10). Для достаточно малого $\varepsilon \neq 0$ (зависящего от P_X и K_X), векторы $R_X^{(\varepsilon)}$, определенные в (11), на самом деле являются вероятностными мерами в $\mathcal{P}_{\mathcal{X}}$.[4] Таким образом, аппроксимация в (14) остается верной, поскольку она относится к режиму, когда $\varepsilon \to 0$.

Непосредственной проверкой можно также убедиться, что f-дивергенции, для которых $f''(1) > 0$, локально симметричны, т.е.

$$D_f\big(R_X^{(\varepsilon)} \,\|\, P_X\big) = D_f\big(P_X \,\|\, R_X^{(\varepsilon)}\big) + o(\varepsilon^2).$$

Поэтому они напоминают стандартную евклидову метрику в "окрестности" вероятностных мер вокруг заданной вероятностной меры в $\mathcal{P}_{\mathcal{X}}^{\circ}$. Отметим, что преимущество использования сферических возмущений

$$\Big\{K_X \in \big(\mathbb{R}^{|\mathcal{X}|}\big)^* :\ \sqrt{P_X}K_X^T = 0\Big\}$$

вместо аддитивных возмущений (например, $R_X - P_X$) состоит в том, что они образуют пространство со стандартным евклидовым скалярным произведением. Это позволяет видоизменить (13) с использованием ℓ^2-нормы K_X вместо взвешенной ℓ^2-нормы аддитивного возмущения $K_X \operatorname{diag}\big(\sqrt{P_X}\,\big)$. Это позволяет сделать обозначения более удобными и упростить вычисления – см. наше доказательство теоремы 2. Наконец, заметим, что идеи возмущения, подобные (11), использовались ранее в различных контекстах, некоторые примеры этого можно найти в [42, 43, 46, 47].

[4] Хотя компоненты вектора $R_X^{(\varepsilon)}$ в сумме всегда равны 1, поскольку $\sqrt{P_X}K_X^T = 0$, для бóльших (по абсолютной величине) значений ε некоторые из его компонент могут быть отрицательными.

2.2. Коэффициенты сжатия совместных распределений. Неравенства об обработке данных (7) и (9) можно максимально усилить до так называемых *сильных неравенствах об обработке данных* (СНОД), вводя в них некоторые константы, известные как *коэффициенты сжатия*. Как было отмечено выше, имеется два варианта коэффициентов сжатия: первый зависит от пары, состоящей из вектора вероятностей и стохастической матрицы, т.е. совместного распределения, а второй – только от стохастической матрицы, т.е. условного распределения. В этом пункте мы введем коэффициенты первого типа, а коэффициенты второго типа обсудим позже.

Определение 2 (коэффициент сжатия для совместного распределения [1,6,8, 11,15]). Для любой вероятностной меры на входе $P_X \in \mathcal{P}_\mathcal{X}$ и любой стохастической матрицы $P_{Y|X} = W \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ *коэффициентом сжатия* для фиксированной f-дивергенции называется

$$\eta_f(P_X, P_{Y|X}) \triangleq \sup_{\substack{R_X \in \mathcal{P}_\mathcal{X} \\ 0<D_f(R_X\|P_X)<+\infty}} \frac{D_f(R_X W \,\|\, P_X W)}{D_f(R_X \,\|\, P_X)},$$

где супремум берется по всем вероятностным мерам R_X, удовлетворяющим ограничению $0 < D_f(R_X \,\|\, P_X) < +\infty$. При этом, если X или Y постоянна почти наверное (п.н.), полагаем $\eta_f(P_X, P_{Y|X}) = 0$.

Используя определение 2, из неравенства об обработке данных для f-дивергенций (7) можно вывести следующее СНОД:

$$D_f(R_X W \,\|\, P_X W) \leqslant \eta_f(P_X, P_{Y|X}) D_f(R_X \,\|\, P_X), \tag{15}$$

справедливое для всех $R_X \in \mathcal{P}_\mathcal{X}$ при фиксированных $P_X \in \mathcal{P}_\mathcal{X}$ и $W \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$. Следующее предложение показывает, что неравенство об обработке данных для взаимной f-информации можно улучшить таким же образом.

Предложение 1 (коэффициент сжатия для взаимной f-информации [8, теорема V.2]). *Для любой вероятностной меры на входе $P_X \in \mathcal{P}_\mathcal{X}$, любой стохастической матрицы $P_{Y|X} \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ и любой выпуклой дифференцируемой функции $f\colon (0,\infty) \to \mathbb{R}$ с равномерно ограниченной производной в некоторой окрестности единицы, такой что $f(1) = 0$, имеет место равенство*

$$\eta_f(P_X, P_{Y|X}) = \sup_{\substack{P_{U|X}\colon U \to X \to Y \\ 0<I_f(U;X)<+\infty}} \frac{I_f(U;Y)}{I_f(U;X)},$$

где супремум берется по всем стохастическим матрицам $P_{U|X} \in \mathcal{P}_{\mathcal{U}|\mathcal{X}}$ и конечным алфавитам $\mathcal{U}$ случайной величины U, таким что $U \to X \to Y$ образуют цепь Маркова. (Заметим, что в этой экстремальной задаче достаточно взять $|\mathcal{U}| = 2$.)

Предложение 1 доказано в [8, теорема V.2]. Частный случай этого результата для КЛ-дивергенции был доказан в [4] (для случая конечного алфавита) и в [6] (для произвольного алфавита). Интуитивно вариационная задача в предложении 1 определяет вероятностную модель, которая делает Y как можно более близкой к достаточной статистике для X относительно U (см. комментарий после формулы (9)). Кроме того, этот результат показывает, что при условиях регулярности коэффициент сжатия для любой f-дивергенции изящным образом объединяет неравенства об обработке данных для f-дивергенции и соответствующей взаимной f-информации, будучи наилучшим множителем, который можно добавить в каждое из них. Действительно, когда случайные величины $U \to X \to Y$ образуют цепь Маркова, можно

записать СНОД-версию неравенства (9):

$$I_f(U;Y) \leqslant \eta_f(P_X, P_{Y|X}) I_f(U;X), \tag{16}$$

справедливую для любого условного распределения $P_{U|X} \in \mathcal{P}_{\mathcal{U}|\mathcal{X}}$ при фиксированных $P_X \in \mathcal{P}_{\mathcal{X}}$ и $P_{Y|X} \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$. Заметим, что даже если условия предложения 1 не выполнены, неравенство (16) по-прежнему справедливо (хотя $\eta_f(P_X, P_{Y|X})$ может уже не быть наилучшей возможной мультипликативной константой в (9)).

Два коэффициента сжатия будут особенно важны для нас. Первый – это коэффициент сжатия для КЛ-дивергенции:

$$\eta_{\mathrm{KL}}(P_X, P_{Y|X}) = \sup_{\substack{R_X \in \mathcal{P}_{\mathcal{X}} \\ 0<D(R_X\|P_X)<+\infty}} \frac{D(R_X W \,\|\, P_X W)}{D(R_X \,\|\, P_X)}. \tag{17}$$

Эта величина связана с фундаментальным понятием *гиперсжимаемости* в теории вероятностей и статистике [15]. Гиперсжимаемостью называется явление, когда некоторые операторы условного математического ожидания являются сжимаемыми, даже если функциональное пространство входов имеет (вероятностную) $\mathcal{L}^q$-норму, в то время как функциональное пространство выходов имеет (вероятностную) $\mathcal{L}^p$-норму, где $1 \leqslant q < p$ (см., например, [5]). Это понятие находит применения в теории информации, поскольку гиперсжимаемые величины часто обладают свойствами тензоризации, что позволяет получать однобуквенные характеризации для них. В [5, 15] показано, что величину $\eta_{\mathrm{KL}}(P_X, P_{Y|X})$ можно определить как наклон секущей нижней границы области гиперсжимаемости (соответствующей совместной вероятностной мере $P_{X,Y}$) на бесконечности.

Коэффициент сжатия для КЛ-дивергенции объясняет поразительную дихотомию между экстремальными задачами в определении 2 и предложении 1. Чтобы пояснить этот контраст, вначале рассмотрим частный случай предложения 1 для КЛ-дивергенции и стандартной взаимной информации [4, 6]:

$$\eta_{\mathrm{KL}}(P_X, P_{Y|X}) = \sup_{\substack{P_U, P_{X|U}: U \to X \to Y \\ I(U;X)>0}} \frac{I(U;Y)}{I(U;X)}, \tag{18}$$

где оптимизация проводится (что эквивалентно) по всем $P_U \in \mathcal{P}_{\mathcal{U}}$, таким что $\mathcal{U} = \{0,1\}$ (без ограничения общности, см. [6, Приложение B]), и всем $P_{X|U} \in \mathcal{P}_{\mathcal{X}|\mathcal{U}}$, таким что для маргинальных распределений $P_X = P_U P_{X|U}$. Теперь напомним пример из работы [4], где $\mathcal{U} = \{0,1\}$, $X \sim \mathrm{Bernoulli}(1/2)$ (т.е. $P_X = (1/2, 1/2)$), а $P_{Y|X}$ – матрица "асимметричного канала со стиранием". В этом численном примере супремум в (18) достигается на последовательностях вероятностных мер $\{P^{(k)}_{X|U=0} \in \mathcal{P}_{\mathcal{X}} : k \in \mathbb{N}\}$, $\{P^{(k)}_{X|U=1} \in \mathcal{P}_{\mathcal{X}} : k \in \mathbb{N}\}$ и $\{P^{(k)}_{U} \in \mathcal{P}_{\mathcal{U}} : k \in \mathbb{N}\}$, где $\mathbb{N} \triangleq \{0,1,2,\ldots\}$, удовлетворяющим условиям

$$\lim_{k\to\infty} P^{(k)}_U(1) = 0, \tag{19}$$

$$\lim_{k\to\infty} D\big(P^{(k)}_{X|U=0} \,\|\, P_X\big) = 0 < \liminf_{k\to\infty} D\big(P^{(k)}_{X|U=1} \,\|\, P_X\big), \tag{20}$$

$$\limsup_{k\to\infty} \frac{D\big(P^{(k)}_{Y|U=0} \,\|\, P_Y\big)}{D\big(P^{(k)}_{X|U=0} \,\|\, P_X\big)} < \eta_{\mathrm{KL}}(P_X, P_{Y|X}) = \lim_{k\to\infty} \frac{D\big(P^{(k)}_{Y|U=1} \,\|\, P_Y\big)}{D\big(P^{(k)}_{X|U=1} \,\|\, P_X\big)}, \tag{21}$$

где $P_Y = P_X P_{Y|X}$ и $P^{(k)}_{Y|U=u} = P^{(k)}_{X|U=u} P_{Y|X}$ для $u \in \mathcal{U}$. Этот пример показывает, что в общем случае, хотя максимум в (18) достигается [2] при $I(U;X) \to 0$, супремум

в (17) часто достигается на последовательности вероятностных мер

$$\big\{R_X^{(k)} \in \mathcal{P}_\mathcal{X} \setminus \{P_X\} : \, k \in \mathbb{N}\big\},$$

которая не стремится к P_X (из-за невыпуклости этой экстремальной задачи). На первый взгляд, это противоречит интуиции, поскольку неравенство об обработке данных (7) обращается в равенство при $R_X = P_X$. Однако в теореме 2 (приведенной в п. 3.1) будет показано, что максимизация отношения КЛ-дивергенций с ограничением $D(R_X \,\|\, P_X) \to 0$ на самом деле позволяет достичь $\eta_{\chi^2}(P_X, P_{Y|X})$, что зачастую строго меньше [4], чем $\eta_{\mathrm{KL}}(P_X, P_{Y|X})$. Поэтому имеется резкий контраст между поведением оптимизационных задач в (17) и (18).

Второй важный коэффициент сжатия – это коэффициент сжатия для χ^2-дивергенции:

$$\eta_{\chi^2}(P_X, P_{Y|X}) = \sup_{\substack{R_X \in \mathcal{P}_\mathcal{X} \\ 0<\chi^2(R_X\|P_X)<+\infty}} \frac{\chi^2(R_X W \,\|\, P_X W)}{\chi^2(R_X \,\|\, P_X)}, \tag{22}$$

тесно связанный с обобщением коэффициента корреляции Пирсона между X и Y, известным как *максимальная корреляция Хиршфельда – Гебелейна – Реньи*, или просто максимальная корреляция [9–12]. Дадим определение максимальной корреляции, которая является мерой статистической зависимости, удовлетворяющей семи естественным аксиомам (некоторые из них будут приведены ниже в предложении 3), которым должны удовлетворять такие меры [12].

О п р е д е л е н и е 3 (максимальная корреляция [9–12]). Для двух совместно распределенных случайных величин $X \in \mathcal{X}$ и $Y \in \mathcal{Y}$ максимальной корреляцией между X и Y называется

$$\rho(X;Y) \triangleq \sup_{\substack{f:\,\mathcal{X}\to\mathbb{R},\, g:\,\mathcal{Y}\to\mathbb{R} \\ \mathbf{E}[f(X)]=\mathbf{E}[g(Y)]=0 \\ \mathbf{E}[f(X)^2]=\mathbf{E}[g(Y)^2]=1}} \mathbf{E}[f(X)g(Y)],$$

где супремум берется по всем (измеримым по Борелю) функциям f и g с нулевым средним и единичной дисперсией. При этом, если X или Y постоянна п.н., то не существует функций f и g, удовлетворяющих этим условиям, и по определению полагаем $\rho(X;Y) = 0$.

Можно показать, что коэффициент сжатия для χ^2-дивергенций в точности равен квадрату максимальной корреляции [11]:

$$\eta_{\chi^2}(P_X, P_{Y|X}) = \rho(X;Y)^2. \tag{23}$$

Кроме того, следующее предложение показывает, что максимальную корреляцию можно описать как некоторое сингулярное число; впервые это было показано в [9,12] в несколько других видах (см. также [1,4,13,40] и [48, теорема 3.2.4]).

П р е д л о ж е н и е 2 (максимальная корреляция как сингулярное число [9, 12]). *Для заданных случайных величин $X \in \mathcal{X}$ и $Y \in \mathcal{Y}$ с совместной вероятностной мерой $P_{X,Y}$, состоящей из (P_X, W), можно определить* матрицу дивергенций переходов (*МДП*)

$$B \triangleq \operatorname{diag}\big(\sqrt{P_X}\,\big) W \operatorname{diag}\big(\sqrt{P_Y}\,\big)^\dagger, \tag{24}$$

где через $\dagger$ обозначено псевдообращение Мура – Пенроуза. Тогда максимальная корреляция $\rho(X;Y)$ равна второму по величине сингулярному числу матрицы B.

Для полноты изложения доказательство этого предложения приведено в Приложении А. Из предложения 2 и равенства (23) видно, что коэффициент сжатия для χ^2-дивергенции на самом деле равен квадрату второго собственного числа МДП B. Используя принцип минимакса Куранта – Фишера – Вейля (см. [49, теоремы 4.2.6 и 7.3.8]), это можно представить в виде

$$\eta_{\chi^2}(P_X, P_{Y\,|\,X}) = \max_{\substack{x\in\mathbb{R}^{|\mathcal{X}|}\setminus\{\mathbf{0}\} \\ \sqrt{P_X}x=0}} \frac{\left\|B^T x\right\|_2^2}{\|x\|_2^2}, \tag{25}$$

где через $\mathbf{0}$ обозначен нулевой вектор соответствующей размерности, а $\sqrt{P_X}^T$ – правый сингулярный вектор матрицы B^T, соответствующий ее максимальному сингулярному числу, равному единице (см. Приложение А).

Разложения по сингулярным числам для МДП и их связь с χ^2-дивергенцией хорошо изучены в статистике. Например, в области *анализа соответствий*, основанной Хиршфельдом в 1935 г. [9], рассматриваются зависимости между категорными случайными величинами. В частности, в *простом* анализе соответствий двумерные вероятностные меры $P_{X,Y}$ рассматриваются как факторная таблица, и зависимость между X и Y раскладывается по так называемым *главным компонентам инерции* с помощью разложения по сингулярным числам матрицы B, см. работы [50; 51, § 2] и библиографию в них. В [9] с помощью этого наблюдения было получено модальное разложение взаимной χ^2-информации (или среднеквадратичной сопряженности Пирсона $\chi^2(P_{X,Y} \,\|\, P_X P_Y)$). Хотя в прошлом анализ соответствий использовался просто как техника визуализации данных, теперь он является частью более широкого инструментария геометрического анализа данных. Недавно в [40] в контексте теории информации и теории оценивания были изучены главные компоненты инерции, являющиеся собственными значениями матрицы Грама $B^T B$. Они обобщают первые главные компоненты инерции (т.е. квадраты максимальной корреляции) до величины, известной как k-корреляция, $k \in \{1, \ldots, \min\{|\mathcal{X}|, |\mathcal{Y}|\} - 1\}$, которая равна $(k+1)$-норме Ки Фаня матрицы $B^T B$ минус 1. Там же доказаны некоторые свойства k-корреляции, такие как выпуклость и неравенство об обработке данных [40, § II], и представлены некоторые приложения. В альтернативном направлении в [52] исследованы нейронные сети, позволяющие приближенно выполнить анализ соответствий в больших масштабах.

В то время как в анализе соответствий рассматриваются категорные случайные величины, зависимость между общими (некатегорными) случайными величинами изучается в близкой области исследований – анализе и идентификации так называемых *распределений Ланкастера* [53, 54]. Для заданного совместного распределения $\mathrm{P}_{X,Y}$ на произведении измеримых пространств $\mathcal{X}\times\mathcal{Y}$ пусть P_X и P_Y – маргинальные распределения на $\mathcal{X}$ и $\mathcal{Y}$ соответственно, $\mathrm{P}_X\mathrm{P}_Y$ – произведение этих распределений, а $\mathcal{L}^2(\mathcal{X}, \mathrm{P}_X)$ (соответственно, $\mathcal{L}^2(\mathcal{Y}, \mathrm{P}_Y)$) – гильбертово пространство интегрируемых с квадратом вещественнозначных функций на $\mathcal{X}$ (соответственно, на $\mathcal{Y}$) со скалярным произведением, заданным распределением P_X (соответственно, P_Y). Предположим, что $\chi^2(\mathrm{P}_{X,Y} \,\|\, \mathrm{P}_X\mathrm{P}_Y) < \infty$, откуда следует, что $\mathrm{P}_{X,Y}$ абсолютно непрерывно относительно $\mathrm{P}_X\mathrm{P}_Y$, и пусть $\mathrm{d}\mathrm{P}_{X,Y}/\mathrm{d}\mathrm{P}_X\mathrm{P}_Y$ – производная Радона – Никодима от $\mathrm{P}_{X,Y}$ относительно $\mathrm{P}_X\mathrm{P}_Y$. Для такой постановки в [53] доказано, что существуют ортонормированные базисы $\{f_j \in \mathcal{L}^2(\mathcal{X}, \mathrm{P}_X) :\ 0 \leqslant j < |\mathcal{X}|\}$ и $\{g_k \in \mathcal{L}^2(\mathcal{Y}, \mathrm{P}_Y) :\ 0 \leqslant k < |\mathcal{Y}|\}$ и некоторая последовательность $\{\sigma_k \geqslant 0 :\ 0 \leqslant k < \min\{|\mathcal{X}|, |\mathcal{Y}|\}\}$ неотрицательных корреляций, такие что $\mathrm{P}_{X,Y}$ является распределением Ланкастера, для которого справедливо разложение

$$\frac{\mathrm{d}\mathrm{P}_{X,Y}}{\mathrm{d}\mathrm{P}_X\mathrm{P}_Y}(x,y) = \sum_{k=0}^{\min\{|\mathcal{X}|,|\mathcal{Y}|\}-1} \sigma_k f_k(x) g_k(y). \tag{26}$$

Если $\mathcal{X}$ и $\mathcal{Y}$ конечны, разложение (26) в точности повторяет структуру разложения по сингулярным числам матрицы B, соответствующей $P_{X,Y}$. Дальнейшие ссылки по этой довольно общей области можно найти в [55; 56, п. II-D].

Еще одно направление исследований состоит в изучении вычислительных аспектов разложений МДП. Известным методом вычисления разложений по сингулярным числам МДП является алгоритм *чередующихся условных математических ожиданий* (ACE) (см. оригинальный алгоритм в контексте параметрической регрессии в [57] и его вариант в контексте выделения признаков и понижения размерности в [58]). По сути, алгоритм ACE использует степенной метод, или, в более общем виде, метод ортогональных итераций (см. [59, п. 7.3.2; 60, п. 4.4.3; 61, п. 4.4]), для оценки сингулярных векторов МДП. Оказывается, что такие сингулярные векторы, соответствующие наибольшим сингулярным числам, можно выделять как "более информативные" функции вклада. Этот подход был использован в [62] для получения результатов о скрытых марковских моделях в контексте обработки изображений, а в [63] он был описан как средство *универсального выделения признаков*. В работах [61, п. 4.5; 63] можно найти дальнейшие подробности о связях между разложениями по сингулярным числам МДП и другими известными понятиями в статистике и машинном обучении, такими как *метод главных компонент* [64, 65], *анализ канонической корреляции* [66] и *отображения диффузии* [67].

После того как мы ввели необходимые нам коэффициенты сжатия, приведем теперь несколько свойств коэффициентов сжатия для f-дивергенций; многие из них хорошо известны либо легко доказываются, но некоторые, насколько нам известно, ранее в литературе не встречались.

Предложение 3 (свойства коэффициентов сжатия совместных распределений). *Коэффициент сжатия для f-дивергенции обладает следующими свойствами*:

1. (Нормировка): *Для любой совместной вероятностной меры $P_{X,Y}$, состоящей из $P_X \in \mathcal{P}_\mathcal{X}$ и $P_{Y|X} \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$, выполнено $0 \leqslant \eta_f(P_X, P_{Y|X}) \leqslant 1$;*
2. (Независимость): *Рассмотрим случайные величины X и Y с совместной вероятностной мерой $P_{X,Y}$, состоящей из $P_X \in \mathcal{P}_\mathcal{X}$ и $P_{Y|X} = W \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$. Если X и Y независимы, т.е. W имеет единичный ранг, то $\eta_f(P_X, P_{Y|X}) = 0$. Наоборот, если f строго выпукла в единице и $\eta_f(P_X, P_{Y|X}) = 0$, то X и Y независимы*;
3. (Разложимость): *Рассмотрим произвольную совместную вероятностную меру $P_{X,Y}$, маргинальные вероятностные меры которой таковы, что $P_X \in \mathcal{P}_\mathcal{X}^\circ$ и $P_Y \in \mathcal{P}_\mathcal{Y}^\circ$, и пусть f дважды дифференцируема в единице, причем $f''(1) > 0$. Если мера $P_{X,Y}$ разложима, то $\eta_f(P_X, P_{Y|X}) = 1$. Наоборот, если f также строго выпукла, $f(0) < \infty$ и $\eta_f(P_X, P_{Y|X}) = 1$, то $P_{X,Y}$ разложима. (Это свойство описано в теореме 1 в п. 3.1, где также определено понятие "разложимости".)*
4. (Выпуклость [8, предложение III.3]): *Для любой фиксированной $P_X \in \mathcal{P}_\mathcal{X}^\circ$ функция $\mathcal{P}_{\mathcal{Y}|\mathcal{X}} \ni P_{Y|X} \mapsto \eta_f(P_X, P_{Y|X})$ выпукла по стохастической матрице $P_{Y|X}$*;
5. (Тензоризация [8, теорема III.9]): *Если f порождает субаддитивную и однородную f-энтропию (где f-энтропия любой неотрицательной случайной величины Z, такой что $\mathbf{E}[f(Z)] < \infty$, определяется как $\mathrm{Ent}_f(Z) \triangleq \mathbf{E}[f(Z)] - f(\mathbf{E}[Z])$, см. [8, § II]), и при этом*

   $$\big\{P_{X_i,Y_i} : P_{X_i} \in \mathcal{P}_{\mathcal{X}_i}^\circ \text{ и } P_{Y_i} \in \mathcal{P}_{\mathcal{Y}_i}^\circ \text{ при } i \in \{1, \ldots, n\}\big\}$$

   – независимые совместные вероятностные меры, то

   $$\eta_f\big(P_{X_1^n}, P_{Y_1^n|X_1^n}\big) = \max_{1 \leqslant i \leqslant n} \eta_f(P_{X_i}, P_{Y_i|X_i}),$$

   где $X_1^n = (X_1, \ldots, X_n)$ и $Y_1^n = (Y_1, \ldots, Y_n)$;

6. (Субмультипликативность): *Если $U \to X \to Y$ – дискретные случайные величины с конечным числом значений, образующие цепь Маркова, то*

$$\eta_f(P_U, P_{Y|U}) \leqslant \eta_f(P_U, P_{X|U})\eta_f(P_X, P_{Y|X}).$$

При этом для любой фиксированной совместной вероятностной меры $P_{X,Y}$, такой что X – не постоянная п.н., справедливо

$$\eta_f(P_X, P_{Y|X}) = \sup_{\substack{P_{U|X}\colon U\to X\to Y \\ \eta_f(P_U, P_{X|U})>0}} \frac{\eta_f(P_U, P_{Y|U})}{\eta_f(P_U, P_{X|U})},$$

где супремум берется по всевозможным конечным множествам значений $\mathcal{U}$ случайной величины U и по всем условным распределениям $P_{U|X} \in \mathcal{P}_{\mathcal{U}|\mathcal{X}}$, таким что $U \to X \to Y$ образуют цепь Маркова;

7. (Нижняя граница максимальной корреляции [1, теорема 5; 8, теорема III.3; 7, теорема 2]): *Рассмотрим произвольную совместную вероятностную меру $P_{X,Y}$, маргинальные вероятностные меры которой таковы, что $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ и $P_Y \in \mathcal{P}_{\mathcal{Y}}^{\circ}$. Если f дважды дифференцируема в единице и $f''(1) > 0$, то*

$$\eta_{\chi^2}(P_X, P_{Y|X}) = \rho(X;Y)^2 \leqslant \eta_f(P_X, P_{Y|X}).$$

(*Усиление этого свойства приведено в теореме 2 в п.* 3.1.)

В Приложении В приведены некоторые доказательства, а также соответствующие ссылки, уточняющие эти результаты. Теперь сделаем несколько замечаний. Во-первых, насколько нам известно, п. 3 предложения ранее не появлялся в литературе в такой общности; были известны лишь случаи η_{χ^2} и η_{KL} (см. [13, 15]).

Во-вторых, так как пп. 1–3 предложения 3 показывают, что коэффициенты сжатия являются нормализованными мерами статистической зависимости между случайными величинами, можно рассматривать субмультипликативность из п. 6 как мета-СНОД для коэффициентов сжатия по аналогии с (16). На самом деле, п. 6 также показывает, что коэффициентом сжатия в СНОД для η_f является сама величина η_f. Этот аспект пункта 6, хотя и довольно простой, также, насколько нам известно, никогда в явном виде не встречался в литературе в такой степени общности; лишь случай η_{χ^2} был представлен в [68, лемма 6].

В-третьих, вариант неравенства об обработке данных для η_{KL}, приведенный в [15] (см. также [5, п. II-A]) справедлив и для общих η_f. Действительно, если $U \to \to X \to Y \to V$ – дискретные случайные величины с конечными множествами значений, образующие цепь Маркова, то немедленным следствием пп. 1 и 6 предложения 3 является следующее свойство монотонности:

$$\eta_f(P_U, P_{V|U}) \leqslant \eta_f(P_X, P_{Y|X}). \tag{27}$$

В-четвертых, нижняя грань максимальной корреляции в п. 7 предложения 3 достигается. Например, пусть $f(t) = t\log(t)$, и рассмотрим две равномерные бернуллиевские случайные величины (X, Y) с $P_X = (1/2, 1/2)$ и с условным распределением $P_{Y|X}$, заданным матрицей "двоичного симметричного канала"

$$P_{Y|X} = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} 1-\alpha & \alpha \\ \alpha & 1-\alpha \end{bmatrix} \end{array}, \tag{28}$$

где $\alpha \in [0, 1]$ – вероятность ошибки в канале. В [15] доказано, что в этом случае нижняя грань максимальной корреляции выполнена с равенством:

$$\eta_{\mathrm{KL}}(P_X, P_{Y|X}) = \eta_{\chi^2}(P_X, P_{Y|X}) = (1-2\alpha)^2, \tag{29}$$

где $\eta_{\chi^2}(P_X, P_{Y|X}) = (1-2\alpha)^2$ можно легко вычислить с помощью характеризации максимальной корреляции через сингулярные числа из предложения 2. Еще один пример: пусть $P_{Y|X} = E_\beta \in \mathcal{P}_{\mathcal{X}\cup\{\mathrm{e}\}|\mathcal{X}}$ –матрица "$|\mathcal{X}|$-ичного канала со стиранием" с вероятностью стирания $\beta \in [0,1]$ и символом стирания e:

$$E_\beta = \begin{matrix} \\ \mathcal{X} \end{matrix} \begin{matrix} \\ [\ (1-\beta)I \end{matrix} \begin{matrix} \mathrm{e} \\ \beta \mathbf{1}\ ], \end{matrix} \tag{30}$$

где $I \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$ – единичная матрица, а $\mathbf{1} \in \mathbb{R}^{|\mathcal{X}|}$ – вектор-столбец из всех единиц. Непосредственной проверкой легко убедиться, что

$$D_f(R_X E_\beta \,\|\, P_X E_\beta) = (1-\beta) \times D_f(R_X \,\|\, P_X)$$

для любых $R_X, P_X \in \mathcal{P}_\mathcal{X}$. Поэтому для любой вероятностной меры на входе $P_X \in \mathcal{P}_\mathcal{X}$ и любой f-дивергенции имеем $\eta_f(P_X, P_{Y|X}) = 1-\beta$.

Наконец, отметим, что хотя п. 7 предложения 3 был независимо доказан авторами ранее в [1, теорема 5] с помощью идеи аппроксимации f-дивергенции, эта же идея использовалась в [8, теорема III.3; 7, теорема 2] для доказательства этого результата. Кроме того, эта идея вытекает из доказательства теоремы 5.4 работы [17] (приведенной далее в п. 6 предложения 5).

2.3. Коэффициенты эргодичности. Прежде чем перейти к обсуждению коэффициентов сжатия, зависящих только от стохастических матриц, вкратце опишем более общее понятие коэффициентов эргодичности. Впервые они возникли в контексте изучения эргодичности и скоростей сходимости неоднородных (по времени) цепей Маркова с конечными пространствами состояний (см. [69, § 1]). Они определяются следующим образом.

Определение 4 (коэффициент эргодичности [16, определение 4.6]). *Коэффициентом эргодичности* называется непрерывная скалярная функция $\eta\colon \mathcal{P}_{\mathcal{Y}|\mathcal{X}} \to \to [0,1]$, где $\mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ имеет фиксированную размерность (и снабжено стандартной топологией, индуцированной нормой Фробениуса). Этот коэффициент называется *собственным*, если для любого $W \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ равенство $\eta(W) = 0$ выполнено тогда и только тогда, когда $W = \mathbf{1}P_Y$ для некоторой вероятностной меры $P_Y \in \mathcal{P}_\mathcal{Y}$ (т.е. W имеет единичный ранг).

Полезным свойством собственных коэффициентов эргодичности является взаимосвязь со слабой эргодичностью. Рассмотрим последовательность стохастических по строкам матриц $\{W_k \in \mathcal{P}_{\mathcal{X}|\mathcal{X}} : k \in \mathbb{N}\}$, задающих неоднородную цепь Маркова на пространстве состояний $\mathcal{X}$. Введем обозначение для последовательного произведения этих матриц в количестве $r \geqslant 1$, начиная с номера $p \in \mathbb{N}$:

$$T_{(p,r)} \triangleq \prod_{i=0}^{r-1} W_{p+i}. \tag{31}$$

Цепь Маркова $\{W_k \in \mathcal{P}_{\mathcal{X}|\mathcal{X}} : k \in \mathbb{N}\}$ называется *слабо эргодической* (в смысле Колмогорова), если для всех $x_1, x_2, x_3 \in \mathcal{X}$ и всех $p \in \mathbb{N}$ [16, определение 4.4]

$$\lim_{r\to\infty} \left([T_{(p,r)}]_{x_1,x_3} - [T_{(p,r)}]_{x_2,x_3} \right) = 0. \tag{32}$$

Это определение формализует интуитивное представление, что для эргодической цепи Маркова строки таких произведений должны становиться равными при $r \to \infty$. (Отметим, что если предельная строка стохастической матрицы $\lim\limits_{r\to\infty} T_{(p,r)}$ существует для всех $p \in \mathbb{N}$, то цепь Маркова называется *сильно эргодической* [16, определение 4.5].) В следующем предложении утверждается, что слабую эргодичность

можно эквивалентным образом определить через собственные коэффициенты эргодичности.

Предложение 4 (слабая эргодичность [16, лемма 4.1]). *Пусть* $\eta\colon \mathcal{P}_{\mathcal{X}|\mathcal{X}} \to [0,1]$ – *собственный коэффициент эргодичности. Тогда неоднородная цепь Маркова* $\{W_k \in \mathcal{P}_{\mathcal{X}|\mathcal{X}} : k \in \mathbb{N}\}$ *слабо эргодична тогда и только тогда, когда*

$$\forall p \in \mathbb{N}, \quad \lim_{r\to\infty} \eta(T_{(p,r)}) = 0.$$

Чтобы пояснить интуитивный смысл этого результата, заметим, что для слабо эргодической цепи Маркова $T_{(p,r)}$ становится (приблизительно) матрицей единичного ранга при $r \to \infty$. Таким образом, следует ожидать, что $\lim\limits_{r\to\infty} \eta(T_{(p,r)}) = 0$, поскольку собственный коэффициент эргодичности непрерывен и равен нулю, когда его аргумент имеет единичный ранг. Формальное доказательство предложения 4 можно найти в [16, лемма 4.1]. О дальнейшем развитии подобных идей см. работы [69; 16, гл. 3 и 4; 70; 71, гл. 3] и библиографию в них.

Одним из первых и наиболее примечательных примеров собственных коэффициентов эргодичности являются коэффициенты сжатия Добрушина[5]. Для заданной стохастической по строкам матрицы $P_{Y|X} = W \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ ее *коэффициент сжатия Добрушина* определяется как константа Липшица отображения $\mathcal{P}_{\mathcal{X}} \ni P_X \mapsto P_X W$ относительно ℓ^1-нормы (или расстояния по вариации) [14]:

$$\eta_{\mathrm{TV}}(W) \triangleq \sup_{\substack{R_X, P_X \in \mathcal{P}_{\mathcal{X}} \\ R_X \neq P_X}} \frac{\|R_X W - P_X W\|_{\mathrm{TV}}}{\|R_X - P_X\|_{\mathrm{TV}}} = \tag{33}$$

$$= \max_{\substack{v \in (\mathbb{R}^{|\mathcal{X}|})^* \\ \|v\|_1 = 1,\, v\mathbf{1} = 0}} \|vW\|_1 = \tag{34}$$

$$= \max_{R_X, P_X \in \mathcal{P}_{\mathcal{X}}} \|R_X W - P_X W\|_{\mathrm{TV}} = \tag{35}$$

$$= \max_{x, x' \in \mathcal{X}} \left\|P_{Y|X=x} - P_{Y|X=x'}\right\|_{\mathrm{TV}} = \tag{36}$$

$$= 1 - \min_{x, x' \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \min\bigl\{P_{Y|X}(y\,|\,x), P_{Y|X}(y\,|\,x')\bigr\}, \tag{37}$$

где различные эквивалентные характеризации (34), (35), (36) (двухточечная характеризация Добрушина [14]) и (37) (характеризация *аффинности* [73; 29, формула (4.13)]) определения (33) можно либо найти в работе [16, гл. 4.3], либо легко ввести из ее результатов. Формула (37) показывает, что $\eta_{\mathrm{TV}}(W) < 1$ тогда и только тогда, когда W является *скремблирующей матрицей* (т.е. никакие две строки W не ортогональны) [16, с. 82]. (Из этого также следует, что $\eta_{\mathrm{TV}}(W) < 1$ тогда и только тогда, когда *пропускная способность с нулевой ошибкой* для $P_{Y|X}$ равна нулю [74].)

Вдобавок к свойствам собственных коэффициентов эргодичности величина η_{TV} обладает также следующими свойствами:

1. *Непрерывность по Липшицу* [70, теорема 3.4, замечание 3.5]: Для любых $V, W \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ справедливо неравенство $|\eta_{\mathrm{TV}}(V) - \eta_{\mathrm{TV}}(W)| \leqslant \|V - W\|_\infty$, где через $\|\cdot\|_\infty$ обозначена индуцированная ℓ^∞-норма, или максимальная сумма модулей по строке применительно к матрице;

[5] Следуя библиографической дискуссии в [16, с. 144–147], авторами коэффициента сжатия Добрушина (или, эквивалентно, коэффициента эргодичности Добрушина) можно также считать (по крайней мере, частично) Дёблина и Маркова. В литературе этот коэффициент встречался под названием *коэффициент сжатия Дёблина* и возникал в *лемме Маркова о сжатии* (см., например, [72, с. 619]).

2. *Субмультипликативность* [16, лемма 4.3]: Для любых $V \in \mathcal{P}_{\mathcal{X}|\mathcal{U}}$ и $W \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ справедливо неравенство $\eta_{\mathrm{TV}}(VW) \leqslant \eta_{\mathrm{TV}}(V)\eta_{\mathrm{TV}}(W)$;
3. *Граница субдоминантного собственного числа* [69, с. 584, формула (9)]: Для любой $W \in \mathcal{P}_{\mathcal{X}|\mathcal{X}}$ неравенство $\eta_{\mathrm{TV}}(W) \geqslant |\lambda|$ справедливо для любого субдоминантного собственного числа $\lambda \neq 1$ матрицы W.

Благодаря двум последним свойствам η_{TV} становится удобным инструментом для исследования неоднородных цепей Маркова. Как указано в [70, § 1], для однородной цепи Маркова $W \in \mathcal{P}_{\mathcal{X}|\mathcal{X}}$ со стационарной вероятностной мерой $\pi \in \mathcal{P}_{\mathcal{X}}$ хорошо известно, что *модуль второго по абсолютной величине собственного значения* матрицы W, обозначаемый через $\mu(W)$, отвечает за скорость сходимости к стационарному состоянию. Действительно, если $\mu(W) < 1$, то $\mu(W^n) = \mu(W)^n$, и $\lim\limits_{n\to\infty} W^n = \mathbf{1}\pi$ со скоростью, определяемой величиной $\mu(W)$. Однако для неоднородной цепи Маркова в общем случае $\{W_k \in \mathcal{P}_{\mathcal{X}|\mathcal{X}} : k \in \mathbb{N}\}$, $\mu(T_{(0,n)}) \neq \prod\limits_{i=0}^{n-1} \mu(W_i)$, поскольку модуль второго собственного числа не мультипликативен. Последние два свойства η_{TV} показывают, что эта величина является адекватной заменой модуля второго собственного числа при изучении неоднородных цепей Маркова, поскольку

$$\mu(T_{(0,n)}) \leqslant \eta_{\mathrm{TV}}(T_{(0,n)}) \leqslant \prod_{i=0}^{n-1} \eta_{\mathrm{TV}}(W_i).$$

2.4. Коэффициенты сжатия стохастических матриц. Коэффициенты сжатия стохастических матриц для f-дивергенций образуют широкий класс коэффициентов эргодичности. Они определяются аналогично формуле (33), но с использованием других f-дивергенций вместо расстояния по вариации.

Определение 5 (коэффициент сжатия стохастической матрицы [14–17]). Для любой стохастической матрицы $P_{Y|X} = W \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ ее *коэффициент сжатия* для заданной f-дивергенции равен

$$\eta_f(P_{Y|X}) \triangleq \sup_{\substack{R_X, P_X \in \mathcal{P}_{\mathcal{X}} \\ 0<D_f(R_X\|P_X)<+\infty}} \frac{D_f(R_X W \,\|\, P_X W)}{D_f(R_X \,\|\, P_X)} = \sup_{P_X \in \mathcal{P}_{\mathcal{X}}} \eta_f(P_X, P_{Y|X}),$$

где супремум берется по всем вероятностным мерам R_X и P_X, таким что $0 < D_f(R_X \,\|\, P_X) < +\infty$. Если при этом Y постоянна п.н., то по определению полагаем $\eta_f(P_{Y|X}) = 0$.

Из этого определения немедленно вытекают СНОД для коэффициентов сжатия стохастических матриц, аналогичные неравенствам (15) и (16). Кроме того, для коэффициентов сжатия стохастических матриц также справедлив вариант предложения 1. Действительно, из определения 5 и предложения 1, получаем, что для любой стохастической матрицы $P_{Y|X} \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ и любой выпуклой дифференцируемой функции $f\colon (0,\infty) \to \mathbb{R}$ с равномерно ограниченной производной в некоторой окрестности единицы, такой что $f(1) = 0$, справедливо

$$\eta_f(P_{Y|X}) = \sup_{\substack{P_{U,X}\colon U \to X \to Y \\ 0<I_f(U;X)<+\infty}} \frac{I_f(U;Y)}{I_f(U;X)}, \tag{38}$$

где супремум берется по всем совместным вероятностным мерам $P_{U,X}$ (состоящим из $P_U \in \mathcal{P}_{\mathcal{U}}$ и $P_{X|U} \in \mathcal{P}_{\mathcal{X}|\mathcal{U}}$) и конечным алфавитам $\mathcal{U}$ случайной величины U, такой что $U \to X \to Y$ образуют цепь Маркова. Частный случай этого результата для КЛ-дивергенции можно найти в [75, с. 345, задача 15.12] (случай конечного алфавита) и [7] (общий случай).

Имеется два важных примера коэффициентов сжатия стохастических матриц: коэффициент сжатия Добрушина для расстояния по вариации (определенный в (33)) и коэффициент сжатия для КЛ-дивергенции. Как и выше, для заданной $P_{Y|X} \in \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ обозначаем через $\eta_{\mathrm{TV}}(P_{Y|X})$, $\eta_{\mathrm{KL}}(P_{Y|X})$, и $\eta_{\chi^2}(P_{Y|X})$ коэффициент сжатия $P_{Y|X}$ для расстояния по вариации, КЛ-дивергенции и χ^2-дивергенции соответственно. В [15] доказано, что для любой стохастической матрицы $P_{Y|X} \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ верно равенство

$$\eta_{\mathrm{KL}}(P_{Y|X}) = \eta_{\chi^2}(P_{Y|X}). \tag{39}$$

Таким образом, при изучении коэффициентов сжатия стохастических матриц не требуется рассматривать η_{KL} и η_{χ^2} по отдельности. Заметим, что альтернативное доказательство равенства (39) (справедливое для общих измеримых пространств) дано в [7, теорема 3]. Кроме того, менее известное наблюдение состоит в том, что правильное обобщение на произвольную размерность техники доказательства из [76, лемма 1, теорема 1] (где аналитически вычисляется η_{KL} для любой стохастической матрицы размера 2×2) также позволяет доказать равенство (39). Следует отметить, что основным результатом работы [76] является индуктивный подход к оценке сверху для η_{KL} в байесовских сетях (или направленных ациклических графах). Суть этого подхода прекрасно изложена в [7], где также приведены доказательства его обобщения на расстояние по вариации (с помощью представления Гольдштейна расстояния по вариации между двумя совместными распределениями через одновременное максимальное склеивание [77]) и описана его связь с задачей перколяции узлов.

Теперь перечислим некоторые известные свойства коэффициентов сжатия стохастических матриц.

Предложение 5 (свойства коэффициентов сжатия стохастических матриц). *Коэффициент сжатия для f-дивергенции обладает следующими свойствами*:

1. (Нормировка): *Для любой стохастической матрицы $P_{Y|X} \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ справедливо неравенство $0 \leqslant \eta_f(P_{Y|X}) \leqslant 1$;*
2. (Независимость [17, § 4]): *Рассмотрим стохастическую матрицу $P_{Y|X} \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$. Если $P_{Y|X}$ имеет единичный ранг, т.е. X и Y независимы, то $\eta_f(P_{Y|X}) = 0$. Наоборот, если f строго выпукла в единице и $\eta_f(P_{Y|X}) = 0$, то $P_{Y|X}$ имеет единичный ранг*;
3. (Скремблирование [17, теорема 4.2]): *Для заданной стохастической матрицы $P_{Y|X} \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ матрица $P_{Y|X}$ является скремблирующей тогда и только тогда, когда $\eta_f(P_{Y|X}) < 1$*;
4. (Выпуклость [17, § 4; 8, предложение III.3]): *Функция $\mathcal{P}_{\mathcal{Y}|\mathcal{X}} \ni P_{Y|X} \mapsto \eta_f(P_{Y|X})$ выпукла*;
5. (Субмультипликативность [17, § 4]): *Если $U \to X \to Y$ – дискретные случайные величины с конечными множествами значений, образующие цепь Маркова, т.е. стохастические матрицы условных распределений таковы, что $P_{Y|U} = = P_{X|U} P_{Y|X}$, то*

$$\eta_f(P_{Y|U}) \leqslant \eta_f(P_{X|U}) \eta_f(P_{Y|X});$$

6. (Нижняя граница χ^2-дивергенции [17, теорема 5.4; 19, предложение II.6.15]): *Рассмотрим стохастическую матрицу $P_{Y|X} \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$. Если f дважды дифференцируема в единице и $f''(1) > 0$, то*

$$\eta_{\chi^2}(P_{Y|X}) \leqslant \eta_f(P_{Y|X});$$

7. (Верхняя граница расстояния по вариации [17, теорема 4.1; 19, предложение II.4.10]): *Для любой стохастической матрицы $P_{Y|X} \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ справедливо неравенство*

$$\eta_f(P_{Y|X}) \leqslant \eta_{\mathrm{TV}}(P_{Y|X}).$$

Доказательства этих результатов мы опускаем, поскольку они либо аналогичны соответствующим доказательствам в предложении 3, либо даны в указанных ссылках. Пункты 1, 2 и 4 предложения 5 означают, что коэффициенты сжатия стохастических матриц для f-дивергенций часто являются собственными коэффициентами эргодичности. (Действительно, из выпуклости отображения $P_{Y|X} \mapsto \eta_f(P_{Y|X})$ в п. 4 предложения 5 следует, что это отображение непрерывно на внутренней области $\mathcal{P}_{\mathcal{Y}|\mathcal{X}}$.) Заметим, что п. 3 показывает, что $\eta_f(P_{Y|X}) = 1$ тогда и только тогда, когда $\eta_{\mathrm{TV}}(P_{Y|X}) = 1$ [17, теорема 4.2], и непосредственной проверкой легко убедиться, что $\eta_f(P_{Y|X}) = 1$ тогда и только тогда, когда пропускная способность при нулевой ошибке для $P_{Y|X}$ строго положительна [74]. Отметим также, что результат об экстремальном значении, аналогичный п. 6 предложения 3, хотя и менее содержательный, можно вывести из п. 5 предложения 5.

В то время как соотношение (39) показывает, что в п. 6 предложения 5 легко может достигаться равенство, неравенство в п. 7 часто бывает строгим. Например, когда $P_{Y|X} = W$ – стохастическая матрица размера 2×2 с параметрами $a, b \in [0, 1]$

$$W = \begin{bmatrix} 1-a & a \\ b & 1-b \end{bmatrix}, \tag{40}$$

непосредственной проверкой легко убедиться, что $\eta_{\mathrm{KL}}(P_{Y|X}) \leqslant \eta_{\mathrm{TV}}(P_{Y|X})$, где неравенство обычно бывает строгим, поскольку

$$\eta_{\mathrm{KL}}(P_{Y|X}) = 1 - \left(\sqrt{a(1-b)} + \sqrt{b(1-a)}\right)^2, \tag{41}$$

$$\eta_{\mathrm{TV}}(P_{Y|X}) = |1-a-b|, \tag{42}$$

где равенство (41) доказано в [76, теорема 1], а (42) легко получить с помощью (36). Более того, в частном случае, где $P_{Y|X}$ имеет вид (28), получаем [15]

$$\eta_{\mathrm{KL}}(P_{Y|X}) = (1-2\alpha)^2 \leqslant |1-2\alpha| = \eta_{\mathrm{TV}}(P_{Y|X}). \tag{43}$$

С другой стороны, как показано в конце п. 2.2, $\eta_f(P_{Y|X}) = 1 - \beta$ для любой f-дивергенции, когда $P_{Y|X} = E_\beta \in \mathcal{P}_{\mathcal{X}\cup\{\mathrm{e}\}|\mathcal{X}}$.

Ввиду п. 6 и равенства (39) естественно задаться вопросом, существуют ли другие f-дивергенции, коэффициенты сжатия которых (для стохастических матриц) также равны η_{χ^2}. Следующий результат из [18, теорема 1], обобщающий соотношение (39), дает ответ на этой вопрос.

Предложение 6 (коэффициенты сжатия для операторно выпуклых f-дивергенций [18, теорема 1; 19]). *Для любой нелинейной операторно выпуклой функции $f\colon (0,\infty) \to \mathbb{R}$, такой что $f(1) = 0$, и любой стохастической матрицы $P_{Y|X} \in \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ справедливо равенство*

$$\eta_f(P_{Y|X}) = \eta_{\chi^2}(P_{Y|X}).$$

Доказательство теоремы 1 из [18] основано на элегантном интегральном представлении операторно выпуклых функций (см. п. 6.1). Такие представления являются мощным инструментом для доказательства неравенств между коэффициентами сжатия, и мы воспользуемся ими для обобщения предложения 6 в п. 3.4. На

самом деле, п. 7 предложения 5 также можно доказать с помощью интегрального представления (см. [8, теорема III.1]).

В заключение этого пункта укажем на дополнительный обзор по коэффициентам сжатия [7, § 2], где, в частности, имеются ссылки на различные приложения этих идей в существующей литературе.

§ 3. Основные результаты и их обсуждение

В настоящей статье мы в основном будем сравнивать между собой различные коэффициенты сжатия. В частности, мы будем интересоваться следующими основными вопросами:

1. *Когда коэффициенты сжатия совместных распределений достигают своей верхней границы, равной единице*? В теореме 1 (п. 3.1) мы покажем, что это происходит, когда совместные распределения разложимы.
2. *Можно ли достичь нижней границы максимальной корреляции из* п. 7 *предложения* 3 *путем наложения дополнительных ограничений в экстремальной задаче, определяющей коэффициенты сжатия совместных распределений*? Да, можно потребовать, чтобы f-дивергенция на входе была малой, как показано в теореме 2 (п. 3.1).
3. Обычно мы оцениваем $\eta_{\mathrm{KL}}(P_X, P_{Y|X})$ снизу через $\eta_{\chi^2}(P_X, P_{Y|X})$ (предложение 3, п. 7), а сверху – через $\eta_{\mathrm{TV}}(P_{Y|X})$ (предложение 5, п. 7). *Существует ли простая верхняя граница на* $\eta_{\mathrm{KL}}(P_X, P_{Y|X})$ *через* $\eta_{\chi^2}(P_X, P_{Y|X})$? Да, две такие границы приведены в следствии 1 и теореме 4 в п. 3.2.
4. *Можно ли обобщить эту верхнюю границу для КЛ-дивергенции на другие* f*-дивергенции*? Да, более общая граница представлена в теореме 3 (п. 3.2).
5. Когда X и Y совместно гауссовские, с помощью характеризации взаимной информации, данной в (18), можно установить [2, теорема 7], что

   $$\eta_{\mathrm{KL}}(P_X, P_{Y|X}) = \eta_{\chi^2}(P_X, P_{Y|X}).$$

   Существует ли простое доказательство этого результата, опирающееся непосредственно на определение η_{KL}? *Выполняется ли это равенство, если наложить дополнительное ограничение по мощности в экстремальной задаче для* η_{KL}? Да, в п. 3.3 мы рассмотрим гауссовский случай, и в теореме 5 докажем это равенство для η_{KL} с ограничением по мощности. Наше доказательство также устанавливает это известное равенство с помощью определения η_{KL} через КЛ-дивергенцию.
6. Коэффициенты сжатия стохастических матриц тесно связаны с предпорядком меньшего искажения на стохастических матрицах [7, § 6]. *Можно ли обобщить результат предложения* 6 *и получить новые сведения о предпорядке меньшего искажения*? Да, в п. 3.4 мы вводим предпорядок меньшего искажения, и в теореме 6 получаем целый класс его эквивалентных характеризаций. Мы также приводим пример применения теоремы 6 в теореме 7, обобщающей СНОД Самородницкого.

Границы, которые мы выведем в ответ на вопросы 3–5, имеют вид верхней границы в

$$\eta_{\chi^2}(P_X, P_{Y|X}) \leqslant \eta_f(P_X, P_{Y|X}) \leqslant C\eta_{\chi^2}(P_X, P_{Y|X}), \tag{44}$$

где первое неравенство – это просто нижняя граница максимальной корреляции из п. 7 предложения 3, а константа C зависит от $P_{X,Y}$ и f; заметим, что $C = 1$ в постановке вопроса 5. Будем называть такие границы *линейными границами* между коэффициентами сжатия совместных распределений. Наши основные результаты сформулированы в нескольких следующих пунктах.

3.1. Свойства коэффициентов сжатия совместных распределений. В этом пункте, как и в п. 3.2, мы будем предполагать, что заданы случайные величины $X \in \mathcal{X}$ и $Y \in \mathcal{Y}$ с совместной вероятностной мерой $P_{X,Y}$, состоящей из $P_X \in \mathcal{P}_{\mathcal{X}}$ и $P_{Y|X} = W \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$, причем маргинальные вероятностные меры таковы, что $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ и $P_Y = P_X W \in \mathcal{P}_{\mathcal{Y}}^{\circ}$. В этом пункте приведены два результата. В первом из них переформулирован п. 3 предложения 3.

Начнем с необходимого определения: совместная вероятностная мера $P_{X,Y}$ называется *разложимой*, если существуют функции $h\colon \mathcal{X} \to \mathbb{R}$ и $g\colon \mathcal{Y} \to \mathbb{R}$, такие что $h(X) = g(Y)$ п.н. и $\mathrm{Var}(h(X)) > 0$, где через $\mathrm{Var}(\cdot)$ обозначена дисперсия. Эквивалентным образом, $P_{X,Y}$ разложима тогда и только тогда, когда неориентированный двудольный граф с непересекающимися множествами вершин $\mathcal{X}$ и $\mathcal{Y}$ и множеством ребер $\{(x,y) \in \mathcal{X} \times \mathcal{Y} : P_{Y|X}(y\,|\,x) > 0\}$ имеет две или более компоненты связности (см. [15, § 1]). Эта комбинаторная характеризация разложимости имеет место, поскольку h и g существуют тогда и только тогда, когда матрица W имеет блочно-диагональную структуру после соответствующей перестановки ее строк и столбцов (где блоки определяются множествами прообразов h и g), и эти блоки соответствуют компонентам связности ассоциированного двудольного графа.

С помощью понятия разложимости можно описать, когда коэффициенты сжатия совместных распределений равны единице.

Теорема 1 (свойство разложимости). *Пусть задана дважды дифференцируемая в единице выпуклая функция $f\colon (0,\infty) \to \mathbb{R}$, такая что $f(1) = 0$ и $f''(1) > 0$. Тогда справедливо следующее*:

1. *Если $P_{X,Y}$ разложима, то $\eta_f(P_X, P_{Y|X}) = 1$;*
2. *Если f строго выпукла и удовлетворяет условию $f(0) = \lim\limits_{t\to 0^+} f(t) < \infty$, то из равенства $\eta_f(P_X, P_{Y|X}) = 1$ следует, что $P_{X,Y}$ разложима.*

Теорема 1 доказана в Приложении C. Как отмечалось выше, этот результат был ранее известен только для случаев η_{χ^2} и η_{KL} [13, 15]. Заметим также, что в случае $\eta_f(P_X, P_{Y|X}) = 1$ можно проверить, что пропускная способность при нулевой ошибке для $P_{Y|X}$ строго положительна [74].

Наш второй результат – уточнение п. 7 предложения 3, он показывает, что при стремлении f-дивергенции на входе к нулю общие коэффициенты сжатия превращаются в коэффициенты сжатия для χ^2-дивергенции.

Теорема 2 (локальная аппроксимация коэффициентов сжатия). *Пусть дана выпуклая функция $f\colon (0,\infty) \to \mathbb{R}$, строго выпуклая в единице и дважды дифференцируемая в единице, такая что $f(1) = 0$ и $f''(1) > 0$. Тогда*

$$\eta_{\chi^2}(P_X, P_{Y|X}) = \lim_{\delta\to 0^+} \sup_{\substack{R_X \in \mathcal{P}_{\mathcal{X}} \\ 0 < D_f(R_X \| P_X) \leqslant \delta}} \frac{D_f(R_X W \,\|\, P_X W)}{D_f(R_X \,\|\, P_X)}.$$

Доказательство приведено в Приложении D; заметим, что частный случай теоремы 2 для КЛ-дивергенции был представлен вместе с наброском доказательства в [1, теорема 3]. Теперь уместно сделать несколько замечаний. Во-первых, отметим, что в доказательстве п. 7 предложения 3 в Приложении B (как и в независимых доказательствах в [8, теорема III.2] и [7, теорема 2]) уже содержится идея о том, что оптимизация $\eta_f(P_X, P_{Y|X})$ по локальным возмущениям P_X дает $\eta_{\chi^2}(P_X, P_{Y|X})$ в силу (14) и (25). Однако это доказательство (с незначительными видоизменениями) показывает лишь, что величина $\eta_{\chi^2}(P_X, P_{Y|X})$ ограничена сверху правой частью равенства из теоремы 2. Хотя интуитивно может быть и ясно, что эта верхняя грани достигается с равенством, формальное доказательство содержит несколько технических деталей, как показано в Приложении D.

Во-вторых, теорема 2 очевидным образом показывает, что нижняя граница максимальной корреляции в п. 7 предложения 3 может быть достигнута, когда в задаче оптимизации для $\eta_f(P_X, P_{Y|X})$ наложено дополнительное ограничение, что f-дивергенция на входе мала. Поэтому из теоремы 2 вытекает нижняя граница максимальной корреляции. Эта идея оказалась важна при сравнении $\eta_{\chi^2}(P_X, P_{Y|X})$ и $\eta_{\mathrm{KL}}(P_X, P_{Y|X})$ в статистическом контексте [78, с. 5].

В-третьих, теорему 2 можно рассматривать как минимаксную характеризацию $\eta_{\chi^2}(P_X, P_{Y|X})$, поскольку супремум отношения f-дивергенций является невозрастающей функцией δ, и поэтому предел (при $\delta \to 0^+$) можно заменить на инфимум (по всем $\delta > 0$).

В-четвертых, при выполнении условий предложения 1 и теоремы 2 непосредственной проверкой легко убедиться, что

$$\eta_{\chi^2}(P_X, P_{Y|X}) = \lim_{\delta\to 0^+} \sup_{\substack{P_{U|X}:\, U\to X\to Y \\ 0<I_f(U;X)\leqslant\delta}} \frac{I_f(U;Y)}{I_f(U;X)}, \tag{45}$$

где супремум берется по всем стохастическим матрицам $P_{U|X} \in \mathcal{P}_{\mathcal{U}|\mathcal{X}}$, таким что $\mathcal{U} = \{0,1\}$, $U \sim \mathrm{Bernoulli}(1/2)$ и $U \to X \to Y$ образуют цепь Маркова. Таким образом, ограничение на малость f-дивергенции на входе в определении $\eta_f(P_X, P_{Y|X})$ через f-дивергенцию соответствует условиям малой $I_f(U;X)$ и $U \sim \mathrm{Bernoulli}(1/2)$ в (45).

Наконец, рассмотрим траекторию вероятностных мер на входе

$$R_X^{(\varepsilon)} = P_X + \varepsilon K_X^* \operatorname{diag}\bigl(\sqrt{P_X}\,\bigr),$$

где $\varepsilon > 0$ достаточно мало, а $K_X^* \in \bigl(\mathbb{R}^{|\mathcal{X}|}\bigr)^*$ – левый сингулярный вектор, соответствующий второму по величине сингулярному числу МДП B (см. (25)). Как показывает доказательство в Приложении D, эта траектория удовлетворяет условию $\lim\limits_{\varepsilon\to 0} D_f\bigl(R_X^{(\varepsilon)} \,\|\, P_X\bigr) = 0$ и достигает значения $\eta_{\chi^2}(P_X, P_{Y|X})$ в теореме 2:

$$\lim_{\varepsilon\to 0} \frac{D_f\bigl(R_X^{(\varepsilon)} W \,\|\, P_X W\bigr)}{D_f\bigl(R_X^{(\varepsilon)} \,\|\, P_X\bigr)} = \eta_{\chi^2}(P_X, P_{Y|X}). \tag{46}$$

Соответствующей траекторией условных распределений для (45) является $P_{X|U}^{(\varepsilon)} \in \in \mathcal{P}_{\mathcal{X}|\mathcal{U}}$ со строками

$$\Bigl\{P_{X|U=u}^{(\varepsilon)} = P_X + (2u-1)\varepsilon K_X^* \operatorname{diag}\bigl(\sqrt{P_X}\,\bigr) :\ u \in \{0,1\}\Bigr\},$$

где $\varepsilon > 0$ достаточно мало. Эта траектория удовлетворяет условию

$$\lim_{\varepsilon\to 0} I_f(U; X^{(\varepsilon)}) = 0$$

и достигает значения $\eta_{\chi^2}(P_X, P_{Y|X})$ в (45):

$$\lim_{\varepsilon\to 0} \frac{I_f(U;Y^{(\varepsilon)})}{I_f(U;X^{(\varepsilon)})} = \eta_{\chi^2}(P_X, P_{Y|X}), \tag{47}$$

где $U \sim \mathrm{Bernoulli}(1/2)$, $U \to X^{(\varepsilon)} \to Y^{(\varepsilon)}$ – случайные величины, образующие цепь Маркова с совместной вероятностной мерой $\bigl(P_U, P_{X|U}^{(\varepsilon)}, P_{Y|X}\bigr)$, $P_{Y|U}^{(\varepsilon)} = P_{X|U}^{(\varepsilon)} P_{Y|X}$, а маргинальная вероятностная мера для $X^{(\varepsilon)}$ равна P_X.

3.2. Линейные границы между коэффициентами сжатия. Напомним, что мы рассматриваем заданную совместную вероятностную меру $P_{X,Y}$ с маргинальными вероятностными мерами $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ и $P_Y \in \mathcal{P}_{\mathcal{Y}}^{\circ}$. Наш следующий результат описывает линейную верхнюю границу на $\eta_f(P_X, P_{Y|X})$ через $\eta_{\chi^2}(P_X, P_{Y|X})$ для определенного класса f-дивергенций.

Теорема 3 (граница на коэффициенты сжатия). *Пусть задана выпуклая функция $f\colon (0,\infty) \to \mathbb{R}$, трижды дифференцируемая в единице, такая что $f(1) = 0$, $f''(1) > 0$ и $f(0) < \infty$, удовлетворяющая условию (75) для любого $t \in (0,\infty)$ (см. п. 4.1). Предположим также, что разностное отношение $g\colon (0,\infty) \to \mathbb{R}$, определяемое как $g(t) = \dfrac{f(t) - f(0)}{t}$, выпукло вверх. Тогда*

$$\eta_f(P_X, P_{Y|X}) \leqslant \frac{f'(1) + f(0)}{f''(1) \min\limits_{x\in\mathcal{X}} P_X(x)} \eta_{\chi^2}(P_X, P_{Y|X}).$$

Теорема 3 доказана в п. 4.2. Условия на f гарантируют, что получаемая f-дивергенция обладает свойствами КЛ-дивергенции, требуемыми в доказательстве теоремы 4 (см. ниже). Таким образом, аналогичная техника доказательства работает и для теоремы 3. Непосредственным частным случаем для КЛ-дивергенции, впервые доказанным в [1, теорема 10], является

Следствие 1 (граница на КЛ-коэффициенты сжатия [1, теорема 10]). *Справедливо неравенство*

$$\eta_{\mathrm{KL}}(P_X, P_{Y|X}) \leqslant \frac{\eta_{\chi^2}(P_X, P_{Y|X})}{\min\limits_{x\in\mathcal{X}} P_X(x)}.$$

Следствие 1 можно вывести из теоремы 3, убедившись, что функция $f(t) = t\log(t)$ удовлетворяет условиям теоремы 3 (см. [79]). Подробное доказательство см. в Приложении Е. Константу в верхней границе на $\eta_{\mathrm{KL}}(P_X, P_{Y|X})$ из следствия 1 можно улучшить, как показывает следующая

Теорема 4 (улучшенная граница на КЛ-коэффициенты сжатия). *Справедливо неравенство*

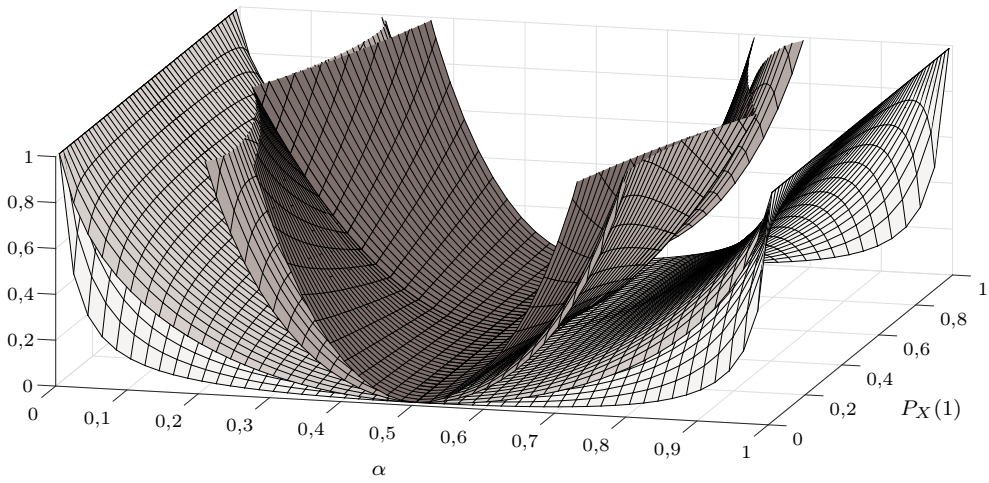
$$\eta_{\mathrm{KL}}(P_X, P_{Y|X}) \leqslant \frac{2\eta_{\chi^2}(P_X, P_{Y|X})}{\varphi\Big(\max\limits_{A\subseteq\mathcal{X}} \min\{P_X(A), 1 - P_X(A)\}\Big) \min\limits_{x\in\mathcal{X}} P_X(x)},$$

где функция $\varphi\colon [0,1/2] \to \mathbb{R}$ определена в (68) (см. п. 4.1).

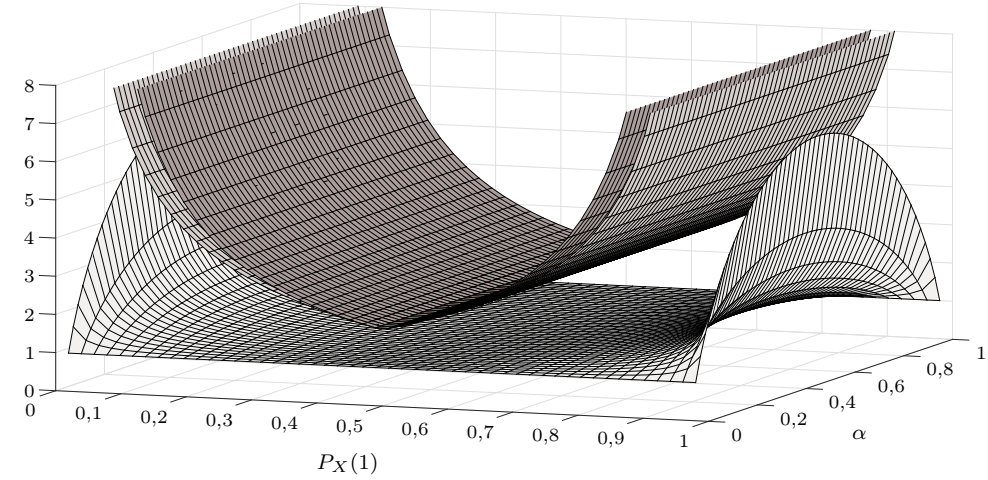
Теорема 4 также доказана в п. 4.2, и в силу соотношения (70) в п. 4.1 эта граница точнее, чем граница из следствия 1. Теперь уместно сделать несколько замечаний по поводу следствия 1 и теорем 3 и 4.

Во-первых, как показано на рис. 1а, верхние границы из этих утверждений могут быть строго меньше тривиальной верхней границы, равной единице. Например, когда (X, Y) – равномерные бернуллиевские случайные величины с $P_X = (1/2, 1/2)$ и $P_{Y|X}$, описанные в (28), для некоторого $\alpha \in [0,1]$ (что соответствует сечению вдоль $P_X(1) = 1/2$ на рис. 1а), верхние границы из следствия 1 и теоремы 4 обе равны

$$\frac{2\eta_{\chi^2}(P_X, P_{Y|X})}{\varphi\Big(\max\limits_{A\subseteq\mathcal{X}} \min\{P_X(A), 1 - P_X(A)\}\Big) \min\limits_{x\in\mathcal{X}} P_X(x)} = \frac{\eta_{\chi^2}(P_X, P_{Y|X})}{\min\limits_{x\in\mathcal{X}} P_X(x)} = 2(1-2\alpha)^2, \quad (48)$$



(a) Графики функций $\eta_{\chi^2}(P_X, P_{Y|X})$ (нижняя сетка), $\eta_{\mathrm{KL}}(P_X, P_{Y|X})$ (вторая снизу) и линейных верхних границ на $\eta_{\mathrm{KL}}(P_X, P_{Y|X})$. Верхняя сетка соответствует верхней границе из следствия 1, а вторая сверху – более точной верхней границе из теоремы 4.



(b) Графики верхних границ на отношение $\eta_{\mathrm{KL}}(P_X, P_{Y|X})/\eta_{\chi^2}(P_X, P_{Y|X})$, соответствующее нижней сетке. Границе $1/\min\limits_{x\in\mathcal{X}} P_X(x)$ из следствия 1 соответствует верхняя сетка, а более точной границе $2/\Big(\varphi\Big(\max\limits_{A\subseteq\mathcal{X}} \min\{P_X(A), 1-P_X(A)\}\Big)\min\limits_{x\in\mathcal{X}} P_X(x)\Big)$ из теоремы 4 – вторая сверху.

Рис. 1. Графики границ на коэффициенты сжатия из следствия 1 и теоремы 4 для двух бернуллиевских случайных величин (X, Y) с совместной вероятностной мерой, состоящей из $P_X = (P_X(0), P_X(1))$ и $P_{Y|X}$, заданных в (28), для вероятности перехода в канале $\alpha \in [0, 1]$.

что следует из (29) и того факта, что

$$\max_{A\subseteq\mathcal{X}} \min\{P_X(A), 1-P_X(A)\} = \frac{1}{2}.$$

Эта верхняя граница точнее, чем тривиальная граница, равная 1, когда

$$2(1-2\alpha)^2 < 1 \quad \Longleftrightarrow \quad \frac{2-\sqrt{2}}{4} < \alpha < \frac{2+\sqrt{2}}{4}. \tag{49}$$

Заметим также, что эта верхняя граница не достигается с равенством при таком сценарии, поскольку $(1-2\alpha)^2 = \eta_{\mathrm{KL}}(P_X, P_{Y|X}) = \eta_{\chi^2}(P_X, P_{Y|X})$, как показано в (29).

Во-вторых, наши доказательства теорем 3 и 4 опираются на обобщения хорошо известного *неравенства Пинскера* (или неравенства Чисара – Кемпермана – Кульбака – Пинскера; см. [80, § V]), оценивающие сверху расстояние по вариации через КЛ-дивергенцию и другие f-дивергенции. Таким образом, возникает естественный вопрос: являются ли эти границы более точными, чем граница расстояния по вариации из п. 7 предложения 5? Как показывает следующий пример, в некоторых режимах наши границы точнее. Пусть (X, Y) – равномерные бернуллиевские случайные величины с $P_X = (1/2, 1/2)$ и $P_{Y|X}$, заданными в (28), для некоторого $\alpha \in [0, 1]$. Тогда (48) задает верхние границы из следствия 1 и теоремы 4, и граница расстояния по вариации имеет вид

$$\eta_{\mathrm{KL}}(P_X, P_{Y|X}) \leqslant \eta_{\mathrm{KL}}(P_{Y|X}) \leqslant \eta_{\mathrm{TV}}(P_{Y|X}) = |1-2\alpha| \tag{50}$$

с учетом определения 5, п. 7 предложения 5 и соотношения (43). Следовательно, наша граница (48) точнее, чем граница через η_{TV}, когда

$$2(1-2\alpha)^2 < |1-2\alpha| \quad \Longleftrightarrow \quad \frac{1}{4} < \alpha < \frac{1}{2} \quad \text{или} \quad \frac{1}{2} < \alpha < \frac{3}{4}. \tag{51}$$

Поскольку наши верхние границы могут быть больше единицы (см. (49)), мы не можем надеяться превзойти границу через η_{TV} ($\leqslant 1$) во всех режимах. С другой стороны, преимуществом наших верхних границ является то, что они "подходят" к нижней границе через η_{χ^2} в п. 7 предложения 3; полезное применение этого будет показано в п. 4.3.

В-третьих, интуитивно можно ожидать, что граница между коэффициентами сжатия должна зависеть от мощности $|\mathcal{X}|$ или $|\mathcal{Y}|$. Поскольку минимальная вероятность во всех наших верхних границах соответствует $1/\min\limits_{x\in\mathcal{X}} P_X(x) \geqslant |\mathcal{X}|$, на первый взгляд можно было бы интерпретировать ее как "моделирующую" $|\mathcal{X}|$. К сожалению, эта интуитивная догадка неверна. Численные результаты, представленные графически на рис. 1b, показывают, что отношение

$$\eta_{\mathrm{KL}}(P_X, P_{Y|X}) / \eta_{\chi^2}(P_X, P_{Y|X})$$

значительно увеличивается возле граничного значения $\mathcal{P}_{\mathcal{X}}$, когда любая из компонент P_X близка к нулю. Этот эффект, хотя и неудивительный с учетом поведения вероятностных симплексов в их граничных точках относительно КЛ-дивергенции в качестве расстояния, адекватно учитывается верхними границами в следствии 1 и теореме 4, поскольку величина $1/\min\limits_{x\in\mathcal{X}} P_X(x)$ возрастает, когда любая из компонент вероятностной меры на входе стремится к нулю (см. рис. 1b). Очевидно, линейные верхние границы на $\eta_f(P_X, P_{Y|X})$, выражаемые только через $|\mathcal{X}|$ или $|\mathcal{Y}|$, этот эффект отражать не могут. Это подтверждает необходимость присутствия минимальной вероятности в наших линейных границах.

Наконец, заметим, что неравенство $1/\min\limits_{x\in\mathcal{X}} P_X(x) \geqslant |\mathcal{X}|$ не препятствует возможности того, что $1/\min\limits_{x\in\mathcal{X}} P_X(x)$ будет гораздо больше, чем $|\mathcal{X}|$. Таким образом, при больших $|\mathcal{X}|$ наши границы могут стать слабыми (см. пример в п. 4.4). В результате границы из теоремы 3, следствия 1 и теоремы 4, как правило, представляют интерес в следующих случаях:

1. $|\mathcal{X}|$ и $|\mathcal{Y}|$ малы: рис. 1 показывает, что наши границы могут быть весьма точными, когда $|\mathcal{X}| = |\mathcal{Y}| = 2$;
2. Слабая зависимость, т.е. $I(X;Y)$ мала: такая ситуация естественным образом возникает при анализе эргодичности цепей Маркова – см. п. 4.3;
3. Произведение распределений: если соответствующая совместная вероятностная мера является произведением, можно использовать свойство тензоризации коэффициентов сжатия (п. 5 предложения 3) – см. п. 4.4.

3.3. Коэффициенты сжатия совместно гауссовских случайных величин. В этом пункте рассмотрим коэффициенты сжатия для КЛ- и χ^2-дивергенций, соответствующие двумерным гауссовским распределениям. Пусть X и Y – совместно гауссовские случайные величины. Их совместное распределение имеет один из трех возможных видов:

1. Величины X или Y постоянны п.н., и коэффициенты сжатия определяются как $\eta_{\mathrm{KL}}(P_X, P_{Y|X}) = \eta_{\chi^2}(P_X, P_{Y|X}) = 0$;
2. $aX + bY = c$ п.н. для некоторых констант $a, b, c \in \mathbb{R}$, таких что $a \neq 0$ и $b \neq 0$. Тогда непосредственной проверкой легко убедиться, что $\rho(X;Y)$=1, откуда получаем $\eta_{\mathrm{KL}}(P_X, P_{Y|X}) = \eta_{\chi^2}(P_X, P_{Y|X}) = 1$, поскольку определения 3, (23) и п. 7 предложения 3 справедливы для общих случайных величин (см. [7, формулы (9) и (13)]);
3. Существует совместная плотность распределения $P_{X,Y}$ относительно меры Лебега на $\mathbb{R}^2$. Если X и Y независимы, то $\eta_{\mathrm{KL}}(P_X, P_{Y|X}) = \eta_{\chi^2}(P_X, P_{Y|X}) = 0$, поскольку п. 2 предложения 3 справедлив в общем случае. Поэтому будем предполагать, что X и Y зависимы.

Нас будет интересовать последний невырожденный случай. Для простоты будем также предполагать, что X и Y имеют нулевое среднее. Тогда совместное распределение величин X и Y можно представить в "инновационном виде": $Y = \gamma X + W$ для некоторой константы $\gamma \neq 0$ и гауссовской случайной величины W с нулевым средним и ненулевой дисперсией, не зависящей от X. Так как для любой плотности распределения R_X справедливо $D_f(R_X \,\|\, P_X) = D_f(R_{\gamma X} \,\|\, P_{\gamma X})$, то $\eta_{\mathrm{KL}}(P_X, P_{Y|X}) = \eta_{\mathrm{KL}}(P_{\gamma X}, P_{Y|\gamma X})$ и $\eta_{\chi^2}(P_X, P_{Y|X}) = \eta_{\chi^2}(P_{\gamma X}, P_{Y|\gamma X})$ (где $R_{\gamma X}$ – производная плотность распределения, соответствующая R_X). Поэтому без ограничения общности положим $\gamma = 1$ и будем рассматривать классическую модель *аддитивного белого гауссовского шума* (АБГШ) [41, гл. 9]:

$$Y = X + W, \quad X \perp\!\!\!\perp W, \tag{52}$$

где входом является $X \sim \mathcal{N}(0, \sigma_X^2)$ с $\sigma_X^2 > 0$ (т.е. X имеет гауссовскую плотность распределения P_X со средним 0 и дисперсией σ_X^2), гауссовским шумом является $W \sim \mathcal{N}(0, \sigma_W^2)$ с $\sigma_W^2 > 0$, причем X и W независимы. Это соотношение также задает условные плотности распределения

$$P_{Y|X} = \left\{P_{Y|X=x} = \mathcal{N}(x, \sigma_W^2) : \; x \in \mathbb{R}\right\}$$

и маргинальную плотность распределения $P_Y = \mathcal{N}(0, \sigma_X^2 + \sigma_W^2)$.

Напомним, что для любой пары функций плотности распределения относительно меры Лебега Leb на $\mathbb{R}$ их КЛ- и χ^2-дивергенции определяются (аналогично (3) и (4)) как

$$D(R_X \,\|\, S_X) \triangleq \begin{cases} \displaystyle\int_{\mathbb{R}} R_X(x) \log\left(\frac{R_X(x)}{S_X(x)}\right) d\,\mathrm{Leb}(x), \\ \qquad\qquad \mathrm{Leb}(\{x \in \mathbb{R} : \; R_X(x) > 0, \; S_X(x) = 0\}) = 0, \\ +\infty \qquad\quad \text{в противном случае,} \end{cases} \tag{53}$$

$$\chi^2(R_X \,\|\, S_X) \triangleq \begin{cases} \displaystyle\int_{\mathbb{R}} \frac{(R_X(x)-S_X(x))^2}{S_X(x)}\, d\,\mathrm{Leb}(x), \\ \qquad\qquad \mathrm{Leb}(\{x\in\mathbb{R}:\ R_X(x)>0,\ S_X(x)=0\})=0, \\ +\infty \qquad\quad \text{в противном случае,} \end{cases} \tag{54}$$

где используются интегралы Лебега и приняты соглашения $0\log(0/t)=0$ для любых $t\geqslant 0$ (следуя соображениям непрерывности) и $(0-0)^2/0=0$. Для совместно гауссовской плотности распределения $P_{X,Y}$, описанной в (52), с учетом определений (53) и (54) коэффициенты сжатия для КЛ- и χ^2-дивергенций имеют вид (см. (17) и (22))

$$\eta_{\mathrm{KL}}(P_X, P_{Y|X}) = \sup_{R_X:\, 0<D(R_X\|P_X)<+\infty} \frac{D(R_Y\,\|\,P_Y)}{D(R_X\,\|\,P_X)}, \tag{55}$$

$$\eta_{\chi^2}(P_X, P_{Y|X}) = \sup_{R_X:\, 0<\chi^2(R_X\|P_X)<+\infty} \frac{\chi^2(R_Y\,\|\,P_Y)}{\chi^2(R_X\,\|\,P_X)}, \tag{56}$$

где супремумы берутся по всем плотностям распределения R_X (отличающихся от P_X на множестве ненулевой меры Лебега)[6], а через R_Y обозначена маргинальная плотность распределения для Y после прохождения R_X через модель АБГШ $P_{Y|X}$. В частности, $R_Y = R_X * P_W$, где $P_W=\mathcal{N}(0,\sigma_W^2)$, а $*$ означает операцию *свертки*. Далее, определим *коэффициент сжатия для КЛ-дивергенции при ограничении на среднюю мощность* (или ограничении на второй момент) $p\geqslant\sigma_X^2$ как

$$\eta_{\mathrm{KL}}^{(p)}(P_X, P_{Y|X}) \triangleq \sup_{\substack{R_X:\, \mathbf{E}_{R_X}[X^2]\leqslant p \\ 0<D(R_X\|P_X)<+\infty}} \frac{D(R_Y\,\|\,P_Y)}{D(R_X\,\|\,P_X)}, \tag{57}$$

где супремум берется по всем плотностям распределения R_X, удовлетворяющим ограничению на среднюю мощность $\mathbf{E}_{R_X}[X^2]\leqslant p$. Отметим, что при $p=+\infty$ получаем стандартный коэффициент сжатия из (55).

Из литературы хорошо известно, что $\eta_{\mathrm{KL}}(P_X, P_{Y|X}) = \eta_{\chi^2}(P_X, P_{Y|X})$ для совместно гауссовской плотности распределения $P_{X,Y}$, определенной в (52). Например, в [2, теорема 7] этот результат доказан в контексте теории инвестиционного портфеля, в [81, с. 2] доказано его обобщение в контексте гауссовской гиперсжимаемости, а в [78, § 5.2, п. 5] он доказан в попытке аксиоматизации η_{KL}. В то время как доказательства в [2, теоремы 6 и 7; 78, § 5.2, п. 5] используют характеризацию η_{KL} через взаимную информацию (18) (см. [7, теорема 4]), в § 5 мы даем альтернативное доказательство этого результата, непосредственно использующее определение η_{KL} через КЛ-дивергенцию в (55). При этом наше доказательство также показывает, что $\eta_{\mathrm{KL}}^{(p)}(P_X, P_{Y|X})$ равно $\eta_{\chi^2}(P_X, P_{Y|X})$ для любого $p\in[\sigma_X^2,\infty]$. Хотя этот последний результат вытекает из нашего доказательства, ранее в литературе, насколько нам известно, он не встречался. Формально этот результат представляет следующая

Теорема 5 (гауссовские коэффициенты сжатия). *Для заданной совместно гауссовской плотности распределения $P_{X,Y}$, определенной в* (52), *с маргинальной плотностью распределения $P_X=\mathcal{N}(0,\sigma_X^2)$ на входе и условными плотностями распределения $P_{Y|X}=\{P_{Y|X=x}=\mathcal{N}(x,\sigma_W^2):\ x\in\mathbb{R}\}$, такими что $\sigma_X^2,\sigma_W^2>0$,*

[6] Если P_X – общая вероятностная мера, а $P_{Y|X}$ – марковское ядро между двумя измеримыми пространствами, то коэффициенты сжатия для КЛ- и χ^2-дивергенций определяются в точности как в (17) и (22) с помощью определений КЛ- и χ^2-дивергенций с позиций теории меры [7, § 2]. В (55) при оптимизации по всем вероятностным мерам R_X на $\mathbb{R}$ (с борелевской σ-алгеброй) из ограничения $D(R_X\,\|\,P_X)<+\infty$ вытекает, что R_X должна быть абсолютно непрерывной относительно гауссовского распределения P_X, см. [38, п. 1.6]. Следовательно, супремум в (55) можно брать по всем плотностям распределения R_X, таким что $0<D(R_X\,\|\,P_X)<+\infty$. То же самое относится и к (56).

следующие величины эквивалентны:

$$\eta_{\mathrm{KL}}(P_X, P_{Y|X}) = \eta_{\mathrm{KL}}^{(p)}(P_X, P_{Y|X}) = \eta_{\chi^2}(P_X, P_{Y|X}) = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_W^2},$$

где наложено ограничение на среднюю мощность $p \geqslant \sigma_X^2$.

Как уже сказано, мы доказываем этот результат в § 5. В отличие от теоремы 5, где $\eta_{\mathrm{KL}}(P_X, P_{Y|X})$, $\eta_{\mathrm{KL}}^{(p)}(P_X, P_{Y|X})$ и $\eta_{\chi^2}(P_X, P_{Y|X})$ могут быть строго меньше 1, коэффициенты сжатия для КЛ- и χ^2-дивергенций для марковского ядра АБГШ $P_{Y|X}$ (т.е. в условиях определения 5) равны 1 вне зависимости от того, накладываются ли ограничения на второй момент (см. [6, § 1.2; 82, § 1]).

3.4. Предпорядок меньшего искажения и операторная выпуклость. Последний наш основной результат состоит в эквивалентной характеризации предпорядка меньшего искажения на стохастических матрицах, обобщающей результат предложения 6. Начнем с определения предпорядка меньшего искажения. В постановке с конечным алфавитом из п. 2.1 рассмотрим случайную величину $X \in \mathcal{X}$ на входе и две случайные величины $Y \in \mathcal{Y}$ и $Z \in \mathcal{Z}$ на выходе, где $\mathcal{Z} \triangleq \{1, \ldots, |\mathcal{Z}|\}$, $2 \leqslant |\mathcal{Z}| < +\infty$.

Определение 6 (предпорядок меньшего искажения [20]). Пусть $P_{Y|X} = W \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ и $P_{Z|X} = V \in \mathcal{P}_{\mathcal{Z}|\mathcal{X}}$ – две стохастические по строкам матрицы с одним алфавитом на входе $\mathcal{X}$, т.е. с одинаковым числом строк. Будем говорить, что $P_{Y|X}$ *менее искажающая*, чем $P_{Z|X}$, и обозначать это через $P_{Y|X} \succeq_{\ln} P_{Z|X}$, если

$$\forall R_X, P_X \in \mathcal{P}_{\mathcal{X}}, \quad D(R_X W \,\|\, P_X W) \geqslant D(R_X V \,\|\, P_X V).$$

Непосредственной проверкой легко убедиться, что определение 6 задает предпорядок на стохастических матрицах[7]. Более того, из этого определения следует, что пара вероятностных мер $R_X W$ и $P_X W$ всегда "более различима", чем пара $R_X V$ и $P_X V$, что в действительности интуитивно соответствует тому, что W "менее искажающая", чем V (если рассматривать их как стохастические ядра). Имеется несколько других эквивалентных описаний отношения $\succeq_{\ln}$, например, через кодирование каналов [20, определение B, предложение 2], взаимную информацию [20, предложение 2] и функционал ван Дейка [84, теорема 2]. Подробнее о предпорядке меньшего искажения см. работу [83, пп. I-B, I-D, II-A, IV] и библиографию в ней.

В [7, § 6] показано, что если заданная стохастическая матрица мажорируется в смысле предпорядка меньшего искажения матрицей канала со стиранием, то это тесно связано с коэффициентом сжатия для КЛ-дивергенции для этой стохастической матрицы. Напомним, что через $E_{1-\beta} \in \mathcal{P}_{\mathcal{X}\cup\{\mathrm{e}\}|\mathcal{X}}$ мы обозначаем матрицу $|\mathcal{X}|$-ичного канала со стиранием с вероятностью стирания $1-\beta \in [0,1]$ (согласно определению (30)). Из [7, предложение 15] можно вывести, что для любой стохастической матрицы $P_{Y|X} \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ справедливо

$$\eta_{\mathrm{KL}}(P_{Y|X}) = \min\{\beta \in [0,1] : \; E_{1-\beta} \succeq_{\ln} P_{Y|X}\}, \tag{58}$$

где множество, по которому ведется минимизация, всегда содержит $\beta = 1$, поскольку по сути E_0 – единичная матрица и $E_0 \succeq_{\ln} P_{Y|X}$. В [83, п. IV-A] отмечено, что хотя (58) показывает, что η_{KL} характеризует мажорирование матрицами канала со

[7] Как указано в [83, сноска 1], мы называем отношение меньшего искажения *предпорядком*, а не *частичным порядком*, поскольку мы не рассматриваем классы эквивалентности стохастических по строкам матриц, например, отождествляя стохастические по строкам матрицы, полученные перестановкой столбцов.

стиранием в смысле меньшего искажения, формула (39) показывает, что и η_{χ^2} характеризует это мажорирование. Отсюда возникает вопрос: характеризует ли χ^2-дивергенция предпорядок меньшего искажения в общем случае? В [83, теорема 1] и [85, теорема 1] на этот вопрос дается утвердительный ответ с помощью характеризации отношения $\succeq_{\ln}$ через χ^2-дивергенцию, обобщая тем самым соотношение (39).

Вдохновляясь этими результатами, рассмотрим предложение 6 (см. [18, теорема 1]), показывающее, что $\eta_{\mathrm{KL}}(P_{Y|X}) = \eta_f(P_{Y|X})$ для всех нелинейных операторно выпуклых функций f (определенных в п. 6.1). Как и выше, отсюда возникает вопрос: *характеризуют ли нелинейные операторно выпуклые f-дивергенции предпорядок меньшего искажения в общем случае*? Следующая теорема отвечает на этот вопрос, обобщая как предложение 6, так и [83, теорема 1], и показывая, что нелинейные операторно выпуклые f-дивергенции также характеризуют предпорядок меньшего искажения.

Т е о р е м а 6 (эквивалентные характеризации отношения $\succeq_{\ln}$). *Рассмотрим произвольную нелинейную операторно выпуклую функцию $f\colon (0,\infty) \to \mathbb{R}$, такую что $f(1) = 0$. Тогда для любых стохастических матриц $P_{Y|X} = W \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ и $P_{Z|X} = V \in \mathcal{P}_{\mathcal{Z}|\mathcal{X}}$ с одинаковым алфавитом на входе $\mathcal{X}$ отношение $P_{Y|X} \succeq_{\ln} P_{Z|X}$ имеет место тогда и только тогда, когда*

$$\forall R_X, P_X \in \mathcal{P}_{\mathcal{X}}, \quad D_f(R_X W \,\|\, P_X W) \geqslant D_f(R_X V \,\|\, P_X V).$$

Теорема 6 доказана в п. 6.2 с использованием техники из [18]. Хорошо известная *теорема Лёвнера – Хайнца* утверждает, что функции $f(t) = t\log(t)$ и $f(t) = \dfrac{t^\alpha - 1}{\alpha - 1}$ для любого $\alpha \in (0,1) \cup (1,2]$ являются операторно выпуклыми (см., например, [86, теорема 2.6; 87, теоремы V.2.5 и V.2.10, упражнения V.2.11 и V.2.13], и можно применить свойство аффинного преобразования из п. 6.1). Следовательно, одним классом f-дивергенций, удовлетворяющих условиям теоремы, являются дивергенции Хеллингера порядка $\alpha \in (0,2]$, где случаи $\alpha = 1$ и $\alpha = 2$ отвечают КЛ- и χ^2-дивергенциям соответственно.

Теперь в качестве примера применения теоремы 6 докажем обобщение так называемого *СНОД Самородницкого*. Следуя изложению в [7, п. 6.2], рассмотрим дискретные случайные величины с конечными множествами значений $U \in \mathcal{U}$, $X_1^n = (X_1, \ldots, X_n)$ и $Y_1^n = (Y_1, \ldots, Y_n)$, где $X_i \in \mathcal{X}_i$ и $Y_i \in \mathcal{Y}_i$ для всех $i \in \{1, \ldots, n\}$, а $n \in \mathbb{N} \setminus \{0\}$. Для заданных стохастических матриц $P_{Y_i|X_i} \in \mathcal{P}_{\mathcal{Y}_i|\mathcal{X}_i}$, $i \in \{1, \ldots, n\}$, задающих условные распределения каждой из Y_i при заданном X_i, пусть условное распределение $Y_1^n \in \mathcal{Y}^n = \mathcal{Y}_1 \times \ldots \times \mathcal{Y}_n$ при заданном $X_1^n \in \mathcal{X}^n = \mathcal{X}_1 \times \ldots \times \mathcal{X}_n$ задается произведением стохастических матриц $P_{Y_1^n|X_1^n} \in \mathcal{P}_{\mathcal{Y}^n|\mathcal{X}^n}$:

$$P_{Y_1^n|X_1^n} = P_{Y_1|X_1} \otimes P_{Y_2|X_2} \otimes \ldots \otimes P_{Y_n|X_n}, \tag{59}$$

где через $\otimes$ обозначено кронекерово (или тензорное) произведение. Заметим, что такие $P_{Y_1^n|X_1^n}$ известны как стохастические ядра *без памяти* (см., например, [41, формула (7.27)]). Определим коэффициент сжатия стохастической матрицы $P_{Y_i|X_i}$ для любой f-дивергенции (см. определение 5) как

$$\eta_i \triangleq \eta_f(P_{Y_i|X_i}) \tag{60}$$

для любого $i \in \{1, \ldots, n\}$. В то время как СНОД для $P_{Y_1^n|X_1^n}$ характеризуется коэффициентом сжатия $\eta_f(P_{Y_1^n|X_1^n})$, хотелось бы получить ослабленный вариант этого СНОД через "однобуквенные" коэффициенты сжатия $\{\eta_i : i \in \{1, \ldots, n\}\}$.

Чтобы проиллюстрировать одно такое тензоризованное СНОД, предположим, что $\eta_i = \eta$ для всех $i \in \{1, \ldots, n\}$, и зафиксируем произвольную нелинейную опера-

торно выпуклую функцию $f\colon (0,\infty)\to\mathbb{R}$, такую что $f(1)=0$. С помощью предложения 6 непосредственной проверкой легко убедиться, что утверждения [7, теорема 5 и следствие 6] справедливы для любой f-дивергенции с нелинейной операторно выпуклой функцией f (а не только для КЛ-дивергенции). В результате из [7, следствие 6] вытекает граница тензоризации

$$\eta_f(P_{Y_1^n|X_1^n}) \leqslant 1-(1-\eta)^n, \tag{61}$$

которую можно считать аналогом п. 5 предложения 3 для коэффициентов сжатия стохастических матриц. Таким образом, для любой пары вероятностных мер на входе $R_{X_1^n}, P_{X_1^n}\in\mathcal{P}_{\mathcal{X}^n}$ имеется тензоризованное СНОД

$$D_f(R_{Y_1^n}\,\|\,P_{Y_1^n}) \leqslant (1-(1-\eta)^n) D_f(R_{X_1^n}\,\|\,P_{X_1^n}), \tag{62}$$

где $R_{Y_1^n}=R_{X_1^n}P_{Y_1^n|X_1^n}\in\mathcal{P}_{\mathcal{Y}^n}$ и $P_{Y_1^n}=P_{X_1^n}P_{Y_1^n|X_1^n}\in\mathcal{P}_{\mathcal{Y}^n}$ – вероятностные меры на выходе после прохождения $R_{X_1^n}$ и $P_{X_1^n}$ через $P_{Y_1^n|X_1^n}$ соответственно. Аналогично, для любого совместного распределения P_{U,X_1^n}, такого что $U\to X_1^n\to Y_1^n$ образуют цепь Маркова, имеется тензоризованное СНОД (см. (8) и (38))

$$I_f(U;Y_1^n) \leqslant (1-(1-\eta)^n) I_f(U;X_1^n). \tag{63}$$

Однако, как указано в [7, п. 6.2], "можно было бы указать более сильные тензоризованные границы, если бы мы обладали лучшим знанием" о парах $(R_{X_1^n}, P_{X_1^n})$ или о распределении P_{U,X_1^n}. В этом смысле следующая теорема дает более точные границы на $D_f(R_{Y_1^n}\,\|\,P_{Y_1^n})$ и $I_f(U;Y_1^n)$ через однобуквенные коэффициенты сжатия $\{\eta_i\,:\,i\in\{1,\dots,n\}\}$ и величины, представляющие "среднюю" f-дивергенцию на входе и "среднюю" взаимную f-информацию, содержащиеся в подмножествах случайной величины X_1^n, соответственно.

Теорема 7 (обобщенное СНОД Самородницкого). *Рассмотрим произвольную нелинейную операторно выпуклую функцию $f\colon(0,\infty)\to\mathbb{R}$, такую что $f(1)=0$. Пусть $U\in\mathcal{U}$, $X_1^n\in\mathcal{X}^n$ и $Y_1^n\in\mathcal{Y}^n$ – дискретные случайные величины с заданной стохастической матрицей-произведением $P_{Y_1^n|X_1^n}\in\mathcal{P}_{\mathcal{Y}^n|\mathcal{X}^n}$ из (59). Пусть S – случайное подмножество множества $\{1,\dots,n\}$, полученное независимым выбором каждого элемента $i\in\{1,\dots,n\}$ с вероятностью η_i (заданной в (60)), и пусть S не зависит от (U,X_1^n,Y_1^n). Тогда для любой пары вероятностных мер на входе $R_{X_1^n}, P_{X_1^n}\in\mathcal{P}_{\mathcal{X}^n}$ имеет место неравенство*

$$D_f\big(R_{Y_1^n}\,\|\,P_{Y_1^n}\big) \leqslant \sum_{T\subseteq\{1,\dots,n\}} P_S(T) D_f(R_{X_T}\,\|\,P_{X_T}), \tag{64}$$

где P_S – распределение вероятностей для S, $X_T\triangleq\{X_k\,:\,k\in T\}$ для любого подмножества $T\subseteq\{1,\dots,n\}$, $D_f(R_{X_\varnothing}\,\|\,P_{X_\varnothing})=0$, $R_{Y_1^n}=R_{X_1^n}P_{Y_1^n|X_1^n}\in\mathcal{P}_{\mathcal{Y}^n}$ и $P_{Y_1^n}=P_{X_1^n}P_{Y_1^n|X_1^n}\in\mathcal{P}_{\mathcal{Y}^n}$. Аналогично, для любого совместного распределения P_{U,X_1^n}, такого что $U\to X_1^n\to Y_1^n$ образуют цепь Маркова,

$$I_f(U;Y_1^n) \leqslant I_f(U;X_S,S) = \sum_{T\subseteq\{1,\dots,n\}} P_S(T) I_f(U;X_T), \tag{65}$$

где $I_f(U;X_\varnothing)=0$. Более того, если $\eta_i=\eta$ для всех $i\in\{1,\dots,n\}$, то

$$I_f(U;Y_1^n) \leqslant \sum_{k=0}^{n}\binom{n}{k}\eta^k(1-\eta)^{n-k} I_k, \tag{66}$$

где для каждого $k \in \{0, \dots, n\}$ *величина* I_k *определяется как "средняя" взаимная* f*-информация, содержащаяся в подмножествах* X_1^n *мощности* k:

$$I_k \triangleq \binom{n}{k}^{-1} \sum_{\substack{T \subseteq \{1,\dots,n\} \\ |T|=k}} I_f(U; X_T).$$

Теорема 7 будет доказана в п. 6.3 с помощью теоремы 6, следуя технике доказательства из [7]. Случай КЛ-дивергенции в теореме 7 был впервые доказан Самородницким в [88] с помощью техники линейного программирования при доказательстве частного случая *гипотезы Куртада – Кумара* из [89]. К ее нынешнему виду она была приведена в [7, теорема 20, замечание 6], где также было дано более простое доказательство. Наш результат в теореме 7 обобщает этот случай КЛ-дивергенции на все нелинейные операторно выпуклые f-дивергенции (включая, например, все дивергенции Хеллингера порядка $\alpha \in (0, 2]$, как отмечалось выше). Заметим также, что случай КЛ-дивергенции имеет и другие приложения, такие как усиление *леммы Гербер* (см. [90, п. 2.1]) в [7, замечание 5]. И наконец, как показано в [7, замечание 4], если $\eta_i = \eta$ для всех $i \in \{1, \dots, n\}$ в теореме 7, то аппроксимация распределения $\mathrm{binomial}(n, \eta)$ в (66) с математическим ожиданием $n\eta$ дает

$$I_f(U; Y_1^n) \lesssim I_{n\eta} \tag{67}$$

для любой цепи Маркова $U \to X_1^n \to Y_1^n$; это показывает, что из Y_1^n можно извлечь лишь информацию об U, содержащуюся в подмножествах X_1^n мощности, ограниченной величиной $n\eta$.

§ 4. Доказательства линейных границ между коэффициентами сжатия

В этом параграфе мы докажем теоремы 3 и 4. Основная идея доказательства состоит в оценивании сверху и снизу f-дивергенций в числителе и знаменателе определения 2, соответственно, через χ^2-дивергенции. С этой целью в п. 4.1 мы приведем несколько простых границ между f-дивергенциями и χ^2-дивергенцией, а в п. 4.2 докажем основные результаты.

4.1. Границы на f-дивергенции через χ^2-дивергенцию. Вначале приведем границы между КЛ- и χ^2-дивергенцией. Для вывода нашей нижней границы на КЛ-дивергенцию нам понадобится следующее "зависящее от распределений уточнение неравенства Пинскера", доказанное в [91].

Лемма 1 (зависящее от распределений неравенство Пинскера [91, теорема 2.1]). *Для любых вероятностных мер* $R_X, P_X \in \mathcal{P}_\mathcal{X}$ *справедливо неравенство*

$$D(R_X \,\|\, P_X) \geqslant \varphi\Big(\max_{A \subseteq \mathcal{X}} \min\{P_X(A), 1 - P_X(A)\}\Big) \|R_X - P_X\|_{\mathrm{TV}}^2,$$

где функция $\varphi\colon [0, 1/2] \to \mathbb{R}$ *определяется как*

$$\varphi(p) \triangleq \begin{cases} \dfrac{1}{1-2p} \log\Big(\dfrac{1-p}{p}\Big), & p \in [0, 1/2), \\ 2, & p = 1/2. \end{cases} \tag{68}$$

Более того, в этом неравенстве участвует оптимальная зависящая от распределений константа в том смысле, что для любого фиксированного $P_X \in \mathcal{P}_\mathcal{X}$

$$\inf_{R_X \in \mathcal{P}_\mathcal{X} \setminus \{P_X\}} \frac{D(R_X \,\|\, P_X)}{\|R_X - P_X\|_{\mathrm{TV}}^2} = \varphi\Big(\max_{A \subseteq \mathcal{X}} \min\{P_X(A), 1 - P_X(A)\}\Big).$$

Напомним (см., например, [41, лемма 11.6.1]), что неравенство Пинскера утверждает, что для любых $R_X, P_X \in \mathcal{P}_{\mathcal{X}}$

$$D(R_X \,\|\, P_X) \geqslant 2 \,\|R_X - P_X\|_{\mathrm{TV}}^2 . \tag{69}$$

Следовательно, лемма 1 точнее, чем неравенство Пинскера, поскольку имеет место неравенство $0 \leqslant \max\limits_{A \subseteq \mathcal{X}} \min\{P_X(A), 1 - P_X(A)\} \leqslant 1/2$, и поэтому

$$\varphi\Big(\max_{A \subseteq \mathcal{X}} \min\{P_X(A), 1 - P_X(A)\} \Big) \geqslant 2, \tag{70}$$

где равенство имеет место тогда и только тогда, когда

$$\max_{A \subseteq \mathcal{X}} \min\{P_X(A), 1 - P_X(A)\} = \frac{1}{2}$$

(см. [91, § III]). Следующая лемма использует лемму 1 для оценки КЛ-дивергенции снизу через χ^2-дивергенцию.

Лемма 2 (нижняя граница КЛ-дивергенции). *Для любых вероятностных мер $R_X, P_X \in \mathcal{P}_{\mathcal{X}}$ справедливо*

$$D(R_X \,\|\, P_X) \geqslant \frac{\varphi\Big(\max\limits_{A \subseteq \mathcal{X}} \min\{P_X(A), 1 - P_X(A)\} \Big) \min\limits_{x \in \mathcal{X}} P_X(x)}{2} \chi^2(R_X \,\|\, P_X),$$

где функция $\varphi\colon [0, 1/2] \to \mathbb{R}$ определена в (68).

Доказательство. Если $R_X = P_X$ или $\min\limits_{x \in \mathcal{X}} P_X(x) = 0$, то неравенство выполнено тривиальным образом. (Заметим, что если $\min\limits_{x \in \mathcal{X}} P_X(x) = 0$ и $P_X(x^*) = 0 < R_X(x^*)$ для некоторого $x^* \in \mathcal{X}$, то $D(R_X \,\|\, P_X) = \chi^2(R_X \,\|\, P_X) = +\infty$, и можно считать, что неравенство выполнено.) Итак, без ограничения общности будем предполагать, что $R_X \neq P_X$ и $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$.

Так как χ^2-дивергенция похожа на взвешенную ℓ^2-норму, вначале применим лемму 1 и получим нижнюю границу

$$D(R_X \,\|\, P_X) \geqslant \varphi\Big(\max_{A \subseteq \mathcal{X}} \min\{P_X(A), 1 - P_X(A)\} \Big) \frac{\|R_X - P_X\|_1^2}{4}, \tag{71}$$

где используется характеризация расстояния по вариации через ℓ^1-норму, приведенная в (2). Далее с учетом (4) заметим, что

$$\chi^2(R_X \,\|\, P_X) = \sum_{x \in \mathcal{X}} |R_X(x) - P_X(x)| \left| \frac{R_X(x) - P_X(x)}{P_X(x)} \right| \leqslant \frac{\|R_X - P_X\|_\infty}{\min\limits_{x \in \mathcal{X}} P_X(x)} \,\|R_X - P_X\|_1 .$$

Отсюда следует

$$\begin{aligned}
&\frac{\|R_X - P_X\|_1^2}{\min\limits_{x \in \mathcal{X}} P_X(x)} \geqslant \chi^2(R_X \,\|\, P_X) \frac{\|R_X - P_X\|_1}{\|R_X - P_X\|_\infty} \geqslant \\
&\geqslant \chi^2(R_X \,\|\, P_X) \min_{\substack{S_X, Q_X \in \mathcal{P}_{\mathcal{X}} \\ S_X \neq Q_X}} \frac{\|S_X - Q_X\|_1}{\|S_X - Q_X\|_\infty} = \\
&= 2\chi^2(R_X \,\|\, P_X),
\end{aligned} \tag{72}$$

где использован тот факт, что

$$\min_{\substack{S_X,Q_X\in\mathcal{P}_\mathcal{X}\\ S_X\neq Q_X}} \frac{\|S_X-Q_X\|_1}{\|S_X-Q_X\|_\infty}=2. \tag{73}$$

Для доказательства (73) заметим, что для любых $S_X, Q_X \in \mathcal{P}_\mathcal{X}$ (см., например, [92, лемма 1])

$$\|S_X-Q_X\|_\infty \leqslant \frac{1}{2}\|S_X-Q_X\|_1,$$

поскольку $(S_X-Q_X)\mathbf{1}=0$, и это неравенство на самом деле может быть точным. Например, выберем любую вероятностную меру $Q_X \in \mathcal{P}_\mathcal{X}^\circ$ и положим $x_0 = \arg\min_{x\in\mathcal{X}} Q_X(x)$. Затем возьмем $S_X \in \mathcal{P}_\mathcal{X}$, такую что $S_X(x_0)=Q_X(x_0)+\delta$ для некоторого достаточно малого $\delta>0$, $S_X(x_1)=Q_X(x_1)-\delta$ для некоторого $x_1\in\mathcal{X}\setminus\{x_0\}$ и $S_X(x)=Q_X(x)$ для всех остальных $x\in\mathcal{X}\setminus\{x_0,x_1\}$. При таком выборе S_X и Q_X получаем $\|S_X-Q_X\|_\infty=\delta=\frac{1}{2}\|S_X-Q_X\|_1$.

Наконец, объединяя (71) и (72), получаем

$$D(R_X\,\|\,P_X)\geqslant \frac{\varphi\Big(\max\limits_{A\subseteq\mathcal{X}}\min\{P_X(A),1-P_X(A)\}\Big)\min\limits_{x\in\mathcal{X}}P_X(x)}{2}\chi^2(R_X\,\|\,P_X),$$

что завершает доказательство. ▲

Заметим, что если применить (70) к лемме 2, или, что то же самое, использовать стандартное неравенство Пинскера (69) вместо леммы 1 в доказательстве леммы 2, то получим хорошо известное более слабое неравенство (см., например, [34, формула (338)])

$$D(R_X\,\|\,P_X)\geqslant \min_{x\in\mathcal{X}}P_X(x)\chi^2(R_X\,\|\,P_X) \tag{74}$$

для любых $R_X, P_X \in \mathcal{P}_\mathcal{X}$.

Стоит отметить, что систематическим методом вывода оптимальных не зависящих от распределений границ между любыми двумя f-дивергенциями является метод *совместной области значений Харремоеса – Вайды* [93][8]. Однако для вывода нижних границ на КЛ-дивергенцию через χ^2-дивергенцию мы не можем использовать эту технику, поскольку таких общих нижних границ не существует (когда оба распределения на входе изменяются) [38, п. 7.3]. С другой стороны, зависящие от распределений границы легко можно находить, используя подходящую технику в каждом конкретном случае. Наше доказательство леммы 2 использует один из таких подходов, основанный на неравенстве Пинскера.

Хотелось бы попытаться улучшить лемму 2, используя лучшие нижние границы на КЛ-дивергенцию через расстояние по вариации. Так, наилучшей возможной границей снизу на КЛ-дивергенцию через расстояние по вариации является нижняя граница их совместной области значений Харремоеса – Вайды (см. [93, рис. 1]). Эта нижняя граница, известная как *точная нижняя граница Вайды*, дает наименьшую возможную КЛ-дивергенцию для любого значения расстояния по вариации и полностью описывается параметрической формулой в [94, теорема 1] (см. также [38, п. 7.2.2]). Хотя точная нижняя граница Вайды дает нелинейную границу снизу на КЛ-дивергенцию через χ^2-дивергенцию, эту нижнюю границу трудно применить в сочетании с леммой 3 (приведенной ниже) для вывода нелинейной верхней

[8] "Не зависящей от распределений" границей между двумя f-дивергенциями называется граница, зависящая от распределений на входе только через соответствующие f-дивергенции.

границы на отношение КЛ-дивергенций через отношение χ^2-дивергенций (см. доказательство теоремы 4 в п. 4.2). По этой причине мы прибегаем к использованию простых линейных границ между КЛ- и χ^2-дивергенцией, что приводит к линейной границе в теореме 4.

Другой, более тонкой причиной использования χ^2-дивергенции для доказательства линейной нижней границы на КЛ-дивергенцию является возможность применить лемму 1. Хотя неравенство Пинскера и является наилучшей нижней границей на КЛ-дивергенцию через квадрат расстояния по вариации по всем парам вероятностных мер на входе (см., например, [94, формула (9)]), коэффициенты сжатия в п. 3.2 имеют фиксированную маргинальную вероятностную меру P_X. Поэтому можно использовать зависящее от распределений улучшение неравенства Пинскера из леммы 1 для вывода более точной границы, чем (74).

Теперь приведем границу сверху на КЛ-дивергенцию через χ^2-дивергенцию, которая тривиально следует из неравенства Йенсена. Эта граница была выведена в контексте изучения эргодичности цепей Маркова в [95] и затем повторно получена при исследовании неравенств, связанных с f-дивергенциями (см. [96, 97], а также [98, теорема 5]).

Лемма 3 (верхняя граница КЛ-дивергенции [95]). *Для заданных вероятностных мер $P_X, R_X \in \mathcal{P}_\mathcal{X}$ справедливо неравенство*

$$D(R_X \,\|\, P_X) \leqslant \log\bigl(1 + \chi^2(R_X \,\|\, P_X)\bigr) \leqslant \chi^2(R_X \,\|\, P_X).$$

Доказательство. Для полноты изложения приведем доказательство леммы (см. [96]). Предположим без ограничения общности, что не существует $x \in \mathcal{X}$, такого что $R_X(x) > P_X(x) = 0$. (Если это не так, то $\chi^2(R_X \,\|\, P_X) = +\infty$, и неравенство тривиальным образом выполнено.) Итак, рассматривая в качестве $\mathcal{X}$ носитель P_X, полагаем, что $P_X \in \mathcal{P}_\mathcal{X}^\circ$ (гарантируя тем самым, что все дальнейшие величины будут конечными). Так как функция $x \mapsto \log(x)$ выпукла вверх, то из неравенства Йенсена получаем

$$\begin{aligned} D(R_X \,\|\, P_X) &= \sum_{x\in\mathcal{X}} R_X(x) \log\left(\frac{R_X(x)}{P_X(x)}\right) \leqslant \log\left(\sum_{x\in\mathcal{X}} \frac{R_X(x)^2}{P_X(x)}\right) = \\ &= \log\bigl(1 + \chi^2(R_X \,\|\, P_X)\bigr) \leqslant \chi^2(R_X \,\|\, P_X), \end{aligned}$$

где первое равенство следует из (3), третье – из (4) после некоторых преобразований, а последнее – из хорошо известного неравенства $\log(1+x) \leqslant x$ для всех $x > -1$. ▲

Отметим, что первая нелинейная граница в лемме 3 покрывает метод совместной области значений Харремоеса – Вайды [38, п. 7.3]. Хотя она и точнее, чем вторая линейная граница, в доказательстве теоремы 4 (как объяснялось выше) мы используем именно вторую. Вторая граница также была получена в [99, лемма 6.3].

Теперь приведем границы между общими f-дивергенциями и χ^2-дивергенцией. Для вывода нашей нижней границы на f-дивергенции вначале сформулируем обобщение неравенства Пинскера для f-дивергенций, доказанное в [79].

Лемма 4 (обобщенное неравенство Пинскера для f-дивергенции [79, теорема 3]). *Пусть задана выпуклая функция $f\colon (0,\infty) \to \mathbb{R}$, трижды дифференцируемая в единице, такая что $f(1) = 0$ и $f''(1) > 0$, удовлетворяющая условию*

$$\bigl(f(t) - f'(1)(t-1)\bigr)\left(1 - \frac{f'''(1)}{3f''(1)}(t-1)\right) \geqslant \frac{f''(1)}{2}(t-1)^2 \tag{75}$$

для любого $t \in (0,\infty)$. Тогда для любых $R_X, P_X \in \mathcal{P}_\mathcal{X}$ справедливо

$$D_f(R_X \,\|\, P_X) \geqslant 2f''(1)\,\|R_X - P_X\|_{\mathrm{TV}}^2\,.$$

Более того, в этом неравенстве участвует оптимальная константа в том смысле, что

$$\inf_{\substack{R_X,P_X\in\mathcal{P}_\mathcal{X}\\ R_X\neq P_X}} \frac{D_f(R_X\,\|\,P_X)}{\|R_X-P_X\|_{\mathrm{TV}}^2} = 2f''(1).$$

Заметим, что функция $f(t)=t\log(t)$ удовлетворяет условиям леммы 4, причем $f''(1)=1$, как показано в Приложении E; отсюда вытекает стандартное неравенство Пинскера (69). Поскольку условие (75) может быть нелегко проверить для других f-дивергенций, в [79, следствие 4] приведены достаточные условия для неравенства (75). (Эти условия можно проверить и получить вариант неравенства Пинскера для, например, *дивергенций Реньи* порядка $\alpha\in(0,1)$ [79, следствие 6].) В следующей лемме с помощью леммы 4 устанавливается нижняя граница на определенные f-дивергенции через χ^2-дивергенцию, что соответствует лемме 2 (или, точнее, неравенству (74), так как оно вытекает из стандартного неравенства Пинскера).

Лемма 5 (нижняя граница f-дивергенции). *Пусть задана выпуклая функция $f\colon(0,\infty)\to\mathbb{R}$, трижды дифференцируемая в единице, такая что $f(1)=0$ и $f''(1)>0$, удовлетворяющая условию (75) для любого $t\in(0,\infty)$. Тогда для любых вероятностных мер $R_X\in\mathcal{P}_\mathcal{X}$ и $P_X\in\mathcal{P}_\mathcal{X}^\circ$ справедливо неравенство*

$$D_f(R_X\,\|\,P_X)\geqslant f''(1)\min_{x\in\mathcal{X}}P_X(x)\chi^2(R_X\,\|\,P_X).$$

Доказательство. Будем следовать доказательству леммы 2, внося необходимые изменения. Предположим без ограничения общности, что $R_X\neq P_X$. Обобщенное неравенство Пинскера для f-дивергенций из леммы 4 дает

$$D_f(R_X\,\|\,P_X)\geqslant\frac{f''(1)}{2}\|R_X-P_X\|_1^2,$$

используя характеризацию расстояния по вариации через ℓ^1-норму, данную в (2). Требуемый результат получается применением (72) к этому неравенству. ▲

Заметим, что если в лемме 5 положить $f(t)=t\log(t)$, получим неравенство (74).

Наконец, приведем верхнюю границу на некоторые f-дивергенции через χ^2-дивергенцию, аналогичную лемме 3. Эта верхняя граница была доказана в [8, лемма A.2] в предположении, что f дифференцируема, но как мы увидим ниже, нужно лишь проверить дифференцируемость в единице. (Было бы поучительно еще раз посмотреть на доказательство леммы 3, чтобы увидеть, как нижеследующее доказательство обобщает ее на f-дивергенции.)

Лемма 6 (верхняя граница f-дивергенции [8, лемма A.2]). *Пусть дана выпуклая функция $f\colon(0,\infty)\to\mathbb{R}$, дифференцируемая в единице, такая что $f(1)=0$ и $f(0)<\infty$, и пусть разностное отношение $g\colon(0,\infty)\to\mathbb{R}$, определяемое как*

$$g(t)=\frac{f(t)-f(0)}{t},$$

выпукло вверх. Тогда для любых вероятностных мер $R_X,P_X\in\mathcal{P}_\mathcal{X}$ справедливо неравенство

$$D_f(R_X\,\|\,P_X)\leqslant(f'(1)+f(0))\chi^2(R_X\,\|\,P_X).$$

Доказательство. Для полноты изложения приведем доказательство из [8]. Как и в доказательстве леммы 3, можно предполагать без ограничения общности, что $P_X\in\mathcal{P}_\mathcal{X}^\circ$, так что все рассматриваемые далее величины конечны. Тогда имеет

место следующая цепочка равенств и неравенств:

$$
\begin{aligned}
&D_f(R_X \,\|\, P_X) = \sum_{x\in\mathcal{X}} P_X(x) f\Big(\frac{R_X(x)}{P_X(x)}\Big) = \\
&= f(0) + \sum_{x\in\mathcal{X}} R_X(x) g\Big(\frac{R_X(x)}{P_X(x)}\Big) \leqslant f(0) + g\bigg(\sum_{x\in\mathcal{X}} \frac{R_X(x)^2}{P_X(x)}\bigg) = \\
&= f(0) + g\big(1+\chi^2(R_X \,\|\, P_X)\big) \leqslant \qquad (76)\\
&\leqslant f(0) + g(1) + g'(1)\chi^2(R_X \,\|\, P_X) = (f'(1)+f(0))\chi^2(R_X \,\|\, P_X),
\end{aligned}
$$

где во втором равенстве используется соглашение $0g(0)=0$, третье следует из неравенства Йенсена, так как функция $g\colon (0,\infty)\to\mathbb{R}$ выпукла вверх, пятое также вытекает из выпуклости $g\colon (0,\infty)\to\mathbb{R}$, как показано в [100, п. 3.1.3], а последнее неравенство справедливо, поскольку $g(1) = -f(0)$ (так как $f(1)=0$) и

$$
\begin{aligned}
&g'(1) = \lim_{\delta\to 0} \frac{g(1+\delta)+f(0)}{\delta} = \lim_{\delta\to 0} \frac{f(1+\delta)+\delta f(0)}{\delta(1+\delta)} = \\
&= \bigg(\lim_{\delta\to 0}\frac{1}{1+\delta}\bigg)\bigg(f(0) + \lim_{\delta\to 0}\frac{f(1+\delta)}{\delta}\bigg) = f'(1)+f(0),
\end{aligned}
$$

что завершает доказательство. ▲

Заметим, что выражение (76) является аналогом более точной (нелинейной) границы из леммы 3. Кроме того, отметим, что в лемме 6 можно использовать функцию $g(t) = \dfrac{f(t)}{t}$ (в предположении ее выпуклости вверх) вместо разностного отношения. Доказательство останется тем же, но с использованием константы $f'(1)$ вместо $f'(1)+f(0)$. Однако мы для доказательства леммы 6 выбрали разностное отношение ввиду свойства аффинной инвариантности f-дивергенций (см. п. 2.1). Нетрудно проверить, что величина $f'(1)+f(0)$ инвариантна относительно подходящих аффинных сдвигов, что неверно для $f'(1)$. Также отметим, что константа $f''(1)$ в лемме 5 инвариантна относительно соответствующих аффинных сдвигов.

4.2. Доказательства теорем 3 и 4. Напомним (см. начало п. 3.1), что мы рассматриваем совместную вероятностную меру $P_{X,Y}$, состоящую из $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ и $P_{Y|X} = W \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$, где $P_Y = P_X W \in \mathcal{P}_{\mathcal{Y}}^{\circ}$. Теперь с помощью лемм 2 и 3 из п. 4.1 мы можем доказать теорему 4.

Доказательство теоремы 4. Для любой вероятностной меры $R_X \in \mathcal{P}_{\mathcal{X}}$, такой что $R_X \neq P_X$, имеем

$$
\frac{D(R_X W \,\|\, P_Y)}{D(R_X \,\|\, P_X)} \leqslant \frac{2\chi^2(R_X W \,\|\, P_Y)}{\varphi\Big(\max\limits_{A\subseteq\mathcal{X}} \min\{P_X(A), 1-P_X(A)\}\Big) \min\limits_{x\in\mathcal{X}} P_X(x)\chi^2(R_X \,\|\, P_X)}
$$

в силу лемм 2 и 3. Переходя к супремуму по всем $R_X \neq P_X$ в обоих частях неравенства, получаем

$$
\eta_{\mathrm{KL}}(P_X, P_{Y|X}) \leqslant \frac{2\eta_{\chi^2}(P_X, P_{Y|X})}{\varphi\Big(\max\limits_{A\subseteq\mathcal{X}} \min\{P_X(A), 1-P_X(A)\}\Big) \min\limits_{x\in\mathcal{X}} P_X(x)}
$$

с учетом (17) и (22), что и завершает доказательство. ▲

Уместно сделать несколько замечаний. Во-первых, если в этом доказательстве применить (74) вместо леммы 2, получим доказательство следствия 1.

Во-вторых, при доказательстве следствия 1 в [1, теорема 10] была также доказана следующая более слабая граница сверху на $\eta_{\mathrm{KL}}(P_X, P_{Y|X})$ (см. [1, теорема 9]):

$$\eta_{\mathrm{KL}}(P_X, P_{Y|X}) \leqslant \frac{2}{\min\limits_{x\in\mathcal{X}} P_X(x)}\eta_{\chi^2}(P_X, P_{Y|X}), \tag{77}$$

также независимо полученная в [8, формула III.19]. Наше доказательство соотношения (77) в [1, теорема 9] использовало следующий в два раза худший вариант неравенства (74) (см. [1, лемма 6]):

$$D(S_X \,\|\, Q_X) \geqslant \frac{\min\limits_{x\in\mathcal{X}} Q_X(x)}{2}\chi^2(S_X \,\|\, Q_X) \tag{78}$$

для всех $S_X, Q_X \in \mathcal{P}_{\mathcal{X}}$. Это неравенство вытекает из доказательства леммы 2 с использованием границы $\|S_X - Q_X\|_\infty \leqslant \|S_X - Q_X\|_1$ (не учитывающей тот факт, что $(S_X - Q_X)\mathbf{1} = 0$) вместо (73) и последующим применением оценки (70) к получающейся нижней границе на КЛ-дивергенцию. В качестве альтернативы для полноты изложения в Приложении F мы даем доказательство формулы (78) через *дивергенции Брегмана* (см. [1, лемма 6]). То, что границу (78) можно улучшить вдвое до границы (74), было также указано в [34, замечание 33], где отмечалось, что наш результат [1, теорема 9] (приведенный в (77)) можно улучшить вдвое, используя (74) вместо (78). По-видимому, авторы работы [34] не заметили наш результат [1, теорема 10] (представленный в следствии 1), в котором в точности было приведено это улучшение вдвое.

И наконец, отметим, что в [8, п. III-D] также приведены верхние границы на $\eta_{\mathrm{KL}}(P_X, P_{Y|X})$, использующие функцию $\varphi\colon [0, 1/2] \to \mathbb{R}$, что вытекает из улучшенного неравенства Пинскера в [91]. Однако эти верхние границы выражены не через $\eta_{\chi^2}(P_X, P_{Y|X})$.

Теперь, объединяя результаты лемм 5 и 6 из п. 4.1, докажем теорему 3.

Доказательство теоремы 3. Условия теоремы 3 включают в себя все условия лемм 5 и 6. Следовательно, применяя леммы 5 и 6, для любой вероятностной меры $R_X \in \mathcal{P}_{\mathcal{X}}$, такой что $R_X \neq P_X$, получаем

$$\frac{D_f(R_X W \,\|\, P_Y)}{D_f(R_X \,\|\, P_X)} \leqslant \frac{(f'(1) + f(0))\chi^2(R_X W \,\|\, P_Y)}{f''(1)\min\limits_{x\in\mathcal{X}} P_X(x)\chi^2(R_X \,\|\, P_X)}.$$

Переходя к супремуму по всем $R_X \neq P_X$ в обеих частях этого неравенства, получаем

$$\eta_f(P_X, P_{Y|X}) \leqslant \frac{f'(1) + f(0)}{f''(1)\min\limits_{x\in\mathcal{X}} P_X(x)}\eta_{\chi^2}(P_X, P_{Y|X}),$$

где использовались определение 2 и соотношение (22), что и завершает доказательство. ▲

Отметим, что в [8, теорема III.4] приведена альтернативная линейная граница сверху на $\eta_f(P_X, P_{Y|X})$ через $\eta_{\chi^2}(P_X, P_{Y|X})$. Пусть $f\colon (0,\infty) \to \mathbb{R}$ – дважды дифференцируемая выпуклая функция, такая что $f(1) = 0$ и $f(0) < \infty$, строго выпуклая в единице и имеющая невозрастающую вторую производную. Если к тому же предположить, что разностное отношение $t \mapsto \dfrac{f(t) - f(0)}{t}$ выпукло вверх, то справедлива следующая граница [8, теорема III.4]:

$$\eta_f(P_X, P_{Y|X}) \leqslant \frac{2(f'(1) + f(0))}{f''(1/p_\star)}\eta_{\chi^2}(P_X, P_{Y|X}), \tag{79}$$

где $p_\star = \min\limits_{x\in\mathcal{X}} P_X(x)$. Следовательно, если функция f к тому же трижды дифференцируема в единице, имеет $f''(1) > 0$ и удовлетворяет условию (75) для любого $t \in (0,\infty)$, то можно усилить верхнюю границу теоремы 3 до

$$\eta_f(P_X, P_{Y|X}) \leqslant \min\left\{\frac{f'(1)+f(0)}{f''(1)p_\star}, \frac{2(f'(1)+f(0))}{f''(1/p_\star)}\right\}\eta_{\chi^2}(P_X, P_{Y|X}). \tag{80}$$

Заметим, что наша граница из теоремы 3 точнее границы (79) тогда и только тогда, когда

$$\frac{2(f'(1)+f(0))}{f''(1/p_\star)} \geqslant \frac{f'(1)+f(0)}{f''(1)p_\star} \quad\iff\quad 2f''(1)p_\star \geqslant f''(1/p_\star). \tag{81}$$

Одной из функций, удовлетворяющих условиям теоремы 3 и неравенствам (79) и (81), является $f(t) = t\log(t)$. Это и дает улучшение в следствии 1 (получаемом из теоремы 3) по сравнению с неравенством (77) (получаемым из [8, теорема III.4]).

В качестве другого примера рассмотрим функцию

$$f(t) = \frac{t^\alpha - 1}{\alpha - 1}, \quad \alpha \in (0,2]\setminus\{1\},$$

задающую дивергенцию Хеллингера порядка α (см. п. 2.1). Непосредственными вычислениями нетрудно проверить, что эта функция удовлетворяет условиям теоремы 3 и неравенству (79) (см. [79, следствие 6], [8, п. III-B, с. 3362]). В этом случае наша граница из теоремы 3 точнее, чем (79), для всех дивергенций Хеллингера порядка α, удовлетворяющих условию (81), т.е.

$$2f''(1)p_\star = 2\alpha p_\star \geqslant \alpha(1/p_\star)^{\alpha-2} = f''(1/p_\star) \iff p_\star^{\alpha-1} \geqslant \frac{1}{2},$$

или, что равносильно, $0 < \alpha \leqslant 1 + (\log(2)/\log(1/p_\star))$ (где $\alpha = 1$ соответствует КЛ-дивергенции – см. п. 2.1).

4.3. Эргодичность цепей Маркова. В этом пункте мы выведем из следствия 1 результат, иллюстрирующий одно из применений верхних границ на коэффициенты сжатия совместных распределений через $\eta_{\chi^2}(P_X, P_{Y|X})$. Рассмотрим матрицу *примитивного* марковского ядра $W \in \mathcal{P}_{\mathcal{X}|\mathcal{X}}$ на пространстве состояний $\mathcal{X}$, задающую *неприводимую* и *апериодическую* (однородную по времени) дискретную цепь Маркова с единственной стационарной вероятностной мерой $P_X \in \mathcal{P}_\mathcal{X}^\circ$, такой что $P_X W = P_X$ (см. [29, п. 1.3]). Для простоты предположим также, что W *обратима* (т.е. имеют место уравнения детального баланса $P_X(x)[W]_{x,y} = P_X(y)[W]_{y,x}$ для всех $x, y \in \mathcal{X}$ [29, п. 1.6]). Это означает, что матрица W самосопряжена относительно взвешенного скалярного произведения, задаваемого P_X, и все ее собственные значения $1 = \lambda_1(W) > \lambda_2(W) \geqslant \ldots \geqslant \lambda_{|\mathcal{X}|}(W) > -1$ вещественны. Через $\mu(W) \triangleq \max\{|\lambda_2(W)|, |\lambda_{|\mathcal{X}|}(W)|\} \in [0,1)$ обозначается модуль второго по абсолютной величине собственного значения W (см. п. 2.3).

Так как эта цепь Маркова *эргодична*, то $\lim\limits_{n\to\infty} R_X W^n = P_X$ для всех $R_X \in \mathcal{P}_\mathcal{X}$ [29, теорема 4.9]. Отсюда $\lim\limits_{n\to\infty} D(R_X W^n \,\|\, P_X) = 0$ из непрерывности КЛ-дивергенции [38, предложение 3.1]. Оценим скорость, с которой убывает это "расстояние до стационарности" (измеряемое КЛ-дивергенцией). Наивный подход состоит в рекурсивном применении СНОД (15) для КЛ-дивергенции, что дает

$$D(R_X W^n \,\|\, P_X) \leqslant \eta_{\mathrm{KL}}(P_X, W)^n D(R_X \,\|\, P_X) \tag{82}$$

для всех $R_X \in \mathcal{P}_{\mathcal{X}}$ и всех $n \in \mathbb{N}$. С учетом (17) отсюда следует, что

$$\limsup_{n\to\infty} \eta_{\mathrm{KL}}(P_X, W^n)^{\frac{1}{n}} \leqslant \eta_{\mathrm{KL}}(P_X, W), \tag{83}$$

что оказывается довольно слабой границей на скорость в общем случае.

При больших n, поскольку $R_X W^n$ близка к P_X, интуитивно следует ожидать, что $D(R_X W^n \,\|\, P_X)$ ведет себя как χ^2-дивергенция (см. (13) в п. 2.1), и это наводит на мысль, что $\eta_{\mathrm{KL}}(P_X, W^n)$ должна по порядку величины совпадать с $\eta_{\chi^2}(P_X, W)^n$. Это интуитивная догадка строго доказана в [17, § 6]. Действительно, когда $\mu(W)$ строго больше модуля третьего по абсолютной величине собственного значения W, из [17, следствие 6.2] вытекает, что

$$\lim_{n\to\infty} \frac{D(R_X W^n \,\|\, P_X)}{D(R_X W^{n-1} \,\|\, P_X)} \leqslant \mu(W)^2 \tag{84}$$

для всех $R_X \in \mathcal{P}_{\mathcal{X}}$, таких что знаменатель всегда положителен. (Этот предел равен либо 0, либо $\mu(W)^2$.) Поэтому, используя сходимость по Чезаро и масштабирование, получаем

$$\lim_{n\to\infty} \frac{D(R_X W^n \,\|\, P_X)}{D(R_X W^{n-1} \,\|\, P_X)} = \lim_{n\to\infty} \left(\frac{D(R_X W^n \,\|\, P_X)}{D(R_X \,\|\, P_X)} \right)^{\frac{1}{n}} \leqslant \mu(W)^2, \tag{85}$$

что позволяет предположить, что $\limsup_{n\to\infty} \eta_{\mathrm{KL}}(P_X, W^n)^{\frac{1}{n}} \leqslant \mu(W)^2$. Следующий результат показывает, что это неравенство на самом деле точное.

Предложение 7 (скорость сходимости). *Для любой неприводимой апериодической обратимой цепи Маркова с ядром $W \in \mathcal{P}_{\mathcal{X}|\mathcal{X}}$ и стационарной вероятностной мерой $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$*

$$\lim_{n\to\infty} \eta_{\mathrm{KL}}(P_X, W^n)^{\frac{1}{n}} = \eta_{\chi^2}(P_X, W) = \mu(W)^2.$$

Доказательство. Так как W обратима, а P_X – ее инвариантная мера, то МДП

$$B = \operatorname{diag}\big(\sqrt{P_X}\,\big) W \operatorname{diag}\big(\sqrt{P_X}\,\big)^{-1}$$

симметрична и подобна W (см. определение (24)). Следовательно, W и B имеют одинаковые собственные значения, и $\mu(W)$ является вторым по величине сингулярным числом для B. Применяя предложение 2 и (23), получаем $\eta_{\chi^2}(P_X, W) = \mu(W)^2$, что доказывает второе равенство.

Аналогично, $\eta_{\chi^2}(P_X, W^n) = \mu(W^n)^2$, так как W^n обратима для любого $n \geqslant 1$. Отсюда

$$\eta_{\chi^2}(P_X, W^n) = \mu(W^n)^2 = \mu(W)^{2n} = \eta_{\chi^2}(P_X, W)^n, \tag{86}$$

где второе равенство справедливо, поскольку собственные значения W^n являются n-ми степенями собственных значений W. С учетом (86), п. 7 предложения 3 и следствия 1 получаем

$$\eta_{\chi^2}(P_X, W)^n \leqslant \eta_{\mathrm{KL}}(P_X, W^n) \leqslant \frac{\eta_{\chi^2}(P_X, W)^n}{\min\limits_{x\in\mathcal{X}} P_X(x)}.$$

Остается извлечь корни n-й степени и перейти к пределу при $n \to \infty$. ▲

Предложение 7 описывает хорошо понимаемое явление, что $D(R_X W^n \,\|\, P_X)$ убывает со скоростью, задаваемой величиной $\mu(W)^2 = \eta_{\chi^2}(P_X, W)$. В более широком

смысле оно показывает, что границы следствия 1 и теорем 3 и 4 полезны в режиме, когда случайные величины X и Y слабо зависимы (например, X – начальное состояние эргодической обратимой цепи Маркова, а Y – ее состояние после большого числа шагов). В таком режиме эти границы довольно точны и превосходят границу через η_{TV} в п. 7 предложения 5.

4.4. Тензоризация границ между коэффициентами сжатия. В отсутствии слабой зависимости верхние границы следствия 1 и теорем 3 и 4 могут быть слабыми. На самом деле, их можно считать сколь угодно слабыми, поскольку константы в этих границах не тензоризуются в отличие от коэффициентов сжатия (последнее показано в п. 5 предложения 3). Например, если дана $P_{X,Y}$ с $X \sim \mathrm{Bernoulli}(1/2)$, то константа в верхней границе следствия 1 равна $1/\min\limits_{x\in\{0,1\}} P_X(x) = 2$. Если вместо этого рассмотреть последовательность независимых и одинаково распределенных (н.о.р.) согласно $P_{X,Y}$ пар $(X_1,Y_1),\dots,(X_n,Y_n)$, то константа в верхней границе следствия 1 равна $1/\min\limits_{x_1^n\in\{0,1\}^n} P_{X_1^n}(x_1^n) = 2^n$. Но поскольку $\eta_{\mathrm{KL}}(P_{X_1^n}, P_{Y_1^n|X_1^n}) = \eta_{\mathrm{KL}}(P_X, P_{Y|X})$ и $\eta_{\chi^2}(P_{X_1^n}, P_{Y_1^n|X_1^n}) = \eta_{\chi^2}(P_X, P_{Y|X})$ по свойству тензоризации из п. 5 предложения 3, то константа 2^n становится сколь угодно плохой с ростом n. Эту атаку на следствие 1 с помощью н.о.р. пар частично позволяет отразить

Следствие 2 (тензоризованная КЛ-граница на коэффициенты сжатия). *Если пары $(X_1,Y_1),\dots,(X_n,Y_n)$ н.о.р. с совместной вероятностной мерой $P_{X,Y}$, такой что $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ и $P_Y \in \mathcal{P}_{\mathcal{Y}}^{\circ}$, то*

$$\eta_{\mathrm{KL}}\big(P_{X_1^n}, P_{Y_1^n|X_1^n}\big) \leqslant \frac{\eta_{\chi^2}\big(P_{X_1^n}, P_{Y_1^n|X_1^n}\big)}{\min\limits_{x\in\mathcal{X}} P_X(x)}.$$

Доказательство. Это неравенство тривиальным образом вытекает из следствия 1 и свойства тензоризации в п. 5 предложения 3. ▲

В контексте произведений распределений это следствие позволяет использовать в верхней границе следствия 1 более точный множитель $1/\min\limits_{x\in\mathcal{X}} P_X(x)$ вместо $1/\min\limits_{x_1^n\in\mathcal{X}^n} P_{X_1^n}(x_1^n) = \Big(1/\min\limits_{x\in\mathcal{X}} P_X(x)\Big)^n$. Аналогичные уточнения в этом контексте можно также сделать для констант в теоремах 3 и 4. Таким образом, тензоризация может улучшить верхние границы в следствии 1 и теоремах 3, 4.

§ 5. Доказательство эквивалентности между гауссовскими коэффициентами сжатия

В этом параграфе мы докажем теорему 5. Напомним (см. п. 3.3), что мы рассматриваем совместно гауссовские плотности распределения $P_{X,Y}$, заданные в (52), с маргинальными плотностями распределения на входе $P_X = \mathcal{N}(0,\sigma_X^2)$ и условными плотностями распределения $P_{Y|X} = \big\{P_{Y|X=x} = \mathcal{N}(x,\sigma_W^2) : x \in \mathbb{R}\big\}$, где $\sigma_X^2, \sigma_W^2 > 0$. Пусть $\mathcal{T}$ – множество всех функций распределения на $\mathbb{R}$ с ограниченным существенным носителем. Таким образом, $R_X \in \mathcal{T}$ тогда и только тогда, когда существует $C > 0$, такая что $R_X = R_X \mathbb{1}_{[-C,C]}$ почти всюду по мере Лебега, где через $\mathbb{1}_{[-C,C]}\colon \mathbb{R} \to \{0,1\}$ обозначена индикаторная функция на $[-C,C]$:

$$\mathbb{1}_{[-C,C]}(x) = \begin{cases} 1, & x \in [-C,C], \\ 0 & \text{в противном случае.} \end{cases} \tag{87}$$

Вначале докажем следующую полезную лемму.

Лемма 7 (характеризация η_{KL} через функции с ограниченным носителем). *Супремум в (55) можно ограничить на плотности распределения из множест-*

ва $\mathcal{T}$:

$$\eta_{\mathrm{KL}}(P_X, P_{Y|X}) = \sup_{\substack{R_X \in \mathcal{T} \\ D(R_X \| P_X) < +\infty}} \frac{D(R_Y \| P_Y)}{D(R_X \| P_X)},$$

где $R_Y = R_X * P_W$ *для каждой плотности распределения* R_X, $P_W = \mathcal{N}(0, \sigma_W^2)$, *а ограничение* $D(R_X \| P_X) > 0$ *автоматически выполнено для любой* $R_X \in \mathcal{T}$, *поскольку* $P_X = \mathcal{N}(0, \sigma_X^2)$.

Доказательство. Рассмотрим произвольную плотность распределения R_X, такую что $0 < D(R_X \| P_X) < +\infty$, и зададим соответствующую последовательность функций плотности распределения

$$R_X^{(n)} = R_X \mathbb{1}_{[-n,n]} / C_n \in \mathcal{T},$$

где $C_n = \mathbf{E}_{R_X}[\mathbb{1}_{[-n,n]}(X)]$, индексы $n \in \mathbb{N}$ достаточно большие, так что $C_n > 0$, и $\lim\limits_{n\to\infty} C_n = 1$. Заметим, что

$$D\big(R_X^{(n)} \| P_X\big) = \frac{1}{C_n} \mathbf{E}_{R_X}\left[\mathbb{1}_{[-n,n]}(X) \log\left(\frac{R_X(X)}{P_X(X)}\right)\right] - \log(C_n).$$

Очевидно, что $\lim\limits_{n\to\infty} \mathbb{1}_{[-n,n]} \log(R_X/P_X) = \log(R_X/P_X)$ поточечно R_X-п.н., а также $\big|\mathbb{1}_{[-n,n]} \log(R_X/P_X)\big| \leqslant |\log(R_X/P_X)|$ поточечно R_X-п.н., так что выполнено неравенство $\mathbf{E}_{R_X}[|\log(R_X(X)/P_X(X))|] < +\infty$ (где конечность следует из того факта, что $D(R_X \| P_X) < +\infty$). Отсюда по теореме о мажорируемой сходимости получаем

$$\lim_{n\to\infty} D\big(R_X^{(n)} \| P_X\big) = D(R_X \| P_X). \tag{88}$$

Далее, пусть $R_Y^{(n)} = R_X^{(n)} * P_W$, так что для любого $y \in \mathbb{R}$

$$R_Y(y) - C_n R_Y^{(n)}(y) = \mathbf{E}_{R_X}[\mathbb{1}_{\mathbb{R}\setminus[-n,n]}(X) P_W(y - X)].$$

Так как для всех $x, y \in \mathbb{R}$ имеем $\lim\limits_{n\to\infty} \mathbb{1}_{\mathbb{R}\setminus[-n,n]}(x) P_W(y-x) = 0$, а также $0 \leqslant \mathbb{1}_{\mathbb{R}\setminus[-n,n]}(x) P_W(y-x) \leqslant P_W(y-x)$, так что $\mathbf{E}_{R_X}[P_W(y - X)] = R_Y(y) < +\infty$, применяя теорему о мажорируемой сходимости, получаем поточечную сходимость функций плотности распределения $\{R_Y^{(n)}\}$:

$$\forall y \in \mathbb{R}, \quad \lim_{n\to\infty} C_n R_Y^{(n)}(y) = \lim_{n\to\infty} R_Y^{(n)}(y) = R_Y(y).$$

Отсюда следует, что $R_Y^{(n)}$ слабо сходится к R_Y при $n \to \infty$ по лемме Шеффе. Следовательно, согласно слабой полунепрерывности КЛ-дивергенции снизу [38, теорема 3.6, п. 3.5] имеем

$$\liminf_{n\to\infty} D\big(R_Y^{(n)} \| P_Y\big) \geqslant D(R_Y \| P_Y). \tag{89}$$

Объединяя (88) и (89), получаем

$$\liminf_{n\to\infty} \frac{D\big(R_Y^{(n)} \| P_Y\big)}{D\big(R_X^{(n)} \| P_X\big)} \geqslant \frac{D(R_Y \| P_Y)}{D(R_X \| P_X)}. \tag{90}$$

Для завершения доказательства применим рассуждения, основанные на "диагональном методе". Пусть $\{R_{X,m} : m \in \mathbb{N}\}$ – последовательность функций плотности

распределения, таких что $0 < D(R_{X,m} \,\|\, P_X) < +\infty$ для всех $m \in \mathbb{N}$, достигающая супремума в (55):

$$\lim_{m\to\infty} \frac{D(R_{Y,m} \,\|\, P_Y)}{D(R_{X,m} \,\|\, P_X)} = \eta_{\mathrm{KL}}(P_X, P_{Y|X}),$$

где $R_{Y,m} = R_{X,m} * P_W$. Тогда в силу (90) можно построить последовательность $\big\{R_{X,m}^{(n(m))} \in \mathcal{T} :\ m \in \mathbb{N}\big\}$, где каждое $n(m)$ выбирается так, чтобы для любого $m \in \mathbb{N}$

$$\frac{D\big(R_{Y,m}^{(n(m))} \,\|\, P_Y\big)}{D\big(R_{X,m}^{(n(m))} \,\|\, P_X\big)} \geqslant \frac{D(R_{Y,m} \,\|\, P_Y)}{D(R_{X,m} \,\|\, P_X)} - \frac{1}{2^m},$$

где $R_{Y,m}^{(n(m))} = R_{X,m}^{(n(m))} * P_W$. Устремляя $m \to \infty$, получаем

$$\liminf_{m\to\infty} \frac{D\big(R_{Y,m}^{(n(m))} \,\|\, P_Y\big)}{D\big(R_{X,m}^{(n(m))} \,\|\, P_X\big)} \geqslant \eta_{\mathrm{KL}}(P_X, P_{Y|X}).$$

Поскольку супремум в (55) берется по всем плотностям распределения (которые заведомо включают в себя все плотности распределения из $\mathcal{T}$), это неравенство на самом деле является равенством, что и завершает доказательство. ▲

Теперь докажем теорему 5, пользуясь леммой 7, которая гарантирует, что все дифференциальные энтропии, участвующие в нижеследующих рассуждениях, корректно определены и конечны.

Доказательство теоремы 5. Вначале заметим, что

$$\eta_{\chi^2}(P_X, P_{Y|X}) = \rho(X;Y)^2 = \frac{\mathrm{Cov}(X,Y)^2}{\mathrm{Var}(X)\,\mathrm{Var}(Y)} = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_W^2},$$

где первое равенство – это в точности соотношение (23) (справедливое для общих случайных величин [11]), второе вытекает из седьмой аксиомы Реньи, гласящей, что $\rho(X;Y)$ равно абсолютной величине коэффициента корреляции Пирсона совместно гауссовских величин X и Y [12], а последнее получается непосредственными вычислениями.

Теперь докажем, что для любого $p \geqslant \sigma_X^2$

$$\eta_{\mathrm{KL}}(P_X, P_{Y|X}) \geqslant \eta_{\mathrm{KL}}^{(p)}(P_X, P_{Y|X}) \geqslant \eta_{\chi^2}(P_X, P_{Y|X}).$$

Первое неравенство очевидно ввиду (55) и (57). Для второго неравенства положим $R_X = \mathcal{N}(\sqrt{\delta}, \sigma_X^2 - \delta)$ и $R_Y = R_X * P_W = \mathcal{N}(\sqrt{\delta}, \sigma_X^2 + \sigma_W^2 - \delta)$ для любого $\delta > 0$. Тогда получаем

$$\lim_{\delta\to 0^+} \frac{D(R_Y \,\|\, P_Y)}{D(R_X \,\|\, P_X)} = \lim_{\delta\to 0^+} \frac{\log\left(\frac{\sigma_X^2 + \sigma_W^2}{\sigma_X^2 + \sigma_W^2 - \delta}\right)}{\log\left(\frac{\sigma_X^2}{\sigma_X^2 - \delta}\right)} = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_W^2},$$

где первое равенство получается непосредственными вычислениями (см. [38, формула (1.17)]), а второе следует из правила Лопиталя. Так как $\mathbf{E}_{R_X}[X^2] = \sigma_X^2$ для любого $\delta > 0$, то $\eta_{\mathrm{KL}}^{(p)}(P_X, P_{Y|X}) \geqslant \sigma_X^2/(\sigma_X^2 + \sigma_W^2)$ для любого $p \geqslant \sigma_X^2$.

Поэтому достаточно доказать, что $\eta_{\mathrm{KL}}(P_X, P_{Y|X}) \leqslant \sigma_X^2/(\sigma_X^2 + \sigma_W^2)$. С учетом леммы 7 это равносильно тому, что для любой плотности распределения $R_X \in \mathcal{T}$,

такой что $D(R_X \,\|\, P_X) < +\infty$,

$$\frac{D(R_Y \,\|\, P_Y)}{D(R_X \,\|\, P_X)} \leqslant \frac{\sigma_X^2}{\sigma_X^2 + \sigma_W^2}. \tag{91}$$

Для любой плотности распределения R_X определим ее *дифференциальную энтропию* $h(R_X) \triangleq -\mathbf{E}_{R_X}[\log(R_X(X))]$. Чтобы проверить, что эти величины корректно определены и конечны для $R_X \in \mathcal{T}$, используем рассуждение из [101, лемма 8.3.1, теорема 8.3.3]. Заметим, что для всех $x \in \operatorname{ess\,supp}(R_X)$

$$\log(R_X(x)) = \log\left(\frac{R_X(x)}{P_X(x)}\right) - \frac{1}{2}\log(2\pi\sigma_X^2) - \frac{x^2}{2\sigma_X^2},$$

где через $\operatorname{ess\,supp}(\cdot)$ обозначен существенный носитель измеримой по Борелю функции относительно меры Лебега. Поскольку величина $D(R_X \,\|\, P_X)$ должна быть конечна в формуле (55), а $X^2 \geqslant 0$, можно перейти к математическим ожиданиям относительно R_X:

$$-h(R_X) = D(R_X \,\|\, P_X) - \frac{1}{2}\log(2\pi\sigma_X^2) - \frac{\mathbf{E}_{R_X}[X^2]}{2\sigma_X^2}, \tag{92}$$

откуда следует, что $h(R_X)$ существует всегда, она конечна, если $\mathbf{E}_{R_X}[X^2] < +\infty$, и $h(R_X) = +\infty$, когда $\mathbf{E}_{R_X}[X^2] = +\infty$. Кроме того, если функция плотности распределения $R_X \in \mathcal{T}$ имеет ограниченный существенный носитель, то $\mathbf{E}_{R_X}[X^2] < +\infty$, и $h(R_X)$ определена и конечна.

Пусть $R_X \in \mathcal{T}$ и $R_Y = R_X * P_W$ имеют вторые моменты $\mathbf{E}_{R_X}[X^2] = \sigma_X^2 + q > 0$ и $\mathbf{E}_{R_Y}[Y^2] = \sigma_X^2 + \sigma_W^2 + q > 0$ для некоторого $q > -\sigma_X^2$. Используя (92) и [38, формула (1.20)], получаем

$$D(R_X \,\|\, P_X) = \frac{1}{2}\log(2\pi\sigma_X^2) + \frac{\sigma_X^2 + q}{2\sigma_X^2} - h(R_X) = h(P_X) - h(R_X) + \frac{q}{2\sigma_X^2},$$
$$D(R_Y \,\|\, P_Y) = h(P_Y) - h(R_Y) + \frac{q}{2(\sigma_X^2 + \sigma_W^2)},$$

где $h(R_Y)$ существует и конечна, так как $\mathbf{E}_{R_Y}[Y^2]$ конечно (как показано выше с помощью (92)). Значит, достаточно доказать, что

$$h(P_Y) - h(R_Y) \leqslant \frac{\sigma_X^2}{\sigma_X^2 + \sigma_W^2}(h(P_X) - h(R_X)), \tag{93}$$

что равносильно (91). Перепишем (93) в виде

$$\left(e^{2h(P_Y) - 2h(R_Y)}\right)^{\sigma_X^2 + \sigma_W^2} \leqslant \left(e^{2h(P_X) - 2h(R_X)}\right)^{\sigma_X^2} \iff$$
$$\iff \quad \left(\frac{\frac{1}{2\pi e}e^{2h(P_Y)}}{\frac{1}{2\pi e}e^{2h(R_Y)}}\right)^{\sigma_X^2 + \sigma_W^2} \leqslant \left(\frac{\frac{1}{2\pi e}e^{2h(P_X)}}{\frac{1}{2\pi e}e^{2h(R_X)}}\right)^{\sigma_X^2} \iff$$
$$\iff \quad \left(\frac{N(P_Y)}{N(R_Y)}\right)^{\sigma_X^2 + \sigma_W^2} \leqslant \left(\frac{N(P_X)}{N(R_X)}\right)^{\sigma_X^2},$$

где для любой плотности распределения Q_X, такой что $h(Q_X)$ существует, мы определяем *энтропийную мощность* Q_X как $N(Q_X) \triangleq e^{2h(Q_X)}/(2\pi e)$ [102, п. III-A]. Для $P_X = \mathcal{N}(0, \sigma_X^2)$, $P_W = \mathcal{N}(0, \sigma_W^2)$ и $P_Y = P_X * P_W = \mathcal{N}(0, \sigma_X^2 + \sigma_W^2)$ энтропийные

мощности равны $N(P_X) = \sigma_X^2$, $N(P_W) = \sigma_W^2$ и $N(P_Y) = \sigma_X^2 + \sigma_W^2$ соответственно (см. [38, формула (1.20)]). Применяя *неравенство для энтропийной мощности* к R_X, P_W и $R_Y = R_X * P_W$ [102, теорема 4], получаем

$$N(R_Y) \geqslant N(R_X) + N(P_W) = N(R_X) + \sigma_W^2.$$

Следовательно, достаточно доказать, что

$$\left(\frac{\sigma_X^2 + \sigma_W^2}{N(R_X) + \sigma_W^2}\right)^{\sigma_X^2 + \sigma_W^2} \leqslant \left(\frac{\sigma_X^2}{N(R_X)}\right)^{\sigma_X^2}.$$

Положим $a = \sigma_X^2 + \sigma_W^2$, $b = \sigma_X^2 - N(R_X)$ и $c = \sigma_X^2$. Тогда имеем $a > c > 0$ и $c > b$ (что следует из конечности $h(R_X)$), и поэтому достаточно доказать, что

$$\left(\frac{a}{a-b}\right)^a \leqslant \left(\frac{c}{c-b}\right)^c,$$

что равносильно

$$a > c > 0 \quad \text{и} \quad c > b \quad \Longrightarrow \quad \left(1 - \frac{b}{c}\right)^c \leqslant \left(1 - \frac{b}{a}\right)^a.$$

Это утверждением является вариантом неравенства Бернулли, доказанного в [103, теорема 3.1, пп. (r_7') и (r_7'')], что и завершает доказательство. ▲

§ 6. Доказательства результатов о предпорядке меньшего искажения

Наконец, обратимся к выводу эквивалентных характеризаций отношения $\succeq_{\ln}$ через операторную выпуклость. В п. 6.1 представлены предварительные сведения об операторно выпуклых функциях, в п. 6.2 доказывается теорема 6, а в п. 6.3 – теорема 7.

6.1. Операторно выпуклые функции. Для любого непустого (конечного или бесконечного, открытого или замкнутого) интервала $I \subseteq \mathbb{R}$ обозначим через $\mathbb{C}^{n\times n}_{\mathrm{Herm}}(I)$ множество всех (комплексных) эрмитовых матриц размера $n \times n$, все собственные значения которых принадлежат I, где $\mathbb{C}^{n\times n}_{\mathrm{Herm}}(\mathbb{R})$ – пространство всех эрмитовых матриц. Любую заданную функцию $f\colon I \to \mathbb{R}$ можно продолжить до функции $f\colon \mathbb{C}^{n\times n}_{\mathrm{Herm}}(I) \to \mathbb{C}^{n\times n}_{\mathrm{Herm}}(\mathbb{R})$ следующим образом [87, гл. V.1]:

$$\forall A \in \mathbb{C}^{n\times n}_{\mathrm{Herm}}(I), \quad f(A) \triangleq U \operatorname{diag}((f(\lambda_1), \ldots, f(\lambda_n)))U^H, \tag{94}$$

где $A = U \operatorname{diag}((\lambda_1, \ldots, \lambda_n))U^H$ – спектральное разложение A с вещественными собственными значениями $\lambda_1, \ldots, \lambda_n \in I$, $U \in \mathbb{C}^{n\times n}$ – унитарная матрица, а U^H – ее эрмитово сопряженная. Будем говорить, что f *операторно выпукла*, если для любого $n \geqslant 1$, любой пары матриц $A, B \in \mathbb{C}^{n\times n}_{\mathrm{Herm}}(I)$ и любого $\lambda \in [0,1]$ справедливо соотношение

$$\lambda f(A) + (1-\lambda) f(B) \succeq_{\mathrm{PSD}} f(\lambda A + (1-\lambda)B), \tag{95}$$

где через $\succeq_{\mathrm{PSD}}$ обозначен *частичный порядок Лёвнера* (т.е. для любых матриц $A, B \in \mathbb{C}^{n\times n}_{\mathrm{Herm}}(\mathbb{R})$ отношение $A \succeq_{\mathrm{PSD}} B$ означает, что матрица $A - B$ положительно полуопределена), и непосредственной проверкой легко убедиться, что $\lambda A + (1-\lambda)B \in \mathbb{C}^{n\times n}_{\mathrm{Herm}}(I)$ (см. [87, гл. V.1]). Заметим, что операторно выпуклая функция $f\colon I \to \mathbb{R}$ очевидным образом выпукла, и ее аффинные преобразования со сдвигами $g\colon \{c + x : x \in I\} \to \mathbb{R}$, $g(t) = af(t-c) + b$, также операторно выпуклы для любых $a \geqslant 0$, $b \in \mathbb{R}$ и $c \in \mathbb{R}$.

Удивительное свойство операторно выпуклых функций состоит в том, что они характеризуются *интегральными представлениями* определенного типа – см. *теоремы Лёвнера* в [87, гл. V.4, задача V.5.5]. В частности, для любой операторно выпуклой функции $f\colon (0,\infty)\to\mathbb{R}$, такой что $f(1)=0$, существуют константы $a\in\mathbb{R}$ и $b\geqslant 0$ и конечная положительная мера μ на $(1,\infty)$ (с соответствующей борелевской σ-алгеброй), такие что

$$f(t)=a(t-1)+b(t-1)^2+\int\limits_{(1,\infty)}\frac{(t-1)(\omega t-\omega-1)}{t+\omega-1}\,d\mu(\omega), \tag{96}$$

что следует из [18, формула (7)] и соответствующих ссылок (заметим, что наша f и функция g в [18] связаны как $f(t)=g(t-1)$). Как отмечено в [18], верно и обратное, т.е. функции вида (96) операторно выпуклы. Следующая лемма является прямым следствием формулы (96), которую мы выделили из [18] и представляем здесь в более прозрачном виде, показывающем роль дивергенций Винце – Ле Кама.

Л е м м а 8 (интегральное представление [18, с. 33]). *Рассмотрим произвольную f-дивергенцию, где функция $f\colon (0,\infty)\to\mathbb{R}$ операторно выпукла и $f(1)=0$. Тогда существует константа $b\geqslant 0$ и конечная положительная мера τ на $(0,1)$ (с соответствующей борелевской σ-алгеброй), такие что для любых $R_X, P_X\in\mathcal{P}_{\mathcal{X}}$*

$$D_f(R_X\,\|\,P_X)=b\chi^2(R_X\,\|\,P_X)+\int\limits_{(0,1)}\frac{1+\lambda^2}{\lambda(1-\lambda)}\,\mathrm{LC}_\lambda(R_X\,\|\,P_X)\,d\tau(\lambda).$$

Д о к а з а т е л ь с т в о. Зафиксируем произвольные вероятностные меры $R_X, P_X\in\in\mathcal{P}_{\mathcal{X}}$, и пусть случайная величина X имеет распределение P_X. Тогда, поскольку для f справедливо интегральное представление (96), подставим $t=R_X(X)/P_X(X)$ в (96) и перейдем к математическим ожиданиям:

$$\mathbf{E}\left[f\left(\frac{R_X(X)}{P_X(X)}\right)\right]=b\,\mathbf{E}\left[\left(\frac{R_X(X)}{P_X(X)}-1\right)^2\right]+$$
$$+\int\limits_{(1,\infty)}\mathbf{E}\left[\frac{\left(\frac{R_X(X)}{P_X(X)}-1\right)\left(\omega\frac{R_X(X)}{P_X(X)}-\omega-1\right)}{\frac{R_X(X)}{P_X(X)}+\omega-1}\right]d\mu(\omega),$$

где первое слагаемое в правой части (96) исчезает после перехода к математическому ожиданию (см. свойство аффинной инвариантности в п. 2.1). Отсюда

$$D_f(R_X\,\|\,P_X)=b\chi^2(R_X\,\|\,P_X)+\int\limits_{(1,\infty)}\mathbf{E}\left[\frac{(1+\omega^2)\left(\frac{R_X(X)}{P_X(X)}-1\right)^2}{\omega\left(\frac{R_X(X)}{P_X(X)}+\omega-1\right)}\right]d\mu(\omega),$$

где левая часть вытекает из определения 1, χ^2-дивергенция получается из (4), а последнее слагаемое следует из свойства аффинной инвариантности в п. 2.1 и соотношения

$$\forall t\geqslant 0,\ \forall\omega>1,\quad \frac{(t-1)(\omega t-\omega-1)}{t+\omega-1}=\frac{(1+\omega^2)(t-1)^2}{\omega(t+\omega-1)}-\frac{t-1}{\omega}.$$

Теперь заметим, что замена переменных $\omega = \dfrac{1}{\lambda}$ дает

$$D_f(R_X \,\|\, P_X) = b\chi^2(R_X \,\|\, P_X) + \int\limits_{(0,1)} \mathbf{E}\left[\frac{(1+\lambda^2)\Big(\dfrac{R_X(X)}{P_X(X)} - 1\Big)^2}{\Big(\lambda\dfrac{R_X(X)}{P_X(X)} + 1 - \lambda\Big)}\right] d\tau(\lambda)$$

для некоторой конечной положительной меры τ на $(0,1)$. Наконец, замечая, что подынтегральное выражение в правой части пропорционально дивергенции Винце–Ле Кама (см. п. 2.1), непосредственными преобразованиями получаем требуемое интегральной представление. ▲

Лемма 8 использовалась в [18, с. 33] (в несколько другом виде) для доказательства предложения 6 (см. [18, теорема 1]). Кроме того, в [8, с. 3363] также использовалась ключевая идея из [18, с. 33] для получения альтернативного интегрального представления (через дивергенцию Винце–Ле Кама и χ^2-дивергенцию), аналогичного лемме 8. Однако представление из [8, с. 3363] справедливо лишь для операторно выпуклых функций f, таких что $f(0)$ конечно, в то время как лемма 8 верна и для бесконечных $f(0)$.

6.2. Доказательство теоремы 6. Теперь с помощью леммы 8 докажем теорему 6.

Доказательство теоремы 6. Наше доказательство использует технику доказательств утверждений [18, теорема 1] и [83, теорема 1] (см. также [8, п. III-C]). Зафиксируем произвольную нелинейную операторно выпуклую функцию $f\colon (0,\infty) \to \to \mathbb{R}$, такую что $f(1) = 0$, где из нелинейности следует, что соответствующая f-дивергенция не тождественно равна нулю (см. свойство аффинной инвариантности в п. 2.1). Для любых стохастических матриц $P_{Y\,|\,X} = W \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ и $P_{Z|X} = V \in \mathcal{P}_{\mathcal{Z}|\mathcal{X}}$ вначале установим, что неравенство

$$\forall R_X, P_X \in \mathcal{P}_{\mathcal{X}}, \quad \chi^2(R_X W \,\|\, P_X W) \geqslant \chi^2(R_X V \,\|\, P_X V) \tag{97}$$

имеет место тогда и только тогда, когда

$$\forall R_X, P_X \in \mathcal{P}_{\mathcal{X}}, \quad D_f(R_X W \,\|\, P_X W) \geqslant D_f(R_X V \,\|\, P_X V). \tag{98}$$

Для доказательства необходимости условия 98, применяя лемму 8 и эквивалентное представление дивергенции Винце–Ле Кама из (6), получаем следующее интегральное представление нашей f-дивергенции через χ^2-дивергенцию (см. [18, с. 33]):

$$\forall R_X, P_X \in \mathcal{P}_{\mathcal{X}}, \quad D_f(R_X \,\|\, P_X) = b\chi^2(R_X \,\|\, P_X) + \\ + \int\limits_{(0,1)} \frac{1+\lambda^2}{(1-\lambda)^2}\chi^2\big(R_X \,\|\, \lambda R_X + (1-\lambda)P_X\big)\, d\tau(\lambda), \tag{99}$$

где $b \geqslant 0$ – некоторая константа, а τ – конечная положительная мера на $(0,1)$. В силу (97) также имеем

$$\forall R_X, P_X \in \mathcal{P}_{\mathcal{X}}, \\ \chi^2\big(R_X W \,\|\, (\lambda R_X + (1-\lambda)P_X)W\big) \geqslant \chi^2\big(R_X V \,\|\, (\lambda R_X + (1-\lambda)P_X)V\big) \tag{100}$$

для любого $\lambda \in (0,1)$. Поэтому, используя (97) и (100), а также интегральное представление (99), получаем (98), что и требовалось.

Для доказательства достаточности заметим, что из интегрального представления Лёвнера (96) следует, что f бесконечно дифференцируема и $f''(1) > 0$. В силу (98)

получаем также, что для любого $\lambda \in (0,1)$

$$\forall R_X, P_X \in \mathcal{P}_{\mathcal{X}}, \quad D_f\big(((1-\lambda)P_X + \lambda R_X)W \,\|\, P_X W\big) \geqslant$$
$$\geqslant D_f\big(((1-\lambda)P_X + \lambda R_X)V \,\|\, P_X V\big). \tag{101}$$

Поэтому, умножая обе части неравенства (101) на $2/(f''(1)\lambda^2) > 0$ и переходя к пределу при $\lambda \to 0^+$, из локальной аппроксимации f-дивергенций (равенство (13)) получаем (97) для всех $R_X \in \mathcal{P}_{\mathcal{X}}$ и всех $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$. Хотя наш вариант равенства (13) справедлив в предположении $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$, утверждение (97) верно также и для $P_X \in \mathcal{P}_{\mathcal{X}} \setminus \mathcal{P}_{\mathcal{X}}^{\circ}$ в силу непрерывности χ^2-дивергенции по второму аргументу при фиксированном первом.

Теперь заметим, что эквивалентность между (97) и (98) показывает, что все предпорядки на стохастических матрицах, определяемые через нелинейные операторно выпуклые f-дивергенции согласно (98) (способом, аналогичным определению 6) эквивалентны. Действительно, все они характеризуются χ^2-дивергенцией (см. (97)). Поскольку КЛ-дивергенция является нелинейной операторно выпуклой f-дивергенцией (см. замечание после теоремы 6 в п. 3.4 по поводу теоремы Лёвнера – Хайнца), предпорядок меньшего искажения $\succeq_{\ln}$ эквивалентен предпорядку, задаваемому χ^2-дивергенцией в (97); это в точности является содержанием утверждений [83, теорема 1] и [85, теорема 1]. Следовательно, $\succeq_{\ln}$ эквивалентен предпорядку, задаваемому соотношением (98), для любой нелинейной операторно выпуклой f-дивергенции, что и завершает доказательство. ▲

Теперь докажем предложение 6, чтобы показать, что оно является немедленным следствием теоремы 6.

Доказательство предложения 6. Зафиксируем произвольную нелинейную операторно выпуклую функцию $f\colon (0,\infty) \to \mathbb{R}$, такую что $f(1) = 0$, и любую стохастическую матрицу $P_{Y|X} = W \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$. Согласно теореме 6 для матрицы $|\mathcal{X}|$-ичного канала со стиранием $E_{1-\beta} \succeq_{\ln} P_{Y|X}$ тогда и только тогда, когда

$$\forall R_X, P_X \in \mathcal{P}_{\mathcal{X}}, \quad D_f(R_X W \,\|\, P_X W) \leqslant D_f(R_X E_{1-\beta} \,\|\, P_X E_{1-\beta}) = \beta D_f(R_X \,\|\, P_X),$$

где равенство разбирается в комментариях после формулы (30) в п. 2.2. Эта эквивалентность приводит к следующему обобщению соотношения (58):

$$\eta_f(P_{Y|X}) = \min\{\beta \in [0,1] :\ E_{1-\beta} \succeq_{\ln} P_{Y|X}\}. \tag{102}$$

Поэтому коэффициенты сжатия $\eta_f(P_{Y|X})$ для всех нелинейных операторно выпуклых f равны, и в частности, все они равны $\eta_{\chi^2}(P_{Y|X})$ и $\eta_{\mathrm{KL}}(P_{Y|X})$ (так как функции, соответственно, $f(t) = t^2 - 1$ и $f(t) = t\log(t)$ операторно выпуклы по теореме Лёвнера – Хайнца). ▲

6.3. Доказательство теоремы 7. Чтобы доказать теорему 7, нам понадобится полезная лемма о тензоризации. Пусть $X_i \in \mathcal{X}_i$, $Y_i \in \mathcal{Y}_i$ и $Z_i \in \mathcal{Z}_i$ – дискретные случайные величины с конечными множествами значений, и пусть $P_{Y_i|X_i} \in \mathcal{P}_{\mathcal{Y}_i|\mathcal{X}_i}$ и $P_{Z_i|X_i} \in \mathcal{P}_{\mathcal{Z}_i|\mathcal{X}_i}$ – стохастические матрицы, $i = 1, 2$. В следующей лемме сформулировано свойство тензоризации предпорядка меньшего искажения [7, предложение 16; 104, предложение 5].

Лемма 9 (тензоризация предпорядка меньшего искажения [7, 104]). *Если* $P_{Z_i|X_i} \succeq_{\ln} P_{Y_i|X_i}$ *для* $i = 1, 2$, *то* $P_{Z_1|X_1} \otimes P_{Z_2|X_2} \succeq_{\ln} P_{Y_1|X_1} \otimes P_{Y_2|X_2}$.

Наконец, приведем доказательство теоремы 7 с помощью теоремы 6, леммы 9 и соотношения (58).

Доказательство теоремы 7. Будем следовать схеме доказательства из [7]. Вначале заметим, что $\eta_i = \eta_{\mathrm{KL}}(P_{Y_i|X_i})$ для всех $i \in \{1,\ldots,n\}$ согласно предложению 6 (см. [18, теорема 1]). Пусть $Z_i \in \mathcal{Z}_i = \mathcal{X}_i \cup \{\mathrm{e}\}$ – дискретная случайная величина, и пусть $P_{Z_i|X_i} = E_{1-\eta_i} \in \mathcal{P}_{\mathcal{Z}_i|\mathcal{X}_i}$ – матрица $|\mathcal{X}_i|$-ичного канала со стиранием с вероятностью стирания $1-\eta_i$, определенная в (30), для всех $i \in \{1,\ldots,n\}$. Тогда, используя (58) (см. [7, предложение 15]), получаем, что матрица $P_{Z_i|X_i}$ менее искажающая, чем $P_{Y_i|X_i}$, для всех $i \in \{1,\ldots,n\}$. Теперь определим произведение стохастических матриц $P_{Z_1^n|X_1^n} \in \mathcal{P}_{\mathcal{Z}^n|\mathcal{X}^n}$:

$$P_{Z_1^n|X_1^n} = P_{Z_1|X_1} \otimes P_{Z_2|X_2} \otimes \ldots \otimes P_{Z_n|X_n},$$

где $Z_1^n = (Z_1,\ldots,Z_n)$ и $\mathcal{Z}^n = \mathcal{Z}_1 \times \ldots \times \mathcal{Z}_n$. Тогда по лемме 9 получаем, что $P_{Z_1^n|X_1^n}$ менее искажающая, чем $P_{Y_1^n|X_1^n}$.

Чтобы доказать вариант СНОД Самородницкого для f-дивергенции (64), зафиксируем произвольную пару вероятностных мер на входе $R_{X_1^n}, P_{X_1^n} \in \mathcal{P}_{\mathcal{X}^n}$. Тогда по теореме 6 имеем

$$D_f(R_{Y_1^n} \,\|\, P_{Y_1^n}) \leqslant D_f(R_{Z_1^n} \,\|\, P_{Z_1^n}), \tag{103}$$

где $R_{Z_1^n} = R_{X_1^n} P_{Z_1^n|X_1^n} \in \mathcal{P}_{\mathcal{Z}^n}$ и $P_{Z_1^n} = P_{X_1^n} P_{Z_1^n|X_1^n} \in \mathcal{P}_{\mathcal{Z}^n}$. Теперь заметим, что случайную величину Z_1^n можно эквивалентным образом представить как (X_S, S), где случайное подмножество S представляет собой позиции, которые не стираются под действием $P_{Z_1^n|X_1^n}$, а X_S – значения в этих позициях. (Заметим, что $Z_1^n = (\mathrm{e},\ldots,\mathrm{e})$ соответствует множеству $S = \varnothing$.) Таким образом,

$$D_f\big(R_{Z_1^n} \,\|\, P_{Z_1^n}\big) = \sum_{T \subseteq \{1,\ldots,n\}} P_S(T) \sum_{x_T} P_{X_T}(x_T) f\left(\frac{P_S(T) R_{X_T}(x_T)}{P_S(T) P_{X_T}(x_T)}\right) =$$

$$= \sum_{T \subseteq \{1,\ldots,n\}} P_S(T) D_f(R_{X_T} \,\|\, P_{X_T}),$$

где использован тот факт, что S не зависит от X_1^n, а также следующие соглашения: $D_f(R_{X_\varnothing} \,\|\, P_{X_\varnothing}) = 0$, и $P_S(T) D_f(R_{X_T} \,\|\, P_{X_T}) = 0$, если $P_S(T) = 0$. С учетом (103) отсюда получаем (64).

Чтобы доказать вариант СНОД Самородницкого для взаимной f-информации (65), зафиксируем произвольное совместное распределение P_{U,X_1^n} и рассмотрим цепь Маркова $U \to X_1^n \to (Y_1^n, Z_1^n)$, такую что Y_1^n и Z_1^n условно независимы при заданном X_1^n. Это задает совместное распределение P_{U,X_1^n,Y_1^n,Z_1^n} через P_{U,X_1^n}, $P_{Y_1^n|X_1^n}$ и $P_{Z_1^n|X_1^n}$. Для любой заданной $U = u \in \mathcal{U}$ из (103) получаем

$$D_f\big(P_{Y_1^n|U=u} \,\|\, P_{Y_1^n}\big) \leqslant D_f\big(P_{Z_1^n|U=u} \,\|\, P_{Z_1^n}\big),$$

где $P_{Y_1^n|U=u} = P_{X_1^n|U=u} P_{Y_1^n|X_1^n} \in \mathcal{P}_{\mathcal{Y}^n}$ и $P_{Z_1^n|U=u} = P_{X_1^n|U=u} P_{Z_1^n|X_1^n} \in \mathcal{P}_{\mathcal{Z}^n}$ – условные вероятностные меры (на выходе), соответствующие условной вероятностной мере (на входе) $P_{X_1^n|U=u} \in \mathcal{P}_{\mathcal{X}^n}$ (благодаря марковскому свойству), а $P_{Y_1^n} \in \mathcal{P}_{\mathcal{Y}^n}$ и $P_{Z_1^n} \in \mathcal{P}_{\mathcal{Z}^n}$ – маргинальные вероятностные меры для P_{U,X_1^n,Y_1^n,Z_1^n}. С учетом (8), переходя к математическим ожиданиям по маргинальной вероятностной мере P_U, получаем

$$I_f(U; Y_1^n) \leqslant I_f(U; Z_1^n). \tag{104}$$

Далее, используя эквивалентность между Z_1^n и (X_S, S), имеем (как и выше)

$$I_f(U; Z_1^n) = I_f(U; X_S, S) =$$

$$= \sum_{T\subseteq\{1,\dots,n\}} P_S(T) \sum_{u\in\mathcal{U}} P_U(u) \sum_{x_T} P_{X_T}(x_T) f\left(\frac{P_{X_T|U}(x_T\,|\,u)P_U(u)P_S(T)}{P_{X_T}(x_T)P_U(u)P_S(T)}\right) =$$

$$= \sum_{T\subseteq\{1,\dots,n\}} P_S(T) D_f\big(P_{U,X_T}\,\|\,P_U P_{X_T}\big) = \sum_{T\subseteq\{1,\dots,n\}} P_S(T) I_f(U;X_T),$$

где использовано соотношение (8) и тот факт, что S не зависит от (U, X_1^n), а также следующие соглашения: $I_f(U;X_\varnothing) = 0$, и $P_S(T)I_f(U;X_T) = 0$, если $P_S(T) = 0$. С учетом (104) отсюда получаем (65).

И наконец, случай $\eta_i = \eta$ для всех $i \in \{1,\dots,n\}$ в (66) получается подстановкой выражения для $P_S(T)$ в (65) и дальнейшими стандартными выкладками, что и завершает доказательство. ▲

Заметим, что в условиях теоремы 7, если к тому же $\eta_i = \eta$ для всех $i \in \{1,\dots,n\}$, то наше обобщенное СНОД Самородницкого действительно точнее, чем тензоризованное СНОД (63):

$$I_f(U;Y_1^n) \leqslant \sum_{T\subseteq\{1,\dots,n\}} \eta^{|T|}(1-\eta)^{n-|T|} I_f(U;X_T) \leqslant (1-(1-\eta)^n) I_f(U;X_1^n). \tag{105}$$

Это так, поскольку $I_f(U;Z_1^n) \leqslant (1-(1-\eta)^n) I_f(U;X_1^n)$ с учетом соотношения (63) для произведения стохастических матриц $P_{Z_1^n|X_1^n}$.

§ 7. Заключение

Здесь мы вкратце перечислим наши основные достижения и предложим некоторые направления дальнейшего исследования. Вначале мы описали некоторые свойства коэффициентов сжатия для совместных распределений. Точнее, в теореме 1 было показано, что такие коэффициенты сжатия достигают своей верхней границы, равной единице, когда совместное распределение разложимо, а в теореме 2 – что наложение на оптимизационную задачу, определяющую величину $\eta_f(P_X, P_{Y|X})$ дополнительного ограничения "локальной аппроксимации", при котором f-дивергенции на входе становятся сколь угодно малыми, приводит к оптимальному значению $\eta_{\chi^2}(P_X, P_{Y|X})$. Этот последний результат довольно прозрачно объясняет интуитивные соображения по поводу нижней границы максимальной корреляции в п. 7 предложения 3. Затем в теореме 3 мы вывели линейную верхнюю границу на $\eta_f(P_X, P_{Y|X})$ через $\eta_{\chi^2}(P_X, P_{Y|X})$ для некоторого класса f-дивергенций, а в теореме 4 улучшили эту границу для важного частного случая $\eta_{\mathrm{KL}}(P_X, P_{Y|X})$. Подобные границы полезны для режимов со слабой зависимостью, таких как возникающие при анализе эргодичности цепей Маркова (как показано в предложении 7). В духе сравнения коэффициентов сжатия для совместных распределений мы также дали альтернативное доказательство эквивалентности $\eta_{\mathrm{KL}}(P_X, P_{Y|X}) = \eta_{\chi^2}(P_X, P_{Y|X})$ для совместно гауссовских распределений $P_{X,Y}$, описываемых моделями АБГШ в теореме 5 и § 5. Это доказательство продемонстрировало, что при наложении достаточно сильного ограничения на мощность (или ограниченный второй момент) в экстремальной задаче для величины $\eta_{\mathrm{KL}}(P_X, P_{Y|X})$ ее значение не изменяется. Наконец, в области коэффициентов сжатия для стохастических матриц мы обобщили предложение 6 в теореме 6 и установили, что предпорядок меньшего искажения на стохастических матрицах полностью характеризуется любой нелинейной операторно выпуклой f-дивергенцией. Более того, в качестве приложения этой характеризации мы также обобщили СНОД Самородницкого на все нелинейные операторно выпуклые f-дивергенции в теореме 7.

Как было указано в п. 4.4, константы в линейных границах из теорем 3 и 4 "неотчетливо" изменяются с изменением размерности распределения-произведения.

В то время как результаты, подобные следствию 2, частично позволяют справиться с этой проблемой тензоризации, одним из обязательных направлений будущих исследований должно стать получение линейных границ, константы в которых должным образом ведут себя при тензоризации. Еще одно, по-видимому, более конкретное направление будущей работы – это вывод оптимального зависящего от распределений улучшенного варианта леммы 4 (как предложено в [79, замечание, с. 5380]). Такой улучшенный вариант можно было бы использовать для уточнения теоремы 3 так, чтобы из нее вытекала теорема 4 вместо следствия 1. Однако и такой улучшенный вариант не сможет охватить вопрос, связанный с тензоризацией, что гораздо важнее для данных границ.

ПРИЛОЖЕНИЕ A: ДОКАЗАТЕЛЬСТВО ПРЕДЛОЖЕНИЯ 2

Схема доказательства описана в [4], а само доказательство приведено в дипломной работе автора [48, теорема 3.2.4] для случая $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ и $P_Y \in \mathcal{P}_{\mathcal{Y}}^{\circ}$. Мы приводим его здесь для полноты изложения.

Пусть маргинальные вероятностные меры для X и Y таковы, что $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ и $P_Y \in \mathcal{P}_{\mathcal{Y}}^{\circ}$. Вначале покажем, что наибольшее сингулярное число МДП B равно единице. Рассмотрим матрицу

$$M = \operatorname{diag}\bigl(\sqrt{P_Y}\bigr)^{-1} B^T B \operatorname{diag}\bigl(\sqrt{P_Y}\bigr) = \operatorname{diag}(P_Y)^{-1} W^T \operatorname{diag}(P_X) W = VW,$$

где $V = \operatorname{diag}(P_Y)^{-1} W^T \operatorname{diag}(P_X) \in \mathcal{P}_{\mathcal{X}|\mathcal{Y}}$ – стохастическая по строкам обратная матрица вероятностей перехода, соответствующая условному распределению $P_{X|Y}$. Заметим, что M имеет тот же набор собственных значений, что и матрица Грама МДП $B^T B$, поскольку она просто определяется через преобразование подобия. Так как $B^T B$ положительно полуопределена, собственные значения матриц M и $B^T B$ – неотрицательные вещественные числа согласно *спектральной теореме* (см. [49, п. 2.5]). Более того, так как и V, и W стохастические по строкам, их произведение $M = VW$ также стохастическое по строкам. Следовательно, наибольшее собственное значение матриц M и $B^T B$ равно единице по *теореме Перрона – Фробениуса* (см. [49, гл. 8]). Отсюда следует, что наибольшее сингулярное число матрицы B также равно единице. Отметим также, что $\sqrt{P_X}$ и $\sqrt{P_Y}$ – соответственно левый и правый сингулярные векторы матрицы B, соответствующие сингулярному числу, равному единице. Действительно,

$$\sqrt{P_X} B = \sqrt{P_X} \operatorname{diag}\bigl(\sqrt{P_X}\bigr) W \operatorname{diag}\bigl(\sqrt{P_Y}\bigr)^{-1} = \sqrt{P_Y},$$
$$B \sqrt{P_Y}^T = \operatorname{diag}\bigl(\sqrt{P_X}\bigr) W \operatorname{diag}\bigl(\sqrt{P_Y}\bigr)^{-1} \sqrt{P_Y}^T = \sqrt{P_X}^T.$$

Далее, следуя определению 3, пусть $f \in \mathbb{R}^{|\mathcal{X}|}$ и $g \in \mathbb{R}^{|\mathcal{Y}|}$ – векторы-столбцы, задающие множества значений функций $f\colon \mathcal{X} \to \mathbb{R}$ и $g\colon \mathcal{Y} \to \mathbb{R}$ соответственно. Заметим, что математические ожидания в определении 3 можно выразить через B, P_X, P_Y, f и g:

$$\mathbf{E}[f(X)g(Y)] = \bigl(\operatorname{diag}\bigl(\sqrt{P_X}\bigr) f\bigr)^T B \bigl(\operatorname{diag}\bigl(\sqrt{P_Y}\bigr) g\bigr),$$
$$\mathbf{E}[f(X)] = \sqrt{P_X}^T \bigl(\operatorname{diag}\bigl(\sqrt{P_X}\bigr) f\bigr),$$
$$\mathbf{E}[g(Y)] = \sqrt{P_Y}^T \bigl(\operatorname{diag}\bigl(\sqrt{P_Y}\bigr) g\bigr),$$
$$\mathbf{E}[f(X)^2] = \bigl\|\operatorname{diag}\bigl(\sqrt{P_X}\bigr) f\bigr\|_2^2,$$
$$\mathbf{E}[g(Y)^2] = \bigl\|\operatorname{diag}\bigl(\sqrt{P_Y}\bigr) g\bigr\|_2^2.$$

Полагая $a = \operatorname{diag}\big(\sqrt{P_X}\big) f$ и $b = \operatorname{diag}\big(\sqrt{P_Y}\big) g$, из определения 3 получаем

$$\rho(X;Y) = \max_{\substack{a\in\mathbb{R}^{|\mathcal{X}|},\, b\in\mathbb{R}^{|\mathcal{Y}|} \\ \sqrt{P_X}a=\sqrt{P_Y}b=0 \\ \|a\|_2^2=\|b\|_2^2=1}} a^T B b,$$

где оптимизация проводится по всем $a \in \mathbb{R}^{|\mathcal{X}|}$ и $b \in \mathbb{R}^{|\mathcal{Y}|}$, поскольку $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ и $P_Y \in \mathcal{P}_{\mathcal{Y}}^{\circ}$. Так как a и b ортогональны, соответственно, левому и правому сингулярным векторам, соответствующим максимальному сингулярному числу матрицы B, равному единице, эта максимизация дает второе по величине сингулярное число матрицы B с помощью альтернативной версии (см., например, [105, лемма 2]) *принципа минимакса Куранта – Фишера – Вейля* (см. [49, теоремы 4.2.6 и 7.3.8]). Это доказывает, что $\rho(X;Y)$ – второе по величине сингулярное число МДП, когда $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ и $P_Y \in \mathcal{P}_{\mathcal{Y}}^{\circ}$.

Наконец, покажем, что без ограничения общности можно считать, что $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ и $P_Y \in \mathcal{P}_{\mathcal{Y}}^{\circ}$. Когда P_X или P_Y имеют нулевые компоненты, X и Y принимают значения лишь на носителях $\operatorname{supp}(P_X) = \{x \in \mathcal{X} : P_X(x) > 0\} \subseteq \mathcal{X}$ и $\operatorname{supp}(P_Y) = \{y \in \mathcal{Y} : P_Y(y) > 0\} \subseteq \mathcal{Y}$ соответственно, что означает, что $P_X \in \mathcal{P}_{\operatorname{supp}(P_X)}^{\circ}$ и $P_Y \in \mathcal{P}_{\operatorname{supp}(P_Y)}^{\circ}$. Пусть B обозначает "истинную" МДП размера $|\mathcal{X}| \times |\mathcal{Y}|$, соответствующую вероятностной мере $P_{X,Y}$ на $\mathcal{X} \times \mathcal{Y}$, а B_{supp} – "ограниченную на носитель" МДП размера $|\operatorname{supp}(P_X)| \times |\operatorname{supp}(P_Y)|$, соответствующую вероятностной мере $P_{X,Y}$ на $\operatorname{supp}(P_X) \times \operatorname{supp}(P_Y)$. Очевидно, B можно восстановить по B_{supp}, добавляя нулевые строки и столбцы, соответствующие компонентам с нулевой вероятностью в $\mathcal{X}$ и $\mathcal{Y}$ соответственно. Поэтому B и B_{supp} имеют одинаковые ненулевые сингулярные значения (с учетом кратностей), откуда следует, что и второе по величине сингулярное число у них одинаково, что завершает доказательство. ▲

ПРИЛОЖЕНИЕ B: ДОКАЗАТЕЛЬСТВО ПРЕДЛОЖЕНИЯ 3

Свойство 1: Нормировка коэффициентов сжатия очевидна в силу неотрицательности f-дивергенций и соответствующих им неравенств об обработке данных (7). Отметим, что в случае $\eta_{\chi^2}(P_X, P_{Y|X}) = \rho(X;Y)^2$ (где использовалось (23)) условие $0 \leqslant \rho(X;Y) \leqslant 1$ – это третья аксиома Реньи в определении максимальной корреляции [12].

Свойство 2: Дадим простое доказательство этого хорошо известного свойства. Без ограничения общности предположим, что $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$, отбрасывая все компоненты $\mathcal{X}$, имеющие нулевую вероятность. Если в результате $|\mathcal{X}| = 1$, то X – константа п.н., и результат тривиален. Итак, можно считать, что $|\mathcal{X}| \geqslant 2$. Так как W – матрица единичного ранга (все ее строки равны P_Y, т.е. $P_{Y|X=x} = P_Y$ для всех $x \in \mathcal{X}$) тогда и только тогда, когда X и Y независимы, достаточно показать, что W имеет единичный ранг тогда и только тогда, когда $\eta_f(P_X, P_{Y|X}) = 0$.

Для доказательства в одну сторону заметим, что если ранг W равен единице, то все ее строки равны P_Y, и $R_X W = P_Y$ для всех $R_X \in \mathcal{P}_{\mathcal{X}}$. Следовательно, $\eta_f(P_X, P_{Y|X}) = 0$ в силу определения 2, поскольку $D_f(R_X W \,\|\, P_X W) = 0$ для всех вероятностных мер на входе $R_X \in \mathcal{P}_{\mathcal{X}}$.

Для доказательства обратного утверждения применим слегка измененный вариант рассуждения из [48, лемма 3.1.5], которое использовалось для доказательства случая $\eta_{\text{KL}}(P_X, P_{Y|X})$. Для любых $x \in \mathcal{X}$ и $\lambda \in (0,1)$ рассмотрим $R_X = (1-\lambda)\delta_x + \lambda \boldsymbol{u} \in \mathcal{P}_{\mathcal{X}}^{\circ}$, где δ_x – вероятностная дельта-мера Кронекера, такая что $\delta_x(x) = 1$ и $\delta_x(x') = 0$ для $x' \in \mathcal{X} \setminus \{x\}$, $\boldsymbol{u}$ – равномерная вероятностная мера, а λ выбрано так, что $R_X \neq P_X$. Тогда $0 < D_f(R_X \,\|\, P_X) < +\infty$, поскольку f строго выпукла в единице, и $D_f(R_X W \,\|\, P_X W) = 0$, так как $\eta_f(P_X, P_{Y|X}) = 0$. Отсюда

следует, что

$$(1-\lambda)P_{Y\,|\,X=x}+\lambda\boldsymbol{u}W=R_XW=P_XW=P_Y,$$

так как f строго выпукла в единице. Переходя к пределу при $\lambda\to 0$, получаем, что каждая строка W равна P_Y. (Заметим, что невозможно применить этот аргумент просто при $\lambda=0$, или $R_X=\delta_x$, поскольку $f(0)$ может равняться бесконечности.) Значит, W имеет единичный ранг.

Заметим также, что в случае $\eta_{\chi^2}(P_X,P_{Y\,|\,X})$ это свойство максимальной корреляции является четвертой аксиомой Реньи из [12].

Свойство 3: См. теорему 1 и Приложение C. Заметим, что случай $\eta_{\chi^2}(P_X,P_{Y\,|\,X})$ этого результата был доказан в [13, 15] (см. лемму 10), а случай $\eta_{\mathrm{KL}}(P_X,P_{Y\,|\,X})$ – в [15].

Свойство 4: Это доказано в [8, предложение III.3].

Свойство 5: Это доказано в [8, теорема III.9]. Заметим также, что два доказательства свойства тензоризации для $\eta_{\mathrm{KL}}(P_X,P_{Y\,|\,X})$ можно найти в [4], а доказательство свойства тензоризации для $\eta_{\chi^2}(P_X,P_{Y\,|\,X})$ – в [13].

Свойство 6: Для доказательства первой части этого утверждения обозначим через $P_{U,X,Y}$ совместную вероятностную меру для (U,X,Y), состоящую из маргинальной вероятностной меры $P_U\in\mathcal{P}_{\mathcal{U}}$ и условных распределений $P_{X\,|\,U}=S\in\mathcal{P}_{\mathcal{X}|\mathcal{U}}$ и $P_{Y\,|\,X}=W\in\mathcal{P}_{\mathcal{Y}|\mathcal{X}}$ (т.е. стохастических по строкам матриц вероятностей перехода). Тогда $P_{Y\,|\,U}=SW\in\mathcal{P}_{\mathcal{Y}|\mathcal{U}}$ согласно марковскому свойству. Заметим, что для любой вероятностной меры $R_U\in\mathcal{P}_{\mathcal{U}}\setminus\{P_U\}$

$$D_f(R_USW\,\|\,P_USW)\leqslant\eta_f(P_X,P_{Y\,|\,X})\eta_f(P_U,P_{X\,|\,U})D_f(R_U\,\|\,P_U),$$

где $P_X=P_US$ и дважды использовалось СНОД (15). Отсюда

$$\eta_f(P_U,P_{Y\,|\,U})\leqslant\eta_f(P_U,P_{X\,|\,U})\eta_f(P_X,P_{Y\,|\,X})$$

согласно определению 2.

Частный случай этого результата для η_{χ^2} соответствует свойству субмультипликативности второго по величине сингулярного числа МДП. Это свойство субмультипликативности также верно и для i-го по величине сингулярного числа МДП (см. [106, теорема 2]), что полезно в приложениях, связанных с кодированием распределенных источников и каналов. При этом результат из [106, теорема 2] также доказан в [40, теорема 3], где показана связь с главными компонентами инерции и максимальной корреляцией.

Для доказательства второй части утверждения заметим, что для фиксированного $P_{X,Y}$ и любого $P_{U\,|\,X}$, такого что $U\to X\to Y$ образуют цепь Маркова (с совместной вероятностной мерой $P_{U,X,Y}$) и $\eta_f(P_U,P_{X\,|\,U})>0$ (для чего требуется, чтобы X не была постоянной п.н.), справедливо

$$\frac{\eta_f(P_U,P_{Y\,|\,U})}{\eta_f(P_U,P_{X\,|\,U})}\leqslant\eta_f(P_X,P_{Y\,|\,X}),\tag{106}$$

где использовалось установленное выше свойство субмультипликативности. Пусть $U=X$ п.н., так что $P_{U\,|\,X}\in\mathcal{P}_{\mathcal{X}|\mathcal{X}}$ – единичная матрица. Тогда $\eta_f(P_U,P_{X\,|\,U})=1$ и $\eta_f(P_U,P_{Y\,|\,U})=\eta_f(P_X,P_{Y\,|\,X})$ в силу определения 2. Поэтому равенство в (106) достигается, что завершает доказательство.

Отметим, что случай η_{χ^2} этого результата представлен в [68, лемма 6], где также доказано, что в качестве оптимальной стохастической матрицы $P_{U\,|\,X}$ можно взять $P_{Y\,|\,X}$ (так что U является копией Y) вместо единичной матрицы (где $U=X$ п.н.).

Свойство 7: Будем следовать нашей технике доказательства из [1, теорема 5], навеянной подходом из [17, теорема 5.4].

Напомним, что совместная вероятностная мера $P_{X,Y}$ состоит из $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ и $P_{Y|X} = W \in \mathcal{P}_{\mathcal{Y}|\mathcal{X}}$, и обозначим через B соответствующую МДП (см. (24)). Зададим траекторию сферически возмущенных вероятностных мер вида (11):

$$R_X^{(\varepsilon)} = P_X + \varepsilon K_X \operatorname{diag}\bigl(\sqrt{P_X}\,\bigr)$$

где

$$K_X \in \mathcal{S} \triangleq \Bigl\{x \in \bigl(\mathbb{R}^{|\mathcal{X}|}\bigr)^* : \sqrt{P_X}x^T = 0,\ \|x\|_2 = 1\Bigr\}$$

– вектор сферического возмущения. Когда эти вероятностные меры проходят через W, получаем траекторию на выходе

$$R_X^{(\varepsilon)}W = P_Y + \varepsilon K_X B \operatorname{diag}\bigl(\sqrt{P_Y}\,\bigr), \tag{107}$$

где B отображает сферические возмущения на входе в сферические возмущения на выходе [43]. Теперь согласно определению 2 имеем

$$\begin{aligned}
&\eta_f(P_X, P_{Y|X}) = \sup_{\substack{R_X \in \mathcal{P}_{\mathcal{X}} \\ 0<D_f(R_X\|P_X)<+\infty}} \frac{D_f(R_XW\,\|\,P_XW)}{D_f(R_X\,\|\,P_X)} \geqslant \\
&\geqslant \liminf_{\varepsilon\to 0} \sup_{K_X\in\mathcal{S}} \frac{\|K_XB\|_2^2 + o(1)}{\|K_X\|_2^2 + o(1)} \geqslant \sup_{K_X\in\mathcal{S}} \liminf_{\varepsilon\to 0} \frac{\|K_XB\|_2^2 + o(1)}{1 + o(1)} = \\
&= \eta_{\chi^2}(P_X, P_{Y|X}) = \rho(X;Y)^2,
\end{aligned}$$

где второе неравенство вытекает из (14) при сужении супремума на все вероятностные меры вида (11) (где $\varepsilon \neq 0$ – некоторая достаточно малая фиксированная величина) и перехода к пределу при $\varepsilon \to 0$, третье неравенство следует из неравенства о минимаксе, а заключительные равенства – из (25) и (23) соответственно, что и завершает доказательство.

Заметим, что предположения $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ и $P_Y \in \mathcal{P}_{\mathcal{Y}}^{\circ}$, будучи важными для задания вышеупомянутых траекторий вероятностных мер, не существенны для самого результата. Для специального случая $\eta_{\mathrm{KL}}(P_X, P_{Y|X})$ этот результат был впервые доказан в [15], а затем повторно в [3] и [1] – последние два доказательства используют различного типа аргументы, основанные на возмущениях. ▲

ПРИЛОЖЕНИЕ C: ДОКАЗАТЕЛЬСТВО ТЕОРЕМЫ 1

Для доказательства теоремы 1 нам потребуются следующие две леммы; первая из них известна, а вторая является новой.

Лемма 10 (разложимость и максимальная корреляция [13, 15]). *Совместная вероятностная мера $P_{X,Y}$ является разложимой тогда и только тогда, когда $\eta_{\chi^2}(P_X, P_{Y|X}) = \rho(X;Y)^2 = 1$.*

Доказательство. Хотя этот результат и был доказан в [13, 15], для полноты изложения мы приведем здесь доказательство. Пусть $P_{X,Y}$ разложима и существуют функции $h\colon \mathcal{X} \to \mathbb{R}$ и $g\colon \mathcal{Y} \to \mathbb{R}$, такие что $h(X) = g(Y)$ п.н. и $\operatorname{Var}(h(X)) > 0$. Тогда без ограничения общности можно считать, что $\mathbf{E}[h(X)] = 0$ и $\mathbf{E}[h(X)^2] = 1$, откуда $\rho(X;Y) = 1$ согласно определению 3. Таким образом, $\eta_{\chi^2}(P_X, P_{Y|X}) = 1$ в силу (23).

В обратную сторону, пусть $\eta_{\chi^2}(P_X, P_{Y|X}) = 1$, или, что равносильно, $\rho(X;Y) = 1$ (в силу (23)). Пусть $h\colon \mathcal{X} \to \mathbb{R}$ и $g\colon \mathcal{Y} \to \mathbb{R}$ – функции, на которых достигается

$\rho(X;Y)$; такие функции существуют, когда $\mathcal{X}$ и $\mathcal{Y}$ конечны, так как в определении 3 берется экстремум непрерывной целевой функции по компактным множествам. Очевидно, что $h(X)$ и $g(Y)$ имеют нулевое среднее, единичную дисперсию и коэффициент корреляции Пирсона, равный 1. Отсюда следует, что $h(X) = g(Y)$ п.н., с помощью прямого (и хорошо известного) рассуждения, использующего неравенство Коши – Буняковского. Следовательно, $P_{X,Y}$ разложима. ▲

Лемма 11 (одновременная экстремальность). *Пусть задана выпуклая функция $f\colon (0,\infty) \to \mathbb{R}$, дважды дифференцируемая в единице и такая, что $f(1) = 0$ и $f''(1) > 0$. Тогда справедливо следующее*:

1. *Если $\eta_{\chi^2}(P_X, P_{Y|X}) = 1$, то $\eta_f(P_X, P_{Y|X}) = 1$;*
2. *Если f строго выпукла и удовлетворяет условию $f(0) < \infty$, то из равенства $\eta_f(P_X, P_{Y|X}) = 1$ следует $\eta_{\chi^2}(P_X, P_{Y|X}) = 1$.*

Доказательство. Утверждение 1 тривиальным образом следует из пп. 1 и 7 предложения 3.

Утверждение 2: Пусть $\eta_f(P_X, P_{Y|X}) = 1$. Рассмотрим последовательность вероятностных мер на входе $\left\{R_X^{(n)} \in \mathcal{P}_\mathcal{X} :\ 0 < D_f\big(R_X^{(n)} \,\|\, P_X\big) < +\infty,\ n \in \mathbb{N}\right\}$, на которой в пределе достигается $\eta_f(P_X, P_{Y|X})$:

$$\lim_{n\to\infty} \frac{D_f\big(R_X^{(n)} W \,\|\, P_Y\big)}{D_f\big(R_X^{(n)} \,\|\, P_X\big)} = 1.$$

С учетом секвенциональной компактности $\mathcal{P}_\mathcal{X}$ можно считать, что $R_X^{(n)} \to R_X$ для некоторого $R_X \in \mathcal{P}_\mathcal{X}$ при $n \to \infty$ (в смысле ℓ^2-нормы), при необходимости переходя к подпоследовательности. Отсюда получаем следующие две возможности.

Случай 1: Пусть $R_X = P_X$. В этом случае доказательство теоремы 2 в Приложении D показывает, что $\eta_f(P_X, P_{Y|X}) = \eta_{\chi^2}(P_X, P_{Y|X})$. Таким образом, получаем $\eta_{\chi^2}(P_X, P_{Y|X}) = 1$.

Случай 2: Пусть $R_X \neq P_X$. Так как $f(0) < \infty$, f строго выпукла и $P_X \in \mathcal{P}_\mathcal{X}^\circ$, то $0 < D_f(R_X \,\|\, P_X) < +\infty$. Следовательно, получаем

$$\lim_{n\to\infty} \frac{D_f\big(R_X^{(n)} W \,\|\, P_Y\big)}{D_f\big(R_X^{(n)} \,\|\, P_X\big)} = \frac{D_f(R_X W \,\|\, P_Y)}{D_f(R_X \,\|\, P_X)} = 1,$$

используя непрерывность f (вытекающую из ее выпуклости). Теперь заметим, что так как f строго выпукла и $0 < D_f(R_X W \,\|\, P_X W) = D_f(R_X \,\|\, P_X) < +\infty$, то Y является достаточной статистикой для X, позволяющей делать выводы о паре (R_X, P_X) (см. [28, теорема 14] или п. 2.1), откуда, в свою очередь, следует (см. [28, теорема 14] или п. 2.1), что

$$0 < \chi^2(R_X W \,\|\, P_X W) = \chi^2(R_X \,\|\, P_X) < +\infty.$$

Следовательно, $\eta_{\chi^2}(P_X, P_{Y|X}) = 1$ в силу (22). ▲

Доказательство теоремы 1 немедленно следует из лемм 10 и 11. ▲

ПРИЛОЖЕНИЕ D: ДОКАЗАТЕЛЬСТВО ТЕОРЕМЫ 2

Доказательство. Начнем с определения функции $\tau\colon (0,\infty) \to [0,1]$:

$$\tau(\delta) \triangleq \sup_{\substack{R_X \in \mathcal{P}_\mathcal{X} \\ 0 < D_f(R_X \,\|\, P_X) \leqslant \delta}} \frac{D_f(R_X W \,\|\, P_X W)}{D_f(R_X \,\|\, P_X)},$$

так что все, что требуется доказать, это

$$\lim_{n\to\infty} \tau(\delta_n) = \eta_{\chi^2}(P_X, P_{Y|X})$$

для любой убывающей последовательности $\{\delta_n > 0 :\ n \in \mathbb{N}\}$, такой что $\lim\limits_{n\to\infty} \delta_n = 0$. Заметим, что предел в левой части существует, поскольку при $\delta_n \to 0$ супремум по $\tau(\delta_n)$ является невозрастающим и ограничен снизу нулем.

Вначале докажем, что $\lim\limits_{n\to\infty} \tau(\delta_n) \geqslant \eta_{\chi^2}(P_X, P_{Y|X})$, следуя, по существу, доказательству п. 7 предложения 3 в Приложении В. Для этого рассмотрим траекторию сферически возмущенных вероятностных мер вида (11):

$$R_X^{(n)} = P_X + \varepsilon_n K_X \operatorname{diag}\bigl(\sqrt{P_X}\,\bigr),$$

где

$$K_X \in \mathcal{S} = \Bigl\{x \in \bigl(\mathbb{R}^{|\mathcal{X}|}\bigr)^* :\ \sqrt{P_X}x^T = 0,\ \|x\|_2 = 1\Bigr\}$$

– вектор сферического возмущения. Соответствующая траектория вероятностных мер на выходе после прохождения через W имеет вид (107):

$$R_X^{(n)}W = P_Y + \varepsilon_n K_X B \operatorname{diag}\bigl(\sqrt{P_Y}\,\bigr),$$

где B – МДП, соответствующая $P_{X,Y}$ в (24). Потребуем, чтобы все скалярные величины $\{\varepsilon_n \neq 0 :\ n \in \mathbb{N}\}$, определяющие нашу траекторию, удовлетворяли условию $\lim\limits_{n\to\infty} \varepsilon_n = 0$ и были достаточно малыми, так чтобы

$$D_f\bigl(R_X^{(n)} \,\|\, P_X\bigr) = \frac{f''(1)}{2}\varepsilon_n^2 \|K_X\|_2^2 + o(\varepsilon_n^2) \leqslant \delta_n,$$

где использовалось (14) (а также тот факт, что $f''(1)$ существует и строго положительна) и стандартная o-символика Бахмана – Ландау. По определению τ имеем

$$\begin{aligned}
&\sup_{K_X\in\mathcal{S}} \frac{D_f\bigl(R_X^{(n)}W \,\|\, P_XW\bigr)}{D_f\bigl(R_X^{(n)} \,\|\, P_X\bigr)} \leqslant \tau(\delta_n),\\
&\lim_{n\to\infty} \sup_{K_X\in\mathcal{S}} \frac{\frac{f''(1)}{2}\varepsilon_n^2 \|K_XB\|_2^2 + o(\varepsilon_n^2)}{\frac{f''(1)}{2}\varepsilon_n^2 \|K_X\|_2^2 + o(\varepsilon_n^2)} \leqslant \lim_{n\to\infty} \tau(\delta_n),\\
&\lim_{n\to\infty} \sup_{K_X\in\mathcal{S}} \frac{\|K_XB\|_2^2 + o(1)}{1 + o(1)} \leqslant \lim_{n\to\infty} \tau(\delta_n),\\
&\eta_{\chi^2}(P_X, P_{Y|X}) \leqslant \lim_{n\to\infty} \tau(\delta_n),
\end{aligned}$$

где во втором неравенстве использовалось (14) как в числителе, так и в знаменателе, а в последнем неравенстве – характеризация $\eta_{\chi^2}(P_X, P_{Y|X})$ через сингулярное число в (25).

Теперь докажем, что $\lim\limits_{n\to\infty} \tau(\delta_n) \leqslant \eta_{\chi^2}(P_X, P_{Y|X})$. Заметим, что для любого $n \in \mathbb{N}$ существует вероятностная мера $R_X^{(n)} \in \mathcal{P}_{\mathcal{X}}$, обладающая следующими свойствами:

1. $0 < D_f\bigl(R_X^{(n)} \,\|\, P_X\bigr) \leqslant \delta_n$;
2. $0 \leqslant \tau(\delta_n) - \dfrac{D_f\bigl(R_X^{(n)}W \,\|\, P_XW\bigr)}{D_f\bigl(R_X^{(n)} \,\|\, P_X\bigr)} \leqslant \dfrac{1}{2^n}$,

первое из которых справедливо, поскольку $R_X \mapsto D_f(R_X \,\|\, P_X)$ – непрерывное отображение при фиксированной $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ (что следует из выпуклости f), а второе – поскольку $\tau(\delta_n)$ определяется как супремум. Так как $\tau(\delta_n)$ сходится при $n \to \infty$, получаем

$$\lim_{n\to\infty} \frac{D_f\big(R_X^{(n)} W \,\|\, P_X W\big)}{D_f\big(R_X^{(n)} \,\|\, P_X\big)} = \lim_{n\to\infty} \tau(\delta_n). \tag{108}$$

С учетом секвенциальной компактности $\mathcal{P}_{\mathcal{X}}$ можно считать, что $R_X^{(n)}$ сходится при $n \to \infty$ (в смысле ℓ^2-нормы), переходя при необходимости к подпоследовательности. Так как $\lim\limits_{n\to\infty} D_f(R_X^{(n)} \,\|\, P_X) = 0$, то $\lim\limits_{n\to\infty} R_X^{(n)} = P_X$ в силу непрерывности отображения $R_X \mapsto D_f(R_X \,\|\, P_X)$ при фиксированной $P_X \in \mathcal{P}_{\mathcal{X}}^{\circ}$ и того факта, что f-дивергенция (где f строго выпукла в единице) равна нулю тогда и только тогда, когда ее аргументы равны. Зададим векторы сферического возмущения $\big\{K_X^{(n)} \in \mathcal{S} :\ n \in \mathbb{N}\big\}$ в виде

$$R_X^{(n)} = P_X + \varepsilon_n K_X^{(n)} \operatorname{diag}\big(\sqrt{P_X}\,\big),$$

где $\{\varepsilon_n \neq 0 :\ n \in \mathbb{N}\}$ задают необходимые масштабирования, а $\lim\limits_{n\to\infty} \varepsilon_n = 0$ (так как $\lim\limits_{n\to\infty} R_X^{(n)} = P_X$). Соответствующие вероятностные меры на выходе имеют вид (107) с необходимыми поправками, и отношение между выходной и входной f-дивергенциями можно, как и выше, аппроксимировать с помощью (14):

$$\frac{D_f(R_X^{(n)} W \,\|\, P_X W)}{D_f(R_X^{(n)} \,\|\, P_X)} = \frac{\frac{f''(1)}{2}\varepsilon_n^2 \big\|K_X^{(n)} B\big\|_2^2 + o(\varepsilon_n^2)}{\frac{f''(1)}{2}\varepsilon_n^2 \big\|K_X^{(n)}\big\|_2^2 + o(\varepsilon_n^2)} = \frac{\big\|K_X^{(n)} B\big\|_2^2 + o(1)}{1 + o(1)}.$$

С учетом секвенциальной компактности $\mathcal{S}$ можно считать, что $\lim\limits_{n\to\infty} K_X^{(n)} = K_X^{\star} \in \mathcal{S}$, переходя при необходимости к подпоследовательности. Отсюда, переходя к пределу при $n \to \infty$, получаем

$$\lim_{n\to\infty} \tau(\delta_n) = \|K_X^{\star} B\|_2^2 \leqslant \eta_{\chi^2}(P_X, P_{Y\,|\,X}),$$

где равенство следует из (108) и непрерывности отображения $\big(\mathbb{R}^{|\mathcal{X}|}\big)^* \ni x \mapsto \|xB\|_2^2$, а неравенство – из (25), что и завершает доказательство. ▲

ПРИЛОЖЕНИЕ E: ДОКАЗАТЕЛЬСТВО СЛЕДСТВИЯ 1

Доказательство. Выпуклая функция $f\colon (0,\infty) \to \mathbb{R}$, $f(t) = t\log(t)$, очевидно, строго выпукла и трижды дифференцируема в единице, причем $f(1) = 0$, $f'(1) = 1$, $f''(1) = 1 > 0$ и $f'''(1) = -1$. Кроме того, функция $g\colon (0,\infty) \to \mathbb{R}$, $g(t) = \dfrac{f(t) - f(0)}{t} = \log(t)$, где $f(0) = \lim\limits_{t\to 0^+} f(t) = 0$, очевидно, выпукла вверх. Таким образом, для доказательства следствия 1 с помощью теоремы 3 достаточно показать, что f удовлетворяет условию (75) для любого $t \in (0,\infty)$ (см. [79])

$$\big(f(t) - f'(1)(t-1)\big)\left(1 - \frac{f'''(1)}{3f''(1)}(t-1)\right) \geqslant \frac{f''(1)}{2}(t-1)^2,$$

что после упрощения приводится к виду

$$2t(t+2)\log(t) - (5t+1)(t-1) \geqslant 0.$$

Зададим $h\colon (0,\infty)\to\mathbb{R}$, $h(t)=2t(t+2)\log(t)-(5t+1)(t-1)$, и заметим, что

$$h'(t)=4(t+1)\log(t)-8(t-1),$$
$$h''(t)=4\log(t)+\frac{4}{t}-4\geqslant 0,$$

где неотрицательность второй производной следует из хорошо известного неравенства

$$\forall x>0,\quad x\log(x)\geqslant x-1.$$

Так как h выпукла (поскольку ее вторая производная неотрицательна) и $h(1)=$ $=h'(1)=0$, то $t=1$ является глобальной точкой минимума h, и $h(t)\geqslant 0$ для любого $t\in(0,\infty)$, что и требовалось.

Наконец, нетрудно проверить, что константа в следствии 1 равна

$$\frac{f'(1)+f(0)}{f''(1)\min\limits_{x\in\mathcal{X}}P_X(x)}=\frac{1}{\min\limits_{x\in\mathcal{X}}P_X(x)},$$

что и завершает доказательство. ▲

ПРИЛОЖЕНИЕ F: ДОКАЗАТЕЛЬСТВО ГРАНИЦЫ (78)

Доказательство. В [1, лемма 6] приведены два доказательства границы (78). Здесь мы изложим одно из них, использующее идеи выпуклого анализа. В нем применяется тот факт, что КЛ-дивергенция является так называемой дивергенцией Брегмана, соответствующей отрицательной функции энтропии Шеннона, а затем сильная выпуклость отрицательной функции энтропии Шеннона используется для вывода границы на КЛ-дивергенцию. Пусть $H_{\text{neg}}\colon \mathcal{P}_{\mathcal{X}}\to\mathbb{R}$ – отрицательная функция *энтропии Шеннона*, определяемая как

$$\forall Q_X\in\mathcal{P}_{\mathcal{X}},\quad H_{\text{neg}}(Q_X)\triangleq\sum_{x\in\mathcal{X}}Q_X(x)\log(Q_X(x)).$$

Так как *дивергенция Брегмана*, соответствующая функции H_{neg}, является КЛ-дивергенцией (см. [107]), то для всех $S_X\in\mathcal{P}_{\mathcal{X}}$ и $Q_X\in\mathcal{P}^{\circ}_{\mathcal{X}}$ справедливо

$$D(S_X\,\|\,Q_X)=H_{\text{neg}}(S_X)-H_{\text{neg}}(Q_X)-(S_X-Q_X)\nabla H_{\text{neg}}(Q_X),$$

где $\nabla H_{\text{neg}}\colon\mathcal{P}^{\circ}_{\mathcal{X}}\to\mathbb{R}^{|\mathcal{X}|}$ – градиент функции H_{neg}. При этом, так как H_{neg} дважды непрерывно дифференцируема, справедливо

$$\forall Q_X\in\mathcal{P}^{\circ}_{\mathcal{X}},\quad \nabla^2H_{\text{neg}}(Q_X)=\operatorname{diag}(Q_X)^{-1}\succeq_{\text{PSD}}I,$$

где через $\nabla^2H_{\text{neg}}\colon\mathcal{P}^{\circ}_{\mathcal{X}}\to\mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$ обозначен гессиан функции H_{neg}. (Заметим, что матрица $\operatorname{diag}(Q_X)^{-1}-I$ положительно полуопределена как диагональная матрица с неотрицательными элементами на диагонали.) Напомним [100, гл. 9], что дважды непрерывно дифференцируемая выпуклая функция $f\colon S\to R$ на открытой области $S\subseteq\mathbb{R}^n$ называется *сильно выпуклой*, если существует $m>0$, такое что для всех $x\in S$ выполнено $\nabla^2f(x)\succeq mI$. Значит, H_{neg} сильно выпукла на $\mathcal{P}^{\circ}_{\mathcal{X}}$. Следствием этой сильной выпуклости является такая квадратичная нижняя граница [100, гл. 9]:

$$H_{\text{neg}}(S_X)\geqslant H_{\text{neg}}(Q_X)+(S_X-Q_X)\nabla H_{\text{neg}}(Q_X)+\frac{1}{2}\|S_X-Q_X\|_2^2\iff$$
$$\iff\quad D(S_X\,\|\,Q_X)\geqslant\frac{1}{2}\|S_X-Q_X\|_2^2$$

для любых $S_X \in \mathcal{P}_{\mathcal{X}}$ и $Q_X \in \mathcal{P}^{\circ}_{\mathcal{X}}$, где разрешается $S_X \in \mathcal{P}_X \setminus \mathcal{P}^{\circ}_X$ в силу непрерывности H_{neg}. Это в точности то же самое, что получалось при ослаблении границы (71) в доказательстве леммы 2 с помощью соотношений $\|S_X - Q_X\|_1 \geqslant \|S_X - Q_X\|_2$ и (70). Наконец, для любых $S_X \in \mathcal{P}_{\mathcal{X}}$ и $Q_X \in \mathcal{P}^{\circ}_{\mathcal{X}}$ имеем

$$D(S_X \,\|\, Q_X) \geqslant \frac{1}{2}\,\|S_X - Q_X\|_2^2 \geqslant \frac{\min\limits_{x\in\mathcal{X}} Q_X(x)}{2}\chi^2(S_X \,\|\, Q_X),$$

где второе неравенство вытекает из (4). Это тривиальным образом справедливо также и для всех $Q_X \in \mathcal{P}_{\mathcal{X}} \setminus \mathcal{P}^{\circ}_{\mathcal{X}}$. ▲

Первый автор выражает благодарность проф. Юрию Полянскому за весьма полезные обсуждения по поводу теорем 6 и 7, а также в целом за обсуждение вопросов, связанных с коэффициентами сжатия.

СПИСОК ЛИТЕРАТУРЫ

1. *Makur A., Zheng L.* Bounds between Contraction Coefficients // Proc. 53rd Annual Allerton Conf. on Communication, Control, and Computing. Monticello, IL, USA. Sept. 29 – October 2, 2015. P. 1422–1429.
2. *Erkip E., Cover T.M.* The Efficiency of Investment Information // IEEE Trans. Inform. Theory. 1998. V. 44. № 3. P. 1026–1040.
3. *Kamath S., Anantharam V.* Non-interactive Simulation of Joint Distributions: The Hirschfeld–Gebelein–Rényi Maximal Correlation and the Hypercontractivity Ribbon // Proc. 50th Annual Allerton Conf. on Communication, Control, and Computing. Monticello, IL, USA. Oct. 1–5, 2012. P. 1057–1064.
4. *Anantharam V., Gohari A., Kamath S., Nair C.* On Maximal Correlation, Hypercontractivity, and the Data Processing Inequality Studied by Erkip and Cover, `arXiv:1304.6133 [cs.IT]`, 2013.
5. *Anantharam V., Gohari A., Kamath S., Nair C.* On Hypercontractivity and the Mutual Information between Boolean Functions // Proc. 51st Annual Allerton Conf. on Communication, Control, and Computing. Monticello, IL, USA. Oct. 2–4, 2013. P. 13–19.
6. *Polyanskiy Y., Wu Y.* Dissipation of Information in Channels with Input Constraints // IEEE Trans. Inform. Theory. 2016. V. 62. № 1. P. 35–55.
7. *Polyanskiy Y., Wu Y.* Strong Data-Processing Inequalities for Channels and Bayesian Networks // Convexity and Concentration. New York: Springer, 2017. P. 211–249.
8. *Raginsky M.* Strong Data Processing Inequalities and Φ-Sobolev Inequalities for Discrete Channels // IEEE Trans. Inform. Theory. 2016. V. 62. № 6. P. 3355–3389.
9. *Hirschfeld H.O.* A Connection between Correlation and Contingency // Math. Proc. Cambridge Philos. Soc. 1935. V. 31. № 4. P. 520–524.
10. *Gebelein H.* Das statistische Problem der Korrelation als Variations- und Eigenwertproblem und sein Zusammenhang mit der Ausgleichsrechnung // Z. Angew. Math. Mech. 1941. V. 21. № 6. P. 364–379.
11. *Сарманов О.В.* Максимальный коэффициент корреляции (несимметричный случай) // Докл. АН СССР. 1958. Т. 121. № 1. С. 52–55.
12. *Rényi A.* On Measures of Dependence // Acta Math. Acad. Sci. Hungar. 1959. V. 10. № 3–4. P. 441–451.
13. *Witsenhausen H.S.* On Sequences of Pairs of Dependent Random Variables // SIAM J. Appl. Math. 1975. V. 28. № 1. P. 100–113.
14. *Добрушин Р.Л.* Центральная предельная теорема для неоднородных цепей Маркова. I // Теория вероятн. и ее примен. 1956. Т. 1. № 1. С. 72–89.
15. *Ahlswede R., Gács P.* Spreading of Sets in Product Spaces and Hypercontraction of the Markov Operator // Ann. Probab. 1976. V. 4. № 6. P. 925–939.
16. *Seneta E.* Non-negative Matrices and Markov Chains. New York: Springer, 1981.

17. *Cohen J.E., Iwasa Y., Răuţu G., Ruskai M.B., Seneta E., Zbăganu G.* Relative Entropy under Mappings by Stochastic Matrices // Linear Algebra Appl. 1993. V. 179. P. 211–235.
18. *Choi M.-D., Ruskai M.B., Seneta E.* Equivalence of Certain Entropy Contraction Coefficients // Linear Algebra Appl. 1994. V. 208/209. P. 29–36.
19. *Cohen J.E., Kemperman J.H.B., Zbăganu G.* Comparisons of Stochastic Matrices with Applications in Information Theory, Statistics, Economics and Population Sciences. Ann Arbor, MI, USA: Birkhäuser, 1998.
20. *Körner J., Marton K.* Comparison of Two Noisy Channels // Topics in Information Theory (2nd Colloq., Keszthely, Hungary, 1975). Colloq. Math. Soc. János Bolyai. V. 16. Amsterdam: North-Holland, 1977. P. 411–423.
21. *Csiszár I.* Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten // Magyar Tud. Akad. Mat. Kutató Int. Közl. Ser. A. 1963. V. 8. P. 85–108.
22. *Csiszár I.* Information-type Measures of Difference of Probability Distributions and Indirect Observations // Studia Sci. Math. Hungar. 1967. V. 2. P. 299–318.
23. *Ali S.M., Silvey S.D.* A General Class of Coefficients of Divergence of One Distribution from Another // J. Roy. Statist. Soc. Ser. B. 1966. V. 28. № 1. P. 131–142.
24. *Morimoto T.* Markov Processes and the H-Theorem // J. Phys. Soc. Japan. 1963. V. 18. № 3. P. 328–331.
25. *Akaike H.* Information Theory and an Extension of the Maximum Likelihood Principle // Proc. 2nd Int. Symp. on Information Theory. Tsaghkadzor, Armenia, USSR. Sept. 2–8, 1971. Budapest, Hungary: Akad. Kiadó, 1973. P. 267–281.
26. *Ziv J., Zakai M.* On Functionals Satisfying a Data-Processing Theorem // IEEE Trans. Inform. Theory. 1973. V. 19. № 3. P. 275–283.
27. *Zakai M., Ziv J.* A Generalization of the Rate-Distortion Theory and Applications // Information Theory: New Trends and Open Problems. New York: Springer, 1975. P. 87–123.
28. *Liese F., Vajda I.* On Divergences and Informations in Statistics and Information Theory // IEEE Trans. Inform. Theory. 2006. V. 52. № 10. P. 4394–4412.
29. *Levin D.A., Peres Y., Wilmer E.L.* Markov Chains and Mixing Times. Providence, RI, USA: Amer. Math. Soc., 2009.
30. *Kullback S., Leibler R.A.* On Information and Sufficiency // Ann. Math. Statist. 1951. V. 22. № 1. P. 79–86.
31. *Neyman J.* Contribution to the Theory of the χ^2 Test // Proc. 1st Berkeley Symp. on Mathematical Statistics and Probability. Berkeley, CA, USA. Aug. 13–18, 1945; Jan. 27–29, 1946. Berkeley, CA, USA: Univ. of California Press, 1949. P. 239–273.
32. *Nielsen F., Nock R.* On the Chi Square and Higher-Order Chi Distances for Approximating f-Divergences // IEEE Signal Process. Lett. 2014. V. 21. № 1. P. 10–13.
33. *Liese F., Vajda I.* Convex Statistical Distances. Leipzig: Teubner, 1987.
34. *Sason I., Verdú S.* f-Divergence Inequalities // IEEE Trans. Inform. Theory. 2016. V. 62. № 11. P. 5973–6006.
35. *Le Cam L.* Asymptotic Methods in Statistical Decision Theory. New York: Springer, 1986.
36. *Vincze I.* On the Concept and Measure of Information Contained in an Observation // Contributions to Probability: A Collection of Papers Dedicated to Eugène Lukacs. New York: Academic Press, 1981. P. 207–214.
37. *Györfi L., Vajda I.* A Class of Modified Pearson and Neyman Statistics // Statist. Decisions. 2001. V. 19. № 3. P. 239–251.
38. *Polyanskiy Y., Wu Y.* Lecture Notes on Information Theory. Dept. of Electrical Engineering and Computer Science, MIT, Cambridge, USA, 2017. Lect. Notes 6.441.
39. *Csiszár I.* A Class of Measures of Informativity of Observation Channels // Period. Math. Hungar. 1972. V. 2. № 1–4. P. 191–213.
40. *du Pin Calmon F., Makhdoumi A., Médard M., Varia M., Christiansen M., Duffy K.R.* Principal Inertia Components and Applications // IEEE Trans. Inform. Theory. 2017. V. 63. № 8. P. 5011–5038.

41. *Cover T.M., Thomas J.A.* Elements of Information Theory. Hoboken, NJ, USA: John Wiley & Sons, 2006.
42. *Borade S., Zheng L.* Euclidean Information Theory // Proc. IEEE Int. Zurich Seminar on Communications. Zurich, Switzerland. Mar. 12–14, 2008. P. 14–17.
43. *Huang S.-L., Zheng L.* Linear Information Coupling Problems // Proc. 2012 IEEE Int. Symp. on Information Theory (ISIT'2012). Cambridge, MA, USA. July 1–6, 2012. P. 1029–1033.
44. *Amari S., Nagaoka H.* Methods of Information Geometry. Providence, RI, USA: Amer. Math. Soc.; Oxford Univ. Press, 2000.
45. *Csiszár I., Shields P.C.* Information Theory and Statistics: A Tutorial. Hanover, MA, USA: Now Publ., 2005.
46. *Gohari A.A., Anantharam V.* Evaluation of Marton's Inner Bound for the General Broadcast Channel // IEEE Trans. Inform. Theory. 2012. V. 58. № 2. P. 608–619.
47. *Abbe E., Zheng L.* A Coordinate System for Gaussian Networks // IEEE Trans. Inform. Theory. 2012. V. 58. № 2. P. 721–733.
48. *Makur A.* A Study of Local Approximations in Information Theory: Master's Thesis. Dept. of Electrical Engineering and Computer Science, MIT, Cambridge, MA, USA, 2015.
49. *Horn R.A., Johnson C.R.* Matrix Analysis. New York: Cambridge Univ. Press, 2013.
50. *Greenacre M.J.* Theory and Applications of Correspondence Analysis. San Diego, CA, USA: Academic Press, 1984.
51. *Greenacre M., Hastie T.* The Geometric Interpretation of Correspondence Analysis // J. Amer. Statist. Assoc. 1987. V. 82. № 398. P. 437–447.
52. *Hsu H., Salamatian S., Calmon F.P.* Correspondence Analysis Using Neural Networks // Proc. 22nd Int. Conf. on Artificial Intelligence and Statistics (AISTATS'2019). Naha, Japan. April 16–18, 2019. P. 2671–2680.
53. *Lancaster H.O.* The Structure of Bivariate Distributions // Ann. Math. Statist. 1958. V. 29. № 3. P. 719–736.
54. *Lancaster H.O.* The Chi-Squared Distribution. New York: John Wiley & Sons, 1969.
55. *Makur A., Zheng L.* Polynomial Spectral Decomposition of Conditional Expectation Operators // Proc. 54th Annual Allerton Conf. on Communication, Control, and Computing. Monticello, IL, USA. Sept. 27–30, 2016. P. 633–640.
56. *Makur A., Zheng L.* Polynomial Singular Value Decompositions of a Family of Source-Channel Models // IEEE Trans. Inform. Theory. 2017. V. 63. № 12. P. 7716–7728.
57. *Breiman L., Friedman J.H.* Estimating Optimal Transformations for Multiple Regression and Correlation // J. Amer. Statist. Assoc. 1985. V. 80. № 391. P. 580–598.
58. *Makur A., Kozynski F., Huang S.-L., Zheng L.* An Efficient Algorithm for Information Decomposition and Extraction // Proc. 53rd Annual Allerton Conf. on Communication, Control, and Computing. Monticello, IL, USA. Sept. 29 – Oct. 2, 2015. P. 972–979.
59. *Golub G.H., van Loan C.F.* Matrix Computations Baltimore, MD, USA: The Johns Hopkins Univ. Press, 1996.
60. *Demmel J.W.* Applied Numerical Linear Algebra. Philadelphia, PA, USA: SIAM, 1997.
61. *Makur A.* Information Contraction and Decomposition: Sc.D. Thesis. Dept. of Electrical Engineering and Computer Science, MIT, Cambridge, MA, USA, 2019.
62. *Huang S.-L., Makur A., Kozynski F., Zheng L.* Efficient Statistics: Extracting Information from IID Observations // Proc. 52nd Annual Allerton Conf. on Communication, Control, and Computing. Monticello, IL, USA. Oct. 1–3, 2014. P. 699–706.
63. *Huang S.-L., Makur A., Wornell G.W., Zheng L.* On Universal Features for High-Dimensional Learning and Inference, `arXiv:1911.09105 [cs.LG]`, 2019.
64. *Pearson K.* On Lines and Planes of Closest Fit to Systems of Points in Space // Philos. Mag. 1901. V. 2. № 11. P. 559–572.
65. *Hotelling H.* Analysis of a Complex of Statistical Variables into Principal Components // J. Educ. Psychol. 1933. V. 24. № 6. P. 417–441; 498–520.

66. *Hotelling H.* Relations between Two Sets of Variates // Biometrika. 1936. V. 28. № 3/4. P. 321–377.
67. *Coifman R.R., Lafon S.* Diffusion Maps // Appl. Comput. Harmon. Anal. 2006. V. 21. № 1. P. 5–30.
68. *Asoodeh S., Diaz M., Alajaji F., Linder T.* Information Extraction under Privacy Constraints // Information. 2016. V. 7. № 1. Article no. 15 (37 pp.).
69. *Seneta E.* Coefficients of Ergodicity: Structure and Applications // Adv. in Appl. Probab. 1979. V. 11. № 3. P. 576–590.
70. *Ipsen I.C.F., Selee T.M.* Ergodicity Coefficients Defined by Vector Norms // SIAM J. Matrix Anal. Appl. 2011. V. 32. № 1. P. 153–200.
71. *Selee T.M.* Stochastic Matrices: Ergodicity Coefficients, and Applications to Ranking: Ph.D. Thesis. Dept. of Applied Mathematics, North Carolina State Univ., Raleigh, NC, USA, 2009.
72. *Kontorovich A.* Obtaining Measure Concentration from Markov Contraction // Markov Process. Related Fields. 2012. V. 18. № 4. P. 613–638.
73. *Yu B.* Assouad, Fano, and Le Cam // Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics. New York: Springer, 1997. P. 423–435.
74. *Shannon C.E.* The Zero Error Capacity of a Noisy Channel // IRE Trans. Inform. Theory. 1956. V. 2. № 3. P. 8–19.
75. *Csiszár I., Körner J.* Information Theory: Coding Theorems for Discrete Memoryless Systems. New York: Cambridge Univ. Press, 2011.
76. *Evans W.S., Schulman L.J.* Signal Propagation and Noisy Circuits // IEEE Trans. Inform. Theory. 1999. V. 45. № 7. P. 2367–2373.
77. *Goldstein S.* Maximal Coupling // Z. Wahrsch. Verw. Gebiete. 1979. V. 46. № 2. P. 193–204.
78. *Kim H., Gao W., Kannan S., Oh S., Viswanath P.* Discovering Potential Correlations via Hypercontractivity // Entropy. 2017. V. 19. № 11. Article no. 586 (32 pp.).
79. *Gilardoni G.L.* On Pinsker's and Vajda's Type Inequalities for Csiszár's f-Divergences // IEEE Trans. Inform. Theory. 2010. V. 56. № 11. P. 5377–5386.
80. *Verdú S.* Total Variation Distance and the Distribution of Relative Information // Proc. 2014 Information Theory and Applications Workshop (ITA'2014). San Diego, CA, USA. Feb. 9–14, 2014. P. 1–3.
81. *Nair C.* An Extremal Inequality Related to Hypercontractivity of Gaussian Random Variables // Proc. 2014 Information Theory and Applications Workshop (ITA'2014). San Diego, CA, USA. Feb. 9–14, 2014. P. 1–7.
82. *Calmon F.P., Polyanskiy Y., Wu Y.* Strong Data Processing Inequalities for Input Constrained Additive Noise Channels // IEEE Trans. Inform. Theory. 2018. V. 64. № 3. P. 1879–1892.
83. *Makur A., Polyanskiy Y.* Comparison of Channels: Criteria for Domination by a Symmetric Channel // IEEE Trans. Inform. Theory. 2018. V. 64. № 8. P. 5704–5725.
84. *van Dijk M.* On a Special Class of Broadcast Channels with Confidential Messages // IEEE Trans. Inform. Theory. 1997. V. 43. № 2. P. 712–714.
85. *Makur A., Polyanskiy Y.* Less Noisy Domination by Symmetric Channels // Proc. 2017 IEEE Int. Symp. on Information Theory (ISIT'2017). Aachen, Germany. June 25–30, 2017. P. 2463–2467.
86. *Carlen E.* Trace Inequalities and Quantum Entropy: An Introductory Course // Entropy and the Quantum. Providence, RI, USA: Amer. Math. Soc., 2010. P. 73–140.
87. *Bhatia R.* Matrix Analysis. New York: Springer, 1997.
88. *Samorodnitsky A.* On the Entropy of a Noisy Function // IEEE Trans. Inform. Theory. 2016. V. 62. № 10. P. 5446–5464.
89. *Kumar G.R., Courtade T.A.* Which Boolean Functions Are Most Informative? // Proc. 2013 IEEE Int. Symp. on Information Theory (ISIT'2013). Istanbul, Turkey. July 7–12, 2013. P. 226–230.
90. *El Gamal A., Kim Y.-H.* Network Information Theory. New York: Cambridge Univ. Press, 2011.

91. *Ordentlich E., Weinberger M.J.* A Distribution Dependent Refinement of Pinsker's Inequality // IEEE Trans. Inform. Theory. 2005. V. 51. № 5. P. 1836–1840.
92. *Sason I.* Bounds on f-Divergences and Related Distances // CCIT Report № 859. Haifa, Israel: Dept. of Electrical Engineering, Technion – Israel Inst. of Technology, 2014.
93. *Harremoës P., Vajda I.* On Pairs of f-Divergences and Their Joint Range // IEEE Trans. Inform. Theory. June 2011. V. 57. № 6. P. 3230–3235.
94. *Fedotov A.A., Harremoës P., Topsøe F.* Refinements of Pinsker's Inequality // IEEE Trans. Inform. Theory. 2003. V. 49. № 6. P. 1491–1498.
95. *Su F.E.* Methods for Quantifying Rates of Convergence for Random Walks on Groups: Ph.D. Thesis. Dept. of Mathematics, Harvard Univ., Cambridge, MA, USA, 1995.
96. *Dragomir S.S., Gluščević V.* Some Inequalities for the Kullback–Leibler and χ^2-Distances in Information Theory and Applications // Tamsui Oxf. J. Math. Sci. 2001. V. 17. № 2. P. 97–111.
97. *Sason I.* Tight Bounds for Symmetric Divergence Measures and a New Inequality Relating f-Divergences // Proc. 2015 IEEE Information Theory Workshop (ITW'2015). Jerusalem, Israel. April 26 – May 1, 2015. P. 1–5.
98. *Gibbs A.L., Su F.E.* On Choosing and Bounding Probability Metrics // Int. Stat. Rev. 2002. V. 70. № 3. P. 419–435.
99. *Csiszár I., Talata Z.* Context Tree Estimation for Not Necessarily Finite Memory Processes, via BIC and MDL // IEEE Trans. Inform. Theory. 2006. V. 52. № 3. P. 1007–1016.
100. *Boyd S., Vandenberghe L.* Convex Optimization. New York: Cambridge Univ. Press, 2004.
101. *Ash R.B.* Information Theory. New York: John Wiley & Sons, 1965.
102. *Dembo A., Cover T.M., Thomas J.A.* Information Theoretic Inequalities // IEEE Trans. Inform. Theory. 1991. V. 37. № 6. P. 1501–1518.
103. *Li Y.-C., Yeh C.-C.* Some Equivalent Forms of Bernoulli's Inequality: A Survey // Appl. Math. 2013. V. 4. № 7. P. 1070–1093.
104. *Sutter D., Renes J.M.* Universal Polar Codes for More Capable and Less Noisy Channels and Sources // Proc. 2014 IEEE Int. Symp. on Information Theory (ISIT'2014). Honolulu, HI, USA. June 29 – July 4, 2014. P. 1461–1465.
105. *Rakočević V., Wimmer H.K.* A Variational Characterization of Canonical Angles between Subspaces // J. Geom. 2003. V. 78. № 1. P. 122–124.
106. *Kang W., Ulukus S.* A New Data Processing Inequality and Its Applications in Distributed Source and Channel Coding // IEEE Trans. Inform. Theory. 2011. V. 57. № 1. P. 56–69.
107. *Banerjee A., Merugu S., Dhillon I.S., Ghosh J.* Clustering with Bregman Divergences // J. Mach. Learn. Res. 2005. V. 6. P. 1705–1749.

Макур Ануран
Чжэн Личжун
Отделение информационных технологий,
Массачусетский технологический институт, Кэмбридж, США
a_makur@mit.edu
lizhong@mit.edu

Поступила в редакцию
17.10.2019
После доработки
17.10.2019
Принята к публикации
09.03.2020