© 2022 г. В.Е. АНЦИПЕРОВ, канд. физ.-мат. наук
(antciperov@cplire.ru)
(Институт радиотехники и электроники
им. В.А. Котельникова РАН, Москва)

ГЕНЕРАТИВНАЯ МОДЕЛЬ АВТОКОДИРОВЩИКОВ, САМООБУЧАЮЩИХСЯ НА ИЗОБРАЖЕНИЯХ, ПРЕДСТАВЛЕННЫХ ВЫБОРКАМИ ОТСЧЕТОВ[1]

В работе обосновывается концепция автокодировщиков, ориентированных на автоматическую генерацию сжатых изображений. Предлагается решение задачи синтеза подобных автокодировщиков в контексте методов машинного обучения, понимаемого здесь как обучение по выборке из самих же данных. Для этих целей разработано специальное представление изображений с помощью выборок отсчетов контролируемого размера (выборочных представлений). Основываясь на специфике данного представления формализуется порождающая (генеративная) модель автокодировщиков, которая затем конкретизуется до вероятностной параметрической модели отсчетов в виде смеси компонент. На основе концепции рецептивных полей обсуждается редукция общей модели смеси компонент до сеточной модели финитных компонент экспоненциального семейства, допускающего синтез реалистических с вычислительной точки зрения алгоритмов кодирования.

Ключевые слова: машинное обучение, автокодировщики, генеративная модель, смеси распределений, EM алгоритм, рецептивные поля.

DOI: 10.31857/S0005231022120091, **EDN:** KSXBRK

1. Введение

Настоящее время характеризется как эпоха больших данных (Big Data). Большие данные, содержащиеся, например, в сети интернет, включают записи медицинских наблюдений, данные регистрации аудио-, фото- и видеопотоков, сканированные документы, графические презентации и т.д. Эти данные поступают из разных источников, поэтому характеризуются высокой разнородностью (variety). Сверх того, более 80% этих данных являются неструктурированными, что сильно затрудняет их передачу, хранение и использование. Большинство изображений в глобальной сети также относятся к категории неструктурированных данных [1]. Учитывая, что сеть интернет является крупнейшим мировым репозитарием изображений, можно только сожалеть о том, что невозможно в полной мере воспользоваться имеющимися ресурсами.

[1] Работа выполнена за счет бюджетного финансирования в рамках государственного задания в ИРЭ им. В.А. Котельникова РАН (ГЗ “РЭЛДИС”).

Существует несколько подходов к преодолению означенной выше проблемы больших данных [2]. Например, ввиду того, что исходные изображения часто имеют высокое качество, а для многих приложений этого не требуется, целесообразно передавать, хранить и использовать изображения в сжатом виде. Эта нехитрая мысль объединяет тем не менее огромный пласт практических исследований. Отметим, что объем этих исследований в последнее время только возрос в связи с замечательными успехами в областях нейронных сетей, глубокого обучения, искусственного интеллекта и т.д. (см. обзор [3]).

Вместе с тем, несмотря на целый ряд новых идей, привнесенных из области машинного обучения в практику сжатия изображений, большинство новых методов по-прежнему базируются на фундаментальных идеях Клода Шеннона, составляющих основу теории передачи данных с потерями (rate-distortion theory) [4]. Последняя, по сути, посвящена фундаментальному компромиссу между стремлением увеличить скорость передачи некоторого закодированного представления исходных данных и стремлением уменьшить искажения, обусловленные возможными потерями части информации в процессе сжатия / кодирования. Изначально полагалось, что цель теории состоит в том, чтобы найти такой способ кодирования исходных (входных) данных, который максимально минимизировал бы отклонение от них восстановленных (выходных) данных при заданной скорости передачи. Однако известно, что в области сжатия (кодирования) изображений минимизация отклонений сама по себе не обязательно ведет к хорошему качеству восприятия восстановленных изображений. Например, показано, что использование методов кодирования в генеративно-состязательных сетях может приводить к заметному улучшению качества восприятия изображения, хотя искажение исходного изображения может быть не минимальным [5]. В связи с этим в последнее время был предпринят ряд попыток включения в теорию кодирования изображений дополнительных элементов, повышающих результирующее качество их восприятия [6, 7].

Кардинальному пересмотру в новых подходах подверглись классические методы оценивания качества изображений с помощью функций искажения (distortion functions), которые определяются обычно как абсолютные или квадратичные отклонения восстановленной версии изображения от оригинала. Это связано с тем, что подобные простые функциональные метрики мало адекватны особенностям восприятия изображений человеком, и по этим причинам вплоть до недавнего времени визуальное качество изображений оценивалось специалистами, как правило, с помощью категориальных шкал, связанных с субъективными оценками в группах зрителей. По этой причине первые попытки корректировки теории были направлены на поиски тех нетрадиционных метрик, которые были бы объективно связаны с визуальным качеством. Основным требованием к таким метрикам была их высокая корреляция с известными категориальными оценками качества. Среди наиболее известных перцептивных метрик (perceptual metrics [2]) можно отметить метрику структурного подобия (SSIM) [8] / метрику многомасштабного струк-

турного подобия (MS-SSIM) [9], метрику на основе визуальной информации (VIF) [10], метрику на основе пространственного и временного наиболее очевидного искажения (MAD) [11].

Однако наибольших успехов в повышении качества восприятия восстановленных изображений удалось достичь не с помощью усовершенствования метрик искажений, а на пути пересмотра самой концепции искажений. Речь идет о генеративном моделировании (generative modeling), набирающем все большую популярность в машинном обучении [12]. Генеративные модели рассматривают всю совокупность (входных / выходных) данных как набор случайных переменных и в отличие от дискриминантных моделей (discriminative models) ориентированы на их совместные вероятностные распределения, нежели на условные. К генеративным моделям, в частности, относятся генеративно-состязательные сети (GAN) [13], вариационные автокодировщики (VAE) [14], глубокие сети доверия (DBN) [15] и т.д.

Практическая успешность генеративных моделей теоретически выяснена не полностью, однако, часто приходится слышать суждение о том, что это может быть связано с более адекватным моделированием особенностей естественного интеллекта [16]. Действительно, по мере развития перечисленных выше подходов [13–15] каждый новый этап разработки добавлял новые элементы, моделирующие особенности иерархической архитектуры коры, глубокого обучения с подкреплением, рабочей памяти в рекуррентных корковых сетях, долговременной памяти и т.д. В этой связи следует сделать особый акцент на роли представлений данных. Моделирование функций и структуры коры само в значительной степени предопределяют представления входных – промежуточных – выходных данных и их взаимосвязи. Вместе с тем существует ряд примеров, когда именно удачный выбор представлений данных позволил существенно повысить эффективность реализуемых на их основе функций [13–15]. Исходя из этого вопросы выбора представлений данных должны по-видимому играть не последнюю роль в разработке новых методов кодирования изображений, ориентированных на высокое качество восприятия.

В данной работе, основываясь на генеративной модели обучения без учителя [12], предлагается новый подход к задачам кодирования изображений на основе данных самих же изображений посредством рекуррентных автокодировщиков. Предлагаемый подход строится на основе ранее разработанного специального представления изображений (входных данных) с помощью выборки отсчетов контролируемого размера (выборочных представлений) [17, 18]. Поскольку в рамках генеративной модели полное статистическое описание выборочных представлений задается произведением плотностей распределения вероятностей отдельных отсчетов, в основе предлагаемого подхода лежит, по существу, статистическая проблема оценки плотностей распределения вероятностей (density estimation) [19]. В данной работе ограничимся классом параметрических процедур оценивания [19], что подразумевает задание некоторого параметрического семейства распределе-

ний вероятностей. В качестве последнего предлагается использовать модель смеси компонент из семейства экспоненциальных распределений (exponential family distributions) [20]. Соответственно в качестве кодированных изображений (выходных данных) используется совокупность оценок параметров компонент и их весов, вычисляемых по выборочному представлению (входным данным). В данном контексте оптимальное кодирование синтезируется на основе метода максимального правдоподобия. Показано, что решение уравнения максимального правдоподобия для модели смеси компонент приводит к рекуррентной процедуре кодирования, которая на каждой итерации, подобно популярному EM-алгоритму [21], включает два шага. На первом шаге (мягкого) разбиения выборки осуществляется стохастическая ассоциация отсчетов с компонентами смеси, а на втором производится пересчет оценок. Для целей алгоритмической (компьютерной) реализации процедуры кодирования вводится представление модели в виде системы рецептивных полей, представляющих собой локальные плотности распределения вероятностей отсчетов, покрывающих в совокупности все поле изображения. Предложенное представление позволяет значительно снизить требования алгоритма кодирования к вычислительным ресурсам. Кроме того, за счет некоторого упрощения представления, подобного используемому, например, в вариационном EM [22], процедура кодирования сводится к хорошо известным алгоритмам брегмановской мягкой / жесткой сегментации [23], LBG алгоритму векторного квантования [24], K-means кластеризации [25] и самоорганизующимся картам Кохонена [26]. Перечисленные взаимосвязи позволяют рассматривать предлагаемую процедуру как обобщение по нескольким направлениям хорошо известных методов машинного обучения. Теоретические аспекты предложенного подхода перцептивного сжатия изображений иллюстрируются по ходу статьи результатами компьютерного моделирования.

2. Представление изображений выборками случайных отсчетов

В нескольких предыдущих работах (см., например, [17, 18]) было предложено представление изображений наборами случайных отсчетов (photon-counting statistics [27]). Данное представление мотивировано, с одной стороны, прогрессом в развитии SPAD-матриц на основе однофотонных лавинных диодов [28], а с другой стороны, постоянно растущей тенденцией моделирования механизмов зрительного восприятия в задачах цифровой обработки изображений [29]. Отсчетное представление предполагает соответствующее устройство формирования изображения (photon counting imaging system [27]), содержащее большое число точечных детекторов, причем предполагается, что последние настолько малы, что каждый из них может регистрировать одиночные фотоны падающего излучения.

Формально под идеальным устройством формирования изображения будем понимать плоскую 2D-поверхность Ω, на которой вплотную друг к другу расположены идентичные точечные детекторы, или “джоты” в терминах [28],

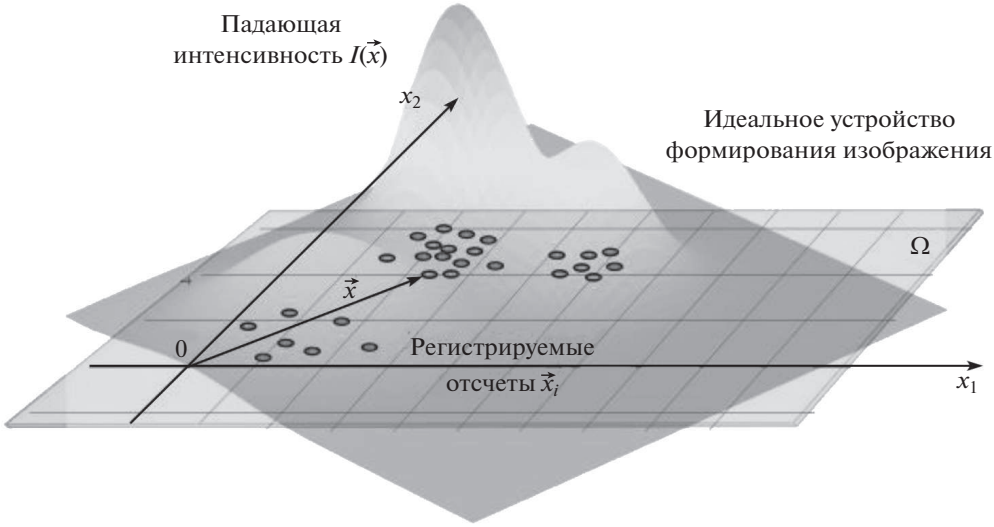

Рис. 1. Идеальное устройство формирования изображения и результат регистрации падающего излучения интенсивности $I(\vec{x})$ в виде набора фотоотсчетов $X = (\vec{x}_1, \ldots, \vec{x}_n)$.

(см. рис. 1), имеющие малую площадь ds светочувствительной поверхности. Соответственно, общее количество детекторов будет $N = S/ds$, где S — общая площадь поверхности Ω. Идеально, в предположении, что S фиксирована, а $ds \to 0$, число детекторов $N \to \infty$. Таким образом, идеальное устройство формирования изображения — это почти непрерывная чувствительная поверхность Ω с координатами $\vec{x} = (x_1, x_2)$, задающими положения идеальных точечных детекторов, как это представлено на рис. 1.

Регистрация падающего на Ω излучения интенсивности $I(\vec{x}, t)$ проявляется, согласно современным физическим представлениям [30], в происходящих на микроуровне случайных событиях, связанных с поглощением фотонов. Эти события, в зависимости от типа используемых фотодетекторов, могут состоять в высвобождении фотоэлектронов из p–n переходов полупроводниковых детекторов, в изменении конформации молекул родопсина рецепторов сетчатки, в образовании атомов металлического Ag из галогенидов серебра в фотопленках и т.д. Все подобные события обычно объединяются общим термином фотоотсчеты, или просто отсчеты (по поводу терминологии см. [27]). В модели идеального устройства формирования изображений отсчеты являются событиями случайными в смысле классической теории вероятностей и задаются пространственными координатами — случайными векторами $\vec{x} \in \Omega$ (совпадающими с координатами соответствующих точечных детекторов) и случайными же моментами времени регистрации $\overline{t} \in (t, t + dt)$. В полуклассическом приближении [30] вероятность отсчета за время $dt \to 0$ равна $P_1(\vec{x}) = \alpha I(\vec{x}, \overline{t}) ds\, dt$, где $\alpha = \eta(h\overline{\nu})^{-1}$, $h\overline{\nu}$ — средняя энергия регистрируемого фотона (h — постоянная Планка, $\overline{\nu}$ — характерная частота излуче-

ния), безразмерный коэффициент $\eta < 1$ является квантовой эффективностью материала детектора.

На основе модели идеального устройства формирования изображений можно сформулировать модель идеального изображения [17, 18]. Под идеальным изображением понимается (упорядоченный) набор $X = (\vec{x}_1, \ldots, \vec{x}_n)$ n случайных отсчетов, регистрируемых идеальным устройством в течение времени экспозиции T. Таким образом, идеальное изображение представляет собой даже при заданной (детерминированной) интенсивности $I(\vec{x}, t)$ случайный объект. При этом случайный характер идеального изображения определяется не только случайными координатами $\vec{x}_i$ отсчетов, но и случайным числом n их общего количества в наборе X.

Полное статистическое описание идеальных изображений в виде всех конечномерных плотностей распределений $\rho(\vec{x}_1, \ldots, \vec{x}_n, n | I(\vec{x}, \overline{t}))$ может быть получено в предположении условной независимости отсчетов $\vec{x}_i$ (при заданной регистрируемой интенсивности $I(\vec{x})$). Стандартный вывод статистического описания, опирающегося на пуассоновскую аппроксимацию при $ds \to 0,\ N \to \infty,\ Nds = S = \text{const}$ совместной вероятности большого ансамбля бернулиевских точечных детекторов (успех — регистрация фотоотсчета с вероятностью $P_1(\vec{x})$, неудача — отсутствие регистрации с вероятностью $P_0(\vec{x}) = 1 - P_1(\vec{x})$), можно найти, например, в [31]. Приведем здесь лишь окончательный результат, отвечающий стационарному случаю ($I(\vec{x}, \overline{t}) = I(\vec{x})$):

$$\begin{gathered}\rho(\vec{x}_1, \ldots, \vec{x}_n, n \mid I(\vec{x})) = \rho(\vec{x}_1, \ldots, \vec{x}_n \mid n, I(\vec{x})) \times P_n(I(\vec{x})) = \\ = \prod_{i=1}^{n} \rho(\vec{x}_i \mid I(\vec{x})) \times P_n(I(\vec{x})), \\ P_n(I(\vec{x})) = \frac{\overline{n}^n}{n!} \exp(-\overline{n}), \quad \overline{n} = \alpha T \iint\limits_{\Omega} I(\vec{x}) ds,\end{gathered} \tag{1}$$

где $\overline{n}$ — среднее число всех отсчетов, $\rho(\vec{x}_i \mid I(\vec{x}))$ — плотность вероятности отсчета:

$$\rho(\vec{x}_i \mid I(\vec{x})) = \frac{I(\vec{x})}{\iint_{\Omega} I(\vec{x}) ds}. \tag{2}$$

Отметим, что полное статистическое описание (1), (2) совпадает с распределением вероятностей событий 2D-точечного пуассоновского процесса на Ω [31], интенсивность событий которого $\lambda(\vec{x})$ с точностью до αT совпадает с интенсивностью $I(\vec{x})$ зарегистрированного излучения.

Из (2) следует, что плотность распределения вероятностей отсчета $\vec{x} \in \Omega$ при регистрации излучения идеальным устройством формирования изображений совпадает с нормированной интенсивностью $I(\vec{x})$ этого излучения на Ω. Отметим в этой связи универсальный характер (2): условная плотность

распределения вероятностей не зависит ни от квантовой эффективности материала детекторов η, ни от спектра (в том числе характерной частоты $\overline{\nu}$ излучения), ни от времени экспозиции T. Более того, она не зависит и от единиц измерения $I(\vec{x})$, будь то абсолютные физические единицы или условные, полученные в результате оцифровки.

Модель идеального изображения и статистическое описание (1), (2) часто используются при низких интенсивностях $I(\vec{x})$ излучений, например, в областях флуоресцентной микроскопии, позитронно-эмиссионной томографии (ПЭТ), однофотонной эмиссионной компьютерной томографии (ОФЭКТ), оптической и инфракрасной астрономии и т.д. [32]. Однако при обычных интенсивностях, соответствующих, например, дневному свету, практическое использование модели идеального изображения оказывается проблематичным. Дело в том, что потоки фотонов, например, от солнца в ясный день огромны — на площадке $S \sim 1mm^2$ на поверхности земли они составляют $\overline{n} \sim 10^{15}$–$10^{16}$ фотонов/с. Очевидно, что работа с такими потоками данных потребует в практических задачах слишком больших ресурсов. Поэтому для задач представления изображений с помощью отсчетов желательно скорректировать приведенную выше модель идеального изображения.

Некоторое время назад было предложено следующее решение проблемы отсчетных представлений [17]. Зафиксируем с самого начала некоторое приемлемое значение размеров представления $k \ll \overline{n}$ и, рассматривая идеальное изображение $X = (\vec{x}_1, \ldots, \vec{x}_n)$ как некоторую генеральную совокупность случайных отсчетов, осуществим из нее, в полном соответствии с традициями классической статистики, случайную выборку в k отсчетов $X_k = (\vec{x}_{j_1}, \ldots, \vec{x}_{j_k})$. Очевидно, такое "выборочное" представление попрежнему, хотя и при гораздо меньшем размере данных, представляет исходное изображение X. Назовем X_k представлением изображения выборкой случайных отсчетов или, короче, выборочным представлением. Статистическое описание выборочных представлений легко следует из (1), (2) в результате интегрирования $\rho(\vec{x}_1, \ldots, \vec{x}_n), n \mid I(\vec{x})$ по невыбранным в X_k отсчетам и суммирования результатов по их числу $l = n - k = 0, 1, \ldots$:

$$\rho(X_k \mid I(\vec{x})) = \prod_{j=1}^{k} \rho(\vec{x}_j \mid I(\vec{x})) \times P_{n \geqslant k}(I(\vec{x})),$$

$$P_{n \geqslant k}(I(\vec{x})) = \sum_{l=0}^{\infty} P_{k+l}(I(\vec{x})), \tag{3}$$

где $P_{n \geqslant k}(I(\vec{x}))$ обозначает вероятность того, что в идеальном изображении содержится более, чем k отсчетов. Учитывая пуассоновский характер вероятностей P_n и их асимптотическое стремление при $n \to \infty$ к гауссовому распределению со средним $\overline{n}$, несложно показать, что в случае $\overline{n} \gg 1$ вероятность $P_{n<k}(I(\vec{x}))$ будет меньше ε, как только $k < 2\varepsilon\overline{n}$ и, соответственно, $P_{n \geqslant k}(I(\vec{x}))$ будет отличаться от единицы менее, чем на ε. Считая далее, что для пред-

ставления изображений выборками случайных отсчетов X_k размеры последних удовлетворяют $k \ll \overline{n}$ и полагая $P_{n \geqslant k}(I(\vec{x})) = 1$, с точностью до малого $\varepsilon = k/2\overline{n}$, получим из (3):

$$\rho(X_k \mid I(\vec{x})) = \prod_{j=1}^{k} \rho(\vec{x}_j \mid I(\vec{x})). \tag{4}$$

Заметим, что (4) формально следует из (2) в предположении, что идеальное устройство формирования изображений при регистрации излучения интенсивности $I(\vec{x})$ практически наверное содержит (гораздо) больше отсчетов, чем размер выборочного представления k. Именно, предполагая в этих обстоятельствах условную независимость произвольных k отсчетов $\vec{x_j} \in \Omega$, придем сразу же к (4), перемножив плотности распределений (2).

Статистическое описание (4) для выборочных представлений $X_k = (\vec{x}_1, \ldots, \vec{x}_k)$ имеет ряд привлекательных свойств (новая нумерация соответствует номерам выбранных отсчетов, а не отсчетам идеального изображения X, как это было выше). Это описание, во-первых, фиксирует условную независимость и одинаковое условное распределение (iid свойство) всех k отсчетов $\{\vec{x}_j\}$. Во-вторых, плотности распределений отдельных отсчетов $\rho(\vec{x}_j \mid I(\vec{x}))$ кратны распределению интенсивности $I(\vec{x})$ по поверхности Ω изображения (см. (2)). И, в-третьих, как это было отмечено выше, описание (4) также обладает свойством универсальности — не зависит ни от квантовой эффективности материала детекторов η, ни от спектра излучения, ни от времени экспозиции T. Таким образом, перечисленные свойства выборочных представлений осуществляют удобную форму представления входных данных для многих хорошо разработанных статистических подходов и методов машинного обучения, включая наивный байесов подход.

Более того, ввиду независимости распределений $\rho(\vec{x}_j \mid I(\vec{x}))$ в (4) от абсолютных значений интенсивности (см. (2)) статистические описания выборочных представлений X_k также не зависят от единиц измерения интенсивности $I(\vec{x})$. В частности, если интенсивность зарегистрированного излучения задана пикселами $\{m_i\}$ некоторого цифрового изображения, описание (4) не зависит от величины порога квантования $Q = \Delta I$, а будет определяться только битовой глубиной пикселов υ (или числом уровней квантования 2^{υ}).

В связи с последним замечанием отметим, что процедуру формирования выборочного представления для цифровых изображений можно по существу свести к нормировке $\pi_i = m_i/\Sigma m_i$ значений пикселей $m_i \sim I_i/Q$ изображения и последующему семплированию k отсчетов из полученного распределения вероятностей $\rho(\vec{x}_j \mid I(\vec{x})) \approx \pi_i$. Если учесть, что в области машинного обучения существует целый арсенал методов семплирования, объединенных общим названием методов Монте-Карло [33], то формирование выборочных представлений цифровых изображений можно с алгоритмической точки зрения рассматривать как применение стандартных процедур.

Рис. 2. Представление изображения “Cameraman” выборками случайных отсчетов: *а* — исходное изображение в формате TIFF; *б*, *в*, *г* — представления выборками размеров, соответственно, 500 000, 1 000 000, 5 000 000 отсчетов.

Для примера на рис. 2 приведены представления выборками случайных отсчетов стандартного тестового изображения “Cameraman”, широко используемого в публикациях по обработке изображений. Изображение “Cameraman” задано изначально в формате TIFF, имеет размеры $s \times s = 512 \times 512$ пикселей (72 dpi), серое, с глубиной цвета $\upsilon = 8$ бит. Семплирование выборочных представлений в $k = 500\,000, 1\,000\,000, 5\,000\,000$ отсчетов осуществлялось одним из самых простых методов — семплированием с отклонением (rejection sampling [33]) при равномерном вспомогательном распределении $g(\vec{x}) = 1/s \times s = 512^{-2}$ и верхнеграничной константе $M = 2^{\upsilon}/\overline{m} = 256/\overline{m}$, где $\overline{m} = \Sigma m_i/s \times s$ — среднее значение пикселей изобра-

жения. Алгоритмическая реализация приведенной процедуры сводится, таким образом, к случайному выбору равномерно распределенных на поверхности изображения Ω с координатами (x_1, x_2) — числами с плавающей точкой случайных векторов $\vec{x}_j$ и включения их в выборку отсчетов X_k при выполнении теста $u_j < m_i$, где i — индекс содержащего $\vec{x}_j$ пикселя изображения, а u_j — реализация случайной, равномерно распределенной на $(0, 2^v)$ величины. Отметим, что для этой процедуры нормировка пикселей не требуется.

В связи с приведенным примером еще раз подчеркнем различие в природе цифровых изображений и их выборочных представлений. Всякое цифровое изображение, например растровое (bitmap), является детерминированным объектом, представляющим собой раз и навсегда зафиксированную реализацию (вообще говоря случайного) процесса регистрации излучения. Любая обработка такого изображения — фильтрация, децимация, выравнивание гистограммы и т.д. всегда будет приводить при заданном алгоритме к одному и тому же результату. Наоборот, как отмечалось выше, выборочные представления $X_k = (\vec{x}_1, \ldots, \vec{x}_k)$ являются случайными объектами и каждая новая процедура их формирования будет давать реализацию, слегка отличную от предыдущих. Соответственно, обработка разных реализаций приведет к несколько отличающимся результатам. Трактовать эти отличия следует статистически, как это делается, например, в теории статистического оценивания. В частности, при увеличении размера выборочного представления k следует ожидать, что относительные отклонения по реализациям будут стремиться к нулю, демонстрируя проявление закона больших чисел. Некоторое представление об асимптотическом поведении выборочных представлений дает рис. 2.

3. Кодирование и декодирование изображений, представленных выборками случайных отсчетов

В области машинного обучения автокодировщики (автоэнкодеры, АЭ) рассматриваются как особый класс искусственных нейронных сетей [34], но для целей данной работы желательно определить их с более общей точки зрения. А именно, будем рассматривать АЭ как некоторый класс информационных систем, понимая под последними интегрированные системы компонент для сбора, хранения и обработки данных. В контексте настоящей работы под данными естественно подразумеваются изображения. Традиционно АЭ имеют симметричную трехуровневую структуру вход — внутреннее представление (код) — выход, симметрия подразумевает подобие выхода входу.

Пары соседних уровней составляют в АЭ две взаимные компоненты: кодер, включающий уровни вход–код, и декодер, включающий уровни код–выход [35]. Цель АЭ — восстановить данные на выходе, соблюдая при этом определенные ограничения, накладываемые на внутреннюю кодировку. Ввиду имеющихся ограничений не разрешается просто копировать данные со ввода на выход. Типичные ограничения связаны с уменьшением размерно-

сти промежуточных данных $\vec{z}$. В свете подходов машинного обучения кодирование $\vec{z}$ осуществляется на основе обучения АЭ без учителя (unsupervised learning) [35]. Отметим, что термин "автокодировщик" часто используется как синоним автоассоциативных, репликативных и др. нейронных сетей.

Для конкретизации АЭ в контексте данной работы возьмем за основу формально-математическую структуру абстрактного АЭ [36]. Структура АЭ включает множество G возможных изображений, полученных при регистрации излучения интенсивности $I(\vec{x})$ и множество F — их внутренних (кодовых) представлений. Также она включает классы операторов $f : G \to F$ (кодеры) и $g : F \to G$ (декодеры), согласованные по размерностям с G и F и с заданными ограничениями. Кроме того, структура АЭ предполагает количественную меру $D(I(\vec{x}), I_r(\vec{x}))$ расхождения (искажения) между изображением $I(\vec{x})$ на входе и некоторым его восстановленным вариантом $I_r(\vec{x})$ на выходе. Обычно эту меру называют функцией потерь [35]. В рамках приведенной структуры задачей АЭ является минимизация функции потерь по отношению к операторам f и g

$$\{f^*, g^*\} = arg\min_{f,g} D\big(I(\vec{x}), g \circ f(I(\vec{x}))\big). \tag{5}$$

Любое решение f^* уравнения (5) рассматривается как желаемое кодирование для последующего оптимального восстановления g^* изображения. К сожалению, решение (5) в общем виде — задача нереальная. Поэтому при изучении практических задач необходимо конкретизировать некоторые элементы общей структуры АЭ. Различные виды АЭ могут быть получены в зависимости от выбора множеств G и F, специальных классов операторов f и g, явного вида функции потерь D, а также наличия дополнительных ограничений, таких как гладкость, размерность и пр.

В рамках данной работы входными данными АЭ являются k-выборки случайных отсчетов — выборочные представления $X_k = (\vec{x}_1, \ldots, \vec{x}_k)$, порожденные плотностью распределения вероятностей координат отсчетов $\rho(\vec{x} \mid I(\vec{x}))$, кратной регистрируемой интенсивности $I(\vec{x})$ (2). Таким образом, вполне разумно выбрать в качестве множества входных изображений G множество плотностей распределения $\{\rho(\vec{x} \mid I(\vec{x}))\}$, $\vec{x} \in \Omega$. Это автоматически подводит к порождающим (генеративным) моделям АЭ [35]. В отличие от дискриминантных АЭ, которые наиболее естественно интерпретировать как регуляризирующие операторы, автокодировщики в генеративной парадигме рассматривают внутренние данные (код) F как латентные переменные, а операцию кодирования как процедуру статистического вывода (нахождения кодовых представлений по заданному выборочному представлению X_k). В связи с этим генеративные модели учатся скорее восстанавливать по обучающим выборочным наблюдениям (выборочному представлению) X_k порождающие их вероятностные распределения, нежели отображать входные данные в выходные. В качестве примеров генеративных подходов отметим вариационные автокодировщики (VAE) [14] и генеративные стохастические сети (GSN) [37].

Чтобы формализовать порождающую модель для выборочных представлений изображений $X_k = (\vec{x}_1, \ldots, \vec{x}_k)$, рассмотрим множество G как некоторое параметрическое семейство $G = \{\rho(\vec{x} \mid \vec{\theta}\}$, $\vec{x} \in \Omega$, $\vec{\theta} \in \Theta \subset \mathbb{R}^p$ плотностей распределения вероятностей отсчета $\vec{x}$. Параметризация G является распространенным подходом, упрощающим общую задачу функциональной оптимизации (5) до задачи оценки оптимальных параметров. Обучение без учителя для генеративной модели АЭ заключается в нахождении тех оптимальных параметров $\vec{\theta}^* \in \Theta$, которые определяют выход АЭ в виде плотности $\rho(\vec{x} \mid \vec{\theta}^*) \in G$ наилучшим образом соответствующей представлению X_k как выборке данных на входе.

Формально представление X_k не является элементом G и не может рассматриваться как вход для АЭ, задаваемый некоторыми параметрами $\vec{\theta}$ (посредством $\rho(\vec{x} \mid \vec{\theta}) \in G$). Однако, рассуждая в духе байесовского подхода, предполагая некоторое распределение параметров $\rho(\vec{\theta} \mid X_k)$ при заданной выборке X_k, можно выразить связанное с обучением вероятностное распределение в виде:

$$\rho\big(\vec{x} \mid X_k = (\vec{x}_1, \ldots, \vec{x}_k)\big) = \iint\limits_{\Theta} \rho(\vec{x} \mid \vec{\theta})\rho(\vec{\theta} \mid X_k)d\vec{\theta}. \tag{6}$$

Далее, согласно (П.8), из Приложения плотность распределения апостериорной вероятности $\rho(\vec{\theta} \mid X_k)$ является, по крайней мере асимптотически, при $k \gg 1$ более узкой функцией θ, чем $\rho(\vec{x} \mid \vec{\theta})$. Поскольку максимум $\rho(\vec{\theta} \mid X_k)$ приходится при этом на оценку максимального правдоподобия $\vec{\theta}_{ML}$, плотность $\rho(\vec{x} \mid \vec{\theta})$ может быть вынесена из последнего интеграла в (6) как $\rho(\vec{x} \mid \vec{\theta}_{ML})$, что приводит к следующему соотношению:

$$\rho\big(\vec{x} \mid X_k = (\vec{x}_1, \ldots, \vec{x}_k)\big) \cong \rho\big(\vec{x} \mid \vec{\theta}_{ML}\,(\vec{x}_1, \ldots, \vec{x}_k)\big). \tag{7}$$

Тем самым, предполагая что на входе имеется некоторая плотность распределения $\rho\big(\vec{x} \mid \vec{\theta}\big)$, задачей АЭ является формирование таких параметров $\vec{\theta}^*$, максимально близких к $\vec{\theta}_{ML}$, которые бы на выходе дали плотность $\rho\big(\vec{x} \mid \vec{\theta}^*\big)$ максимально подобную, в соответствии с (7), входной. Другими словами, задача автокодировщика в генеративной парадигме сводится к решению уравнений максимального правдоподобия Р. Фишера [38]:

$$\vec{\theta}_{ML} = arg\max_{\vec{\theta} \in \Theta} L(\theta; X_k),$$

$$L(\theta; X_k) = \ln\big(\rho(X_k \mid \vec{\theta})\big) = \ln\left(\prod_{j=1}^{k} \rho(\vec{x}_j \mid \vec{\theta})\right) = \sum_{j=1}^{k} \ln\big(\rho(\vec{x}_j \mid \vec{\theta})\big), \tag{8}$$

где традиционно использована функция логарифмического правдоподобия выборочного представления $L(\theta; X_k)$, которая представляет собой ввиду (4)

сумму функций логарифмического правдоподобия по всем отсчетам выборки X_k. Отметим, что формально (8) можно получить, если использовать в качестве функции потерь $D(\ldots)$ (5) дивергенцию Кульбака–Лейблера [39] между эмпирическим распределением $\rho(\vec{x} \mid X_k)$ (6) и параметрическим $\rho(\vec{x} \mid \vec{\theta}) \in G$ [19].

Кажущаяся элегантность сформулированной основной задачи АЭ (8) в рамках генеративной модели связана с переносом акцента с проблемы кодирования входных данных на проблему вычисления оценок максимального правдоподобия. Последняя задача, известная более ста лет, начиная с основополагающих работ Р. Фишера [38], в реальных приложениях может оказаться не проще (с вычислительной точки зрения), чем синтез дискриминантных кодировщиков. По этим причинам сделаем еще один шаг в уточнения генеративной модели АЭ, основанной на специфических особенностях архитектуры, связанных с внутренними переменными $\vec{z} \in F$, составляющими промежуточное представление данных на входе. Для этих целей уточним генеративную модель автокодировщиков $G = \{\rho(\vec{x} \mid \vec{\theta})\}$ как модель смеси компонент, в которой внутренние (скрытые) переменные $\vec{z}$ возникают естественным образом.

4. Генеративная модель автокодировщика в виде семейства смеси компонент

Возьмем в качестве параметрического семейства

$$G = \{\rho(\vec{x} \mid \vec{\theta}\}, \quad \vec{x} \in \Omega, \quad \vec{\theta} \in \Theta \subset \mathbb{R}^p$$

плотностей распределения вероятностей отсчета $\vec{x}$ семейство смесей из K компонент [20] вида

$$\rho(\vec{x} \mid \vec{\theta}) = \sum_{i=1}^{K} w_i \rho_i(\vec{x} \mid \vec{\theta}), \tag{9}$$

где $\{w_i\}$, $w_i \geqslant 0$, $i = 1, \ldots, K$ являются нормированными весами компонент в смеси и связаны условием $\sum\limits_{i=1}^{K} w_i = 1$, а $\{\rho_i(\vec{x} \mid \vec{\theta})\}$, $\rho_i(\vec{x} \mid \vec{\theta}) \geqslant 0$, $i = 1, \ldots, K$ являются плотностями распределения вероятностей $\vec{x} \in \Omega$ для каждой из компонент. С точки зрения статистики, смесь (9) формально можно интерпретировать как маргинальное распределение по $\vec{x}$, если для всех отсчетов наряду с координатами $\vec{x}$ ввести случайные скрытые (латентные) целочисленные переменные $z \in \{1, \ldots, K\}$, ассоциирующие отсчет с компонентой $i = z$ смеси. При этом, интерпретируя веса $\{w_i\}$ как априорное распределение вероятностей z, а плотности $\rho_i(\vec{x} \mid \vec{\theta})$ как условные распределения $\rho(\vec{x} \mid z = i, \vec{\theta})$, слагаемые $w_i \rho_i(\vec{x} \mid \vec{\theta}) = w_i \rho(\vec{x} \mid z = i, \vec{\theta}) = \rho(\vec{x}, z = i \mid \vec{\theta})$ в (9) можно интерпретировать как совместные распределения вероятностей “полных” наборов дан-

ных $\{\vec{x}, z\}$, а $\rho(\vec{x} \mid \vec{\theta})$ как маргинальное по отношению к нему распределение

$$\begin{aligned} \rho(\vec{x}, z \mid \vec{\theta}) &= w_z \rho_z(\vec{x} \mid \vec{\theta}), \\ \rho(\vec{x} \mid \vec{\theta}) &= \sum_{z=1}^{K} \rho(\vec{x}, z \mid \vec{\theta}). \end{aligned} \tag{10}$$

Обычно априорные вероятности $\{w_i\}$ являются частью множества параметров $\vec{\theta}$ смеси. В отношении же оставшихся параметров предполагается, что они могут быть разбиты на K непересекающихся групп $\{\vec{\nu}_i\}$, $i = 1, \ldots, K$, $\vec{\nu}_i \in \Xi \subset \mathbb{R}^q$ так, что от параметров z-й группы, и только от них зависят условные распределения $\rho_z(\vec{x} \mid \vec{\theta}) = \rho_z(\vec{x} \mid \vec{\nu}_z)$. Таким образом, полный набор параметров модели $\vec{\theta} = \{\{w_i\}, \{\vec{\nu}_i\}\} \in \Theta \subset \mathbb{R}^p$, $i = 1, \ldots, K$, $p = K(q+1)$ рассматривается как объединение наборов весов $\{w_i\}$ и параметров $\{\vec{\nu}_i\}$ компонент. Отметим, что в $\vec{\theta}$ по крайней мере часть параметров — $\{w_i\}$ зависимы — связаны условием нормировки. В сделанных предположениях модель смеси (10) уточняется посредством $\rho(\vec{x}, z \mid \vec{\theta}) = w_z \rho_z(\vec{x} \mid \vec{\nu}_z)$.

С учетом (10) задача нахождения максимума функции правдоподобия (8) для смесей принимает вид (в предположении, что помимо условия нормировки для $\{w_i\}$ другие ограничения на параметры отсутствуют):

$$\begin{aligned} \vec{\nabla}_{\vec{\theta}} L(\vec{\theta}^*; X_k) &= \sum_{j=1}^{k} \vec{\nabla}_{\vec{\theta}} \ln\bigl(\rho(\vec{x}_j \mid \vec{\theta}^*)\bigr) = \\ &= \sum_{j=1}^{k} \frac{1}{\rho(\vec{x}_j \mid \vec{\theta}^*)} \sum_{z_j=1}^{K} \vec{\nabla}_{\vec{\theta}} \rho(\vec{x}_j, z_j \mid \vec{\theta}^*) = \\ &= \sum_{j=1}^{k} \sum_{z_j=1}^{K} \frac{\rho(\vec{x}_j, z_j \mid \vec{\theta}^*)}{\rho(\vec{x}_j \mid \vec{\theta}^*)} \vec{\nabla}_{\vec{\theta}} \ln\bigl(\rho(\vec{x}_j, z_j \mid \vec{\theta}^*)\bigr) = \\ &= \sum_{j=1}^{k} \sum_{z_j=1}^{K} \rho(z_j \mid \vec{x}_j, \vec{\theta}^*) \vec{\nabla}_{\vec{\theta}} \ln\bigl(\rho(\vec{x}_j, z_j \mid \vec{\theta}^*)\bigr) = \lambda \vec{\Im}, \end{aligned} \tag{11}$$

где $\rho(z \mid \vec{x}, \vec{\theta}) = \rho(\vec{x}, z \mid \vec{\theta}) / \rho(\vec{x} \mid \vec{\theta})$ — апостериорная вероятность скрытых переменных z, а вектор $\vec{\Im}$ является индикатором частных производных по параметрам–весам: он состоит из нулей, кроме единиц, на местах соответствующих $\partial L(\vec{\theta}^*; X_k)/\partial w_i$ в левой части (11). Кратный $\vec{\Im}$ вектор $\lambda\vec{\Im}$ возникает при применении метода неопределенных множителей Лагранжа для условной максимизации $L(\vec{\theta}; X_k)$ с ограничением нормировки на $\{w_i\}$. Соответственно неопределенный множитель λ находится из условия $\sum\limits_{i=1}^{K} w_i^*(\lambda) = 1$.

Уравнения максимального правдоподобия (11) содержат в левой части линейную комбинацию градиентов логарифмических функций правдоподобия отсчетов (функций вкладов — score functions) $\vec{\nabla}_{\vec{\theta}} l(\vec{\theta}; \vec{x}_j, z_j) =$

$= \vec{\nabla}_{\vec{\theta}} \ln\bigl(\rho(\vec{x}_j, z_j \mid \vec{\theta})\bigr)$, которую обычно проще анализировать, чем саму логарифмическую функцию правдоподобия $L(\vec{\theta}; X_k)$. Действительно, используя $\rho(\vec{x}_j, z_j \mid \vec{\theta}) = w_{z_j} \rho_{z_j}(\vec{x}_j \mid \vec{\nu}_z)$ (см. (10) и далее), получим

$$\frac{\partial l(\vec{\theta}^*; \vec{x}_j, z_j)}{\partial w_z} = \frac{\delta_{z_j z}}{w^*_{z_j}},$$
$$\vec{\nabla}_{\vec{\nu}_z} l(\vec{\theta}^*; \vec{x}_j, z_j) = \vec{\nabla}_{\vec{\nu}_{z_j}} \ln\bigl(\rho_{z_j}(\vec{x}_j \mid \vec{\nu}^*_{z_j})\bigr)\delta_{z_j z}, \tag{12}$$

где $\delta_{z_j z}$ — символ Кронекера.

Подстановка (12) в (11) упрощает уравнения максимального правдоподобия до:

$$\frac{\partial L(\vec{\theta}^*; X_k)}{\partial w_z} = \frac{\sum\limits_{j=1}^{k} \rho(z \mid \vec{x}_j, \vec{\theta}^*)}{w^*_z} = \lambda,$$
$$\vec{\nabla}_{\vec{\nu}_z} L(\vec{\theta}^*; X_k) = \sum_{j=1}^{k} \rho(z \mid \vec{x}_j, \vec{\theta}^*) \vec{\nabla}_{\vec{\nu}_z} \ln\bigl(\rho_z(\vec{x}_j \mid \vec{\nu}^*_z)\bigr) = \vec{0}. \tag{13}$$

Домножая уравнения, содержащие λ каждое на w^*_z и суммируя их по z, получим, вследствие условий нормировки $\{w^*_i\}$ и апостериорных распределений $\rho(z \mid \vec{x}, \vec{\theta})$, что значением неопределенного множителя является $\lambda = k$. Это позволяет освободиться от λ в уравнениях (13) и придать им следующую компактную форму: 0

$$z = 1, \ldots, K,$$
$$w^*_z = \frac{1}{k} \sum_{j=1}^{k} \rho(z \mid \vec{x}_j, \vec{\theta}^*),$$
$$\sum_{j=1}^{k} \rho(z \mid \vec{x}_j, \vec{\theta}^*) \vec{\nabla}_{\vec{\nu}_z} \ln\bigl(\rho_z(\vec{x}_j \mid \vec{\nu}^*_z)\bigr) = \vec{0}. \tag{14}$$

Дальнейшее уточнение уравнений (14) максимально-правдоподобного восстановления изображений по выборке $X_k = (\vec{x}_1, \ldots, \vec{x}_k)$ зависит от выбора вида зависимости плотностей (условного) распределения вероятностей отсчетов $\rho_z(\vec{x}_j \mid \vec{\nu}_z)$ от параметров $\vec{\nu}_z$ для каждой из z-компонент (от конкретизации параметрической модели смеси (10)). Основная идея, лежащая в основе предлагаемой ниже модели, связана с теорией рецептивных полей.

5. Генеративная модель автокодировщика с компонентами смеси в виде рецептивных полей

Теория рецептивных полей позволяет использовать в задачах обработки изображений современные представления о механизмах восприятия челове-

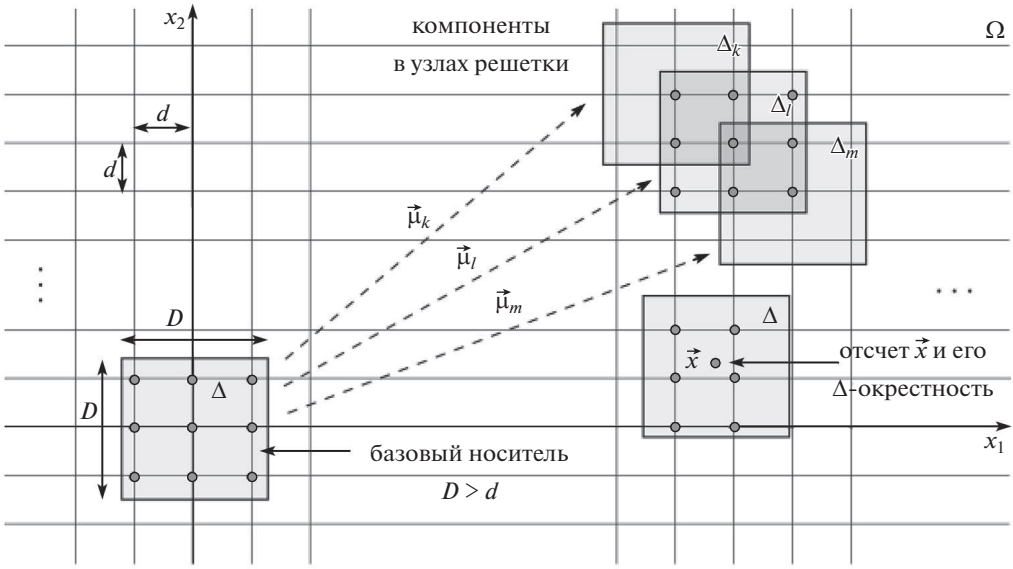

Рис. 3. Простейшая геометрия расположения носителей компонент $\{\Delta_z\}$ в узлах $\{\vec{\mu}_z\}$ прямоугольной сетки, покрывающей Ω. Показаны базовый носитель Δ и Δ-окрестность произвольной точки $\vec{x} \in \Omega$ с принадлежащими им узлами сетки.

ком зрительных образов. Основная концепция теории заключаются в том, что извлекаемая из изображений информация содержится не в значениях интенсивности, зарегистрированной отдельными рецепторами, а формируется на основе анализа кооперативной активности рецеторов, составляющих определенные локальные области [40, 41]. Основными доводами, согласно которым необходимо анализировать интенсивность по областям изображения, а не по отдельным точечным детекторам, являются, с одной стороны, необходимость накопления стимула по группам рецепторов для формирования значимого суммарного сигнала для выходных нейронов сетчатки, а с другой стороны, необходимость сжатия полного потока данных от рецепторов ввиду ограниченной пропускной способности зрительного канала. Подобные локальные области, составляющие зоны ответственности отдельных выходных зрительных нейронов, традиционно называют рецептивными полями. Функциональное описание рецептивных полей и способы их алгоритмического описания можно найти в [42].

В соответствии с концепцией рецептивных полей примем, что компоненты — плотности совместных распределений $\rho_z(\vec{x} \mid \vec{\nu}_z)$ (10) имеют финитные, не зависящие от параметров $\vec{\nu}_z$ носители $\Delta_z = \{\vec{x} \mid \rho_z(\vec{x} \mid \vec{\nu}_z) > 0\}$, расположенные в узлах $\vec{\mu}_z$ некоторой воображаемой сетки, покрывающей область изображения Ω (см. рис. 3). Предполагается, что совокупность всех носителей $\{\Delta_z\}$ составляет покрытие области изображения: $\Omega \subset \bigcup_{z=1}^{K} \Delta_z$. Таким

образом, множество носителей компонент $\{\Delta_z\}$ играет роль совокупности рецептивных полей.

Предположим далее для простоты, что набор компонент является однородным в области Ω — с точностью до расположений узлов $\{\vec{\mu}_z\}$ все условные распределения $\{\rho_z(\vec{x} \mid \vec{\nu}_z)\}$ принадлежат одному общему параметрическому семейству плотностей $g = \{\eta(\vec{\xi} \mid \vec{\nu})\}$, $\nu \in \Xi \subset \mathbb{R}^q$, так что $\rho_z(\vec{x} \mid \vec{\nu}_z) = \eta(\vec{x} - \vec{\mu}_z \mid \vec{\nu}_z)$. В этом случае параметрическое семейство смесей (10) имеет трансляционно-инвариантный по отношению к сетке вид

$$\rho(\vec{x} \mid \vec{\theta}) = \sum_{z=1}^{K} w_z \eta(\vec{x} - \vec{\mu}_z \mid \vec{\nu}_z). \tag{15}$$

Предполагается также, что все элементы семейства $g = \{\eta(\vec{\xi} \mid \vec{\nu})\}$ имеют общий финитный носитель Δ, который будет в дальнейшем называться базовым носителем. Поскольку компоненты (15) являются перемещенными в узлы сетки $\{\vec{\mu}_z\}$ элементами семейства, их носители $\{\Delta_z\}$ также будут перемещенными в узлы сетки копиями базового носителя Δ. Сам же базовый носитель, расположенный в окрестности начала координат $\mathbb{R}^2$, будет предполагаться симметричным в том смысле, что вместе с каждой своей точкой $\vec{\xi} \in \Delta$ он содержит также и $-\vec{\xi}$, в частности, содержит начало координат $\vec{0}$. Характерный размер базового носителя обозначим как D. Примером симметричного базового носителя является прямоугольный носитель Δ размеров $D \times D$ на рис. 3.

Для перекрытия (носителей) соседних компонент необходимым условием является $D > d$, где d — шаг сетки. Это же условие необходимо для принятого предположения о покрытии области Ω объединением носителей всех компонент $\{\Delta_z\}$. Отметим здесь, что для прямоугольной сетки в случае прямоугольной формы базового носителя условие $D > d$ также и достаточно, для круглой формы носителя достаточно $D > \sqrt{2}d$.

При покрытии Ω совокупностью носителей $\{\Delta_z\}$ каждая точка $\vec{x} \in \Omega$ принадлежит хотя бы одному из носителей. Поэтому множество узлов сетки, носители которых содержат $\vec{x}$, не пусто. Обозначим множество индексов этих узлов посредством $\delta_{\vec{x}} = \{z \mid \vec{x} \in \Delta_z\}$ и назовем его сеточным окружением $\vec{x}$. Заметим, что в силу симметричности базового носителя Δ, в $\delta_{\vec{x}}$ попадут те и только те узлы сетки, которые содержатся в области, полученной смещением в точку $\vec{x}$ базового носителя Δ (см. рис. 3). Поскольку в смеси (15) компоненты, носители которых не содержат $\vec{x}$, обращаются в ноль, представление смеси с использованием сеточного окружения $\delta_{\vec{x}}$ можно записать в сжатом виде

$$\rho(\vec{x} \mid \vec{\theta}) = \sum_{z \in \delta_{\vec{x}}} w_z \eta(\vec{x} - \vec{\mu}_z \mid \vec{\nu}_z), \tag{16}$$

где в отличие от (15) область суммирования по скрытым переменным z зависит от $\vec{x}$. При размерах сеточных окружений $|\delta_{\vec{x}}| \ll K$ в смеси (16) для каждой

точки $\vec{x}$ слагаемых будет намного меньше K (как формально предполагается в (15)).

Последним предположением относительно параметрического семейства смесей (16) является уточнение явного вида семейства $g = \{\eta(\vec{\xi} \mid \vec{\nu})\}$. Достаточно гибким и удобным с вычислительной точки зрения выбором представляется семейство экспоненциально-скошенных распределений (family of exponentially tilted densities) [43]. Семейство экспоненциально-скошенных распределений задается некоторой, не зависящей от параметров ν плотностью распределения вероятностей $\rho_0(\vec{\xi})$ и ее экспоненциально-скошенными, зависящими от параметров $\nu \in \Xi \subset \mathbb{R}^2$ версиями:

$$\eta(\vec{\xi} \mid \vec{\nu}) = \exp\{\vec{\xi}^T\vec{\nu} - A(\vec{\nu})\}\rho_0(\vec{\xi}), \tag{17}$$

где T — операция векторного транспонирования.

Отметим, что нормировка плотностей $\rho_0(\vec{\xi})$ и $\eta(\vec{\xi} \mid \vec{0})$ в (17) ведет к условию $A(\vec{0}) = 0$, или $\rho_0 = \eta(\vec{\xi} \mid \vec{0})$. Другими словами, определяющая семейство плотность $\rho_0(\vec{\xi})$ сама является элементом семейства. Более того, как следует из (17), носитель $\rho_0(\vec{\xi})$ задает базовый носитель Δ для всего семейства. В частности, ограниченность и симметрия Δ будут обеспечены, если $\rho_0(\vec{\xi})$, финитная и симметричная плотность, что в дальнейшем предполагается выполненным.

Функция параметров $A(\vec{\nu})$ в (17) обеспечивает нормировку плотностей $\eta(\vec{\xi} \mid \vec{\nu})$ при любых $\vec{\nu}$, что равносильно выполнению условия

$$A(\vec{\nu}) = \ln\left\{\int\limits_{\Delta} \exp(\vec{\xi}^T\vec{\nu})\rho_0(\vec{\xi})d\vec{\xi}\right\}. \tag{18}$$

Из (18) следует ввиду симметричности $\rho_0(\vec{\xi})$, что $A(\vec{\nu})$ также является симметричной функцией. Далее, ввиду финитности $\rho_0(\vec{\xi})$, функция параметров $A(\vec{\nu})$ определена и аналитична на всей плоскости $\mathbb{R}^2$ (и даже на $\mathbb{C}^2$). Из (18) также следует, что $A(\vec{\nu})$ является логарифмом двумерного преобразования Лапласа от $\rho_0(\vec{\xi})$, что характеризует ее как производящую функцию кумулянтов [44] для $\rho_0(\vec{\xi})$ (см. в этой связи Приложение (П.9)). Впрочем, через $A(\vec{\nu})$ достаточно просто выражается и производящая функция кумулянтов $\psi(\vec{\nu}')$ самого распределения $\eta(\vec{\xi} \mid \vec{\nu})$ (17): $\psi(\vec{\nu}') = A(\vec{\nu}' + \vec{\nu}) - A(\vec{\nu})$, что для первых моментов $\vec{\zeta}_{\vec{\nu}}$ и $R_{\vec{\nu}}$ дает:

$$\begin{gathered}\vec{\nabla}_{\vec{\nu}'}\psi(\vec{0}) = \vec{\nabla}A(\vec{\nu}) = \vec{\zeta}_{\vec{\nu}} = \int\limits_{\Delta} \vec{\xi}\eta(\vec{\xi} \mid \vec{\nu})d\vec{\xi}, \\ \vec{\nabla}_{\vec{\nu}'}\vec{\nabla}^T_{\vec{\nu}'}\psi(\vec{0}) = \vec{\nabla}\vec{\nabla}^T A(\vec{\nu}) = R_{\vec{\nu}} = \int\limits_{\Delta} (\vec{\xi} - \vec{\zeta}_{\vec{\nu}})(\vec{\xi} - \vec{\zeta}_{\vec{\nu}})^T\eta(\vec{\xi} \mid \vec{\nu})d\vec{\xi}.\end{gathered} \tag{19}$$

Как следует из (19), матрица Гесса $\vec{\nabla}\vec{\nabla}^T A(\vec{\nu})$ при любых $\vec{\nu}$ является корреляционной матрицей $R_{\vec{\nu}}$ распределения (17), поэтому она всюду положительно определена и, как следствие, всюду строго выпукла. Последнее обстоятельство обеспечивает взаимную однозначность отображения $\vec{\nu} \to \vec{\zeta}_{\vec{\nu}} = \vec{\nabla} A(\vec{\nu})$. Ввиду этой взаимной однозначности можно вместо параметров $\vec{\nu}$, называемых в контексте экспоненциальных семейств натуральными (или каноническими) [44], пользоваться параметрами средних $\vec{\zeta}$. Заметим, что в соответствии $\vec{\nu} \leftrightarrow \vec{\zeta}$ в силу симметрии $A(\vec{\nu})$ ноль отображается в ноль.

Переход к параметрам средних в (17) тесно связан с преобразованием Лежандра [43] функции $A(\vec{\nu})$ (18):

$$A^*(\vec{\zeta}) = \max_{\vec{\nu} \in \Xi} \left(\vec{\zeta}^T \vec{\nu} - A(\vec{\nu}) \right). \tag{20}$$

Максимум в (20) достигается на решении уравнения $\vec{\nabla} A(\vec{\nu}_{\vec{\zeta}}) = \vec{\zeta}$, а его величина $A^*(\vec{\zeta})$ равна значению выражения $\vec{\zeta}^T \vec{\nu}_{\vec{\zeta}} - A(\vec{\nu}_{\vec{\zeta}})$. Отметим, что для симметричной $A(\vec{\nu})$ ее дуальная по Лежандру функция $A^*(\vec{\zeta})$ также симметрична.

Если переобозначить в (20) $\vec{\zeta} \to \vec{\xi}$, можно с учетом явного вида дуальной по Лежандру функции $A^*(\vec{\zeta})$ преобразовать показатель экспоненты в (17) следующим образом:

$$\begin{aligned} \vec{\xi}^T \vec{\nu} - A(\vec{\nu}) = \vec{\xi}^T (\vec{\nu} - \vec{\nu}_{\vec{\xi}}) - A(\vec{\nu}) + A(\vec{\nu}_{\vec{\xi}}) + A^*(\vec{\xi}) = \\ = -B_A(\vec{\nu}, \vec{\nu}_{\vec{\xi}}) + A^*(\vec{\xi}), \end{aligned} \tag{21}$$

где введена связанная с $A(\vec{\nu})$ дивергенция Брегмана (Bregman divergence) [45]:

$$B_A(\vec{\nu}, \vec{\nu}\,') = A(\vec{\nu}) - A(\vec{\nu}\,') - \vec{\nabla}^T A(\vec{\nu}\,')(\vec{\nu} - \vec{\nu}\,'). \tag{22}$$

Аналогично дивергенции $B_A(\ldots)$ (22) можно ввести дивергенцию Брегмана $B_{A^*}(\ldots)$, связанную с дуальной функцией $A^*(\vec{\zeta})$ (20). Оказывается, что обе эти дивергенции также связаны соотношением дуальности $B_A(\vec{\nu}, \vec{\nu}\,') = B_{A^*}(\vec{\zeta}\,', \vec{\zeta})$, где предполагается $\vec{\zeta} = \vec{\nabla} A(\vec{\nu})$, $\vec{\zeta}\,' = \vec{\nabla} A(\vec{\nu}\,')$ [45]. Поэтому, на основе (21), (22) семейство скошенных распределений (17) можно переписать в параметрах средних в виде

$$\eta(\vec{\xi} \mid \vec{\zeta}) = \eta(\vec{\xi} \mid \vec{\nu}_{\vec{\zeta}}) = \exp\{-B_{A^*}(\vec{\xi}, \vec{\zeta})\} \exp\{A^*(\vec{\xi})\} \rho_0(\vec{\xi}). \tag{23}$$

Отметим, что хотя форма (23) распределений семейства не столь явно как (17) выражает деформацию симметричной плотности $\rho_0(\vec{\xi})$, она также содержит некоторую трактовку скошенных распределений, ввиду следующей из (19) интерпретации $\vec{\zeta}$ как среднего от $\vec{\xi}$. Именно форма (23) представляет плотности семейства как произведение факторов $\exp\{-B_{A^*}(\vec{\xi}, \vec{\zeta})\}$ — купола-подобной функции, достигающей максимума в $\vec{\xi} = \vec{\zeta}$ и $\exp\{A^*(\vec{\xi})\} \rho_0(\vec{\xi})$ —

симметричной функции, имеющей максимум в нуле. Эту трактовку можно подчеркнуть еще больше, если воспользоваться приближенным выражением (П.15) из Приложения для фактора $\exp\{A^*(\vec{\xi})\}\rho_0(\vec{\xi}) \simeq 1/2\pi D^2$, где D — стандартное отклонение для плотности $\rho_0(\vec{\xi})$, которое также можно интерпретировать как характерный размер базового носителя Δ:

$$\eta(\vec{\xi} \mid \vec{\zeta}) = \frac{1}{2\pi D^2}\exp\{-B_{A^*}(\vec{\xi}, \vec{\zeta})\}. \tag{24}$$

Приближенная форма (24) в еще большей степени подчеркивает интерпретацию $\eta(\vec{\xi} \mid \vec{\zeta})$ в параметрах средних как смещенное в точку $\vec{\zeta}$ симметричное распределение $\rho_0(\vec{\xi})$. Отметим, что для (не финитных) симметричных гауссовых плотностей, для которых $R_{\vec{\nu}} = R_0 = D^2E$, форма (24) оказывается точной.

Возвращаясь к системе уравнений максимального правдоподобия (14), выразим входящие в эти уравнения градиенты логарифмических функций правдоподобия $\ln\bigl(\rho_z(\vec{x} \mid \vec{\nu}_z)\bigr) = \ln\bigl(\eta(\vec{x} - \vec{\mu}_z \mid \vec{\nu}_z\bigr)$ через средние параметры $\vec{\zeta}_z$ согласно (23) и воспользуемся тем обстоятельством, что $\vec{\zeta}_z = \vec{\nabla}A(\vec{\nu}_z)$ и, следовательно, $\|\vec{\nabla}_{\vec{\nu}_z}\vec{\zeta}_z^{\,T}\| = \vec{\nabla}\vec{\nabla}^T A(\vec{\nu}_z)$

$$\begin{aligned}\vec{\nabla}_{\vec{\nu}_z}\ln\bigl(\eta(\vec{x} - \vec{\mu}_z \mid \vec{\nu}_z\bigr) &= \|\vec{\nabla}_{\vec{\nu}_z}\vec{\zeta}_z^{\,T}\|\vec{\nabla}_{\vec{\zeta}_z}\{-B_{A^*}(\vec{x} - \vec{\mu}_z, \vec{\zeta}_z)\} = \\ &= -\vec{\nabla}\vec{\nabla}^T A(\vec{\nu}_z)\vec{\nabla}_{\vec{\zeta}_z}B_{A^*}(\vec{x} - \vec{\mu}_z, \vec{\zeta}_z) = \vec{x} - \vec{\mu}_z - \vec{\zeta}_z.\end{aligned} \tag{25}$$

где учтено, что согласно (22) $\vec{\nabla}_{\vec{\zeta}}B_{A^*}(\vec{\xi}, \vec{\zeta}) = -\vec{\nabla}\vec{\nabla}^T A^*(\vec{\zeta})(\vec{\xi} - \vec{\zeta})$ и матрицы $\vec{\nabla}\vec{\nabla}^T A^*(\vec{\zeta})$ и $\vec{\nabla}\vec{\nabla}^T A(\vec{\nu})$ взаимнообратны. Отметим, что получившееся в результате простое выражение в правой части (25) является главным аргументом перехода к параметрам средних. Однако здесь следует сделать важное замечание. На самом деле соотношения (25) выполняются не для всех $\vec{x}$, а только для тех, которые принадлежат носителю Δ_z данной компоненты. Для прочих $\vec{x}$ целесообразно положить градиенты (25) равными нулю.

Подставляя (25) в (14), окончательно получим систему уравнений максимального правдоподобия для нахождения оптимальных параметров $\vec{\theta}^* = \{\{w_z^*\}, \{\vec{\nu}_z^*\}\} \in \Theta$

$$\begin{aligned} &z = 1, \ldots, K, \\ &w_z^* = \frac{1}{k}\sum_{\vec{x}_j \in \Delta_z}\rho(z \mid \vec{x}_j, \vec{\theta}^*), \\ &\vec{\zeta}_z^* = \frac{\sum\limits_{\vec{x}_j \in \Delta_z}(\vec{x}_j - \vec{\mu}_z)\rho(z \mid \vec{x}_j, \vec{\theta}^*)}{\sum\limits_{\vec{x}_j \in \Delta_z}\rho(z \mid \vec{x}_j, \vec{\theta}^*)}, \end{aligned} \tag{26}$$

где при суммировании по отсчетам $\vec{x}_j$ учтено, что плотность $\rho(z \mid \vec{x}_j, \vec{\theta}^*) \sim$ $\sim \rho(\vec{x}_j, z \mid \vec{\theta}^*)$ отлична от нуля только в Δ_z.

Система уравнений (26) является относительно простой нелинейной системой, выражающей искомое решение $\vec{\theta}^*$ через функцию от него же и от отсчетов выборки (выборочного представления) $X_k = \{\vec{x}_j\}$:

$$\vec{\theta}^* = H(\vec{\theta}^*, X_k). \tag{27}$$

Простейшим методом решения уравнений вида (27) является метод последовательных приближений (successive approximations) [46], который по ν-приближению $\vec{\theta}^{(\nu)}$ итеративно находит следующее приближение $\vec{\theta}^{(\nu+1)} =$ $= H(\vec{\theta}^{(\nu)}, X_k)$. Существует много методов численного решения подобных уравнений [47]. Особенности каждого из этих методов определяются особенностями итерирующей функции $H(\ldots)$.

В случае системы (26) особенностью функции $H(\vec{\theta}, X_k)$ является то, что она зависит от $\vec{\theta}$ только через плотности апостериорных распределений $\rho(z \mid \vec{x}, \vec{\theta})$. Поэтому естественно разбить вычисления каждой итерации на два шага: на первом вычислить все апостериорные распределения:

$$\mathbf{I} \qquad j = 1, \ldots, k, \ z \in \delta_{\vec{x}_j}$$

$$\begin{gathered}
\rho_j^{(\nu+1)}(z) = \rho\left(z \mid \vec{x}_j, \vec{\theta}^{\,(\nu)}\right) = \frac{w_z^\nu \rho_z\left(\vec{x}_j \mid \vec{\theta}^{\,(\nu)}\right)}{P_j^{(\nu)}} = \frac{w_z^\nu \eta\left(\vec{x}_j - \vec{\mu}_z \mid \ \zeta_z^{(\nu)}\right)}{P_j^{(\nu)}} = \\
= \frac{w_z^\nu}{P_j^{(\nu)}} \exp\left\{-B_{A^*}\left(\vec{x}_j - \vec{\mu}_z, \vec{\zeta}_z^{\,(\nu)}\right)\right\} \exp\left\{A^*(\vec{x}_j - \vec{\mu}_z)\right\} \rho_0\left(\vec{x}_j - \vec{\mu}_z\right),
\end{gathered} \tag{28}$$

$$\begin{gathered}
P_j^{(\nu)} = \rho\left(\vec{x}_j \mid \vec{\theta}^{\,(\nu)}\right) = \sum_{i \in \delta_{\vec{x}_j}} w_i^\nu \eta\left(\vec{x}_j - \vec{\mu}_i \mid \ \zeta_i^{\,(\nu)}\right) = \\
= \sum_{i \in \delta_{\vec{x}_j}} w_i^\nu \exp\left\{-B_{A^*}\left(\vec{x}_j - \vec{\mu}_i, \vec{\zeta}_i^{\,(\nu)}\right)\right\} \exp\left\{A^*(\vec{x}_j - \vec{\mu}_i)\right\} \rho_0\left(\vec{x}_j - \vec{\mu}_i\right),
\end{gathered}$$

а уже на втором шаге пересчитать текущее приближение

$$\mathbf{II} \qquad z = 1, \ldots, K$$

$$\begin{gathered}
w_z^{(\nu+1)} = \frac{1}{k} \sum_{\vec{x}_j \in \Delta_z} \rho_j^{(\nu+1)}(z), \\
\vec{\zeta}_z^{\,(\nu+1)} = \frac{\sum\limits_{\vec{x}_j \in \Delta_z} (\vec{x}_j - \vec{\mu}_z) \rho_j^{(\nu+1)}(z)}{\sum\limits_{\vec{x}_j \in \Delta_z} \rho_j^{(\nu+1)}(z)}.
\end{gathered} \tag{29}$$

Двухшаговая итеративная вычислительная схема **I** (28) и **II** (29) соответствует шагам E и M известного EM-алгоритма для смесей распределений экспоненциального семейства [48]. Известно, что EM-алгоритм, когда число компонент K относительно невелико ($\sim$ 10–100) вполне стабилен и позволяет в обозримое время находить максимально-правдоподобные оценки $\vec{\theta}^*$. К сожалению, при больших объемах данных k и больших размерах моделей K, например, для стандартных цифровых изображений, применение традиционного EM-алгоритма оказывается проблематичным. Проблемы эти связаны с высокими требованиями к объему памяти $k \times K$ и соответственно с большим объемом вычислений, а также с низкой скоростью сходимости (линейной) EM-алгоритма [21]. В предложенной нами модели рецептивных полей, ввиду уменьшения требуемого объема памяти до $k \times |\overline{\delta}|$, где $|\overline{\delta}| \ll K$ — среднее число узлов в решеточных окружениях отсчетов, и сокращения объема вычислений ввиду ограниченного суммирования по компонентам в (28) и отсчетам в (29), требования к ресурсам гораздо меньше, чем для EM-алгоритма в общей версии, особенно при большом числе K компонент.

Вместе с тем, несмотря на очевидную экономию ресурсов, при очень больших $K \sim 10^4$–10^6, представляющих интерес для реальных изображений, объем вычислений в предложенной схеме может оказаться по-прежнему высоким. В этом случае можно несколько снизить объем памяти и вычислений, если воспользоваться аппроксимацией (24). Используя ее, вычисления на шаге **I** (28) можно организовать следующим образом:

$$\mathbf{I} \qquad j = 1, \ldots, k, \;\; z \in \delta_{\vec{x}_j}$$

$$\rho_j^{(\nu+1)}(z) = \frac{w_z^\nu \exp\left\{-B_{A^*}(\vec{x}_j - \vec{\mu}_z, \vec{\zeta}_z^{(\nu)})\right\}}{\sum\limits_{i \in \delta_{\vec{x}_j}} w_i^\nu \exp\left\{-B_{A^*}\left(\vec{x}_j - \vec{\mu}_i, \vec{\zeta}_i^{(\nu)}\right)\right\}} = \tag{30}$$

$$= \mathsf{softmax}_z\left(\left\{\ln(w_i^\nu) - B_{A^*}\left(\vec{x}_j - \vec{\mu}_i, \vec{\zeta}_i^{(\nu)}\right) \mid i \in \delta_{\vec{x}_j}\right\}\right),$$

где $\mathsf{softmax}_z$ — z-компонента softmax-функции [49]. Вместе с шагом **II** (29) шаг **I** (30) почти совпадает со схемой мягкой брегмановской кластеризации (Bregman Soft Clustering), обсуждавшейся в [23]. Отличие состоит лишь в том, что в **I** (30) для расчета апостериорных плотностей $\rho_j^{(\nu+1)}(z)$ используются брегмановские дивергенции между локальными по отношению к компоненте z координатами отсчетов $\vec{x}_j - \vec{\mu}_z$ и локальными же координатами центроида $\vec{\zeta}_z^{(\nu)}$ тех отсчетов $\vec{x}_j$, которые принадлежат носителю Δ_z этой компоненты

$$\vec{\zeta}_z^{(\nu)} = \frac{\sum\limits_{\vec{x}_j \in \Delta_z} \vec{x}_j \rho_j^{(\nu)}(z)}{\sum\limits_{\vec{x}_j \in \Delta_z} \rho_j^{(\nu)}(z)} - \vec{\mu}_z. \tag{31}$$

В схеме же мягкой брегмановской кластеризации [23] используются брегмановские дивергенции $B_{A^*}(\vec{x}, \vec{M}_z)$ между собственно координатами отсчетов $\vec{x}$ и их несмещенными центроидами $\vec{M}_z = \vec{\zeta}_z^{(\nu)} + \vec{\mu}_z$. Поскольку в общем случае дивергенция Брегмана не является трансляционно-инвариантной функцией: $B_{A^*}(\vec{x}_j - \vec{\mu}_z, \vec{\zeta}_z^{(\nu)}) \neq B_{A^*}(\vec{x}_j, \vec{\zeta}_z^{(\nu)} + \vec{\mu}_z)$, обсуждаемые схемы вычислений не тождественны. Они совпадают только в одном частном случае квадратичных по разности аргументов дивергенций Брегмана, соответствующих случаю гауссовских компонент. Таким образом, предложенная схема вычисления максимально-правдоподобных оценок параметров $\vec{\theta}^*$ отличается свойством локальности обучения в отличие от известных процедур мягкой брегмановской кластеризации [23].

Оказывается, можно еще более сэкономить вычислительные ресурсы, если воспользоваться "жестким" приближением значений softmax функции в (30), аппроксимировав нулями все ее z-компоненты кроме максимальной, для которой выбирается значение 1. При этом шаг **I** (30) сведется к задаче максимизации:

$$\mathbf{I} \qquad j = 1, \ldots, k,\ z \in \delta_{\vec{x}_j}$$

$$i^* = arg \max_{i \in \delta_{\vec{x}_j}} \left\{ \ln(w_i^\nu) - B_{A^*}\left(\vec{x}_j - \vec{\mu}_i, \vec{\zeta}_i^{(\nu)}\right) \right\}, \tag{32}$$

$$\rho_j^{(\nu+1)}(z) = \delta_{i^* z}.$$

Жесткая аппроксимация $\rho_j^{(\nu+1)}(z)$ (32) приводит к однозначной ассоциации отсчетов $\vec{x}_j$ с узлами сетки $\vec{\mu}_{i^*}$, или, что эквивалентно, разбиению отсчетов выборки $X_k = (\vec{x}_1, \ldots, \vec{x}_k)$ на K непересекающихся подгрупп $\left\{X_{k_z}^{(\nu+1)} = \right.$ $\left. = \left(\vec{x}_{j_1}, \ldots, \vec{x}_{j_{k_z}}\right)\right\}$, где k_z — число отсчетов $\vec{x}_j$ объединенных в группу $X_{k_z}^{(\nu+1)}$ согласно ассоциации $\rho_j^{(\nu+1)}(z) = 1$ (некоторые из $X_{k_z}^{(\nu+1)}$ могут быть пустыми). В терминах результирующего разбиения вычисления на шаге **II** (29) также упрощаются и принимают вид

$$\mathbf{II} \qquad z = 1, \ldots, K$$

$$w_z^{(\nu+1)} = \frac{k_z}{k}, \tag{33}$$

$$\vec{\zeta}_z^{(\nu+1)} = \frac{1}{k_z} \sum_{\vec{x}_j \in X_{k_z}^{(\nu+1)}} (\vec{x}_j - \vec{\mu}_z).$$

Интересно в этой связи отметить, что если исключить веса $\{w_z\}$ из числа параметров $\vec{\theta}$, предписав им равномерное распределение $w_z = 1/K$, то схема

жестких вычислений **I** (32)–**II** (33) принимает форму:

$$
\begin{gathered}
\mathbf{I} \qquad j = 1, \ldots, k, \quad z \in \delta_{\vec{x}_j} : \\
z = arg \min_{i \in \delta_{\vec{x}_j}} \left\{ B_{A^*} \left(\vec{x}_j - \vec{\mu}_i, \vec{\zeta}_i^{(\nu)} \right) \right\}, \\
k_z = k_z + 1, \; X_{k_z}^{(\nu+1)} = X_{k_z}^{(\nu+1)} \bigcup \vec{x}_j, \\
\mathbf{II} \qquad z = 1, \ldots, K : \\
\vec{\zeta}_z^{(\nu+1)} = \frac{1}{k_z} \sum_{\vec{x}_j \in X_{k_z}^{(\nu+1)}} \vec{x}_j - \vec{\mu}_z, \\
k_z = 0, \; X_{k_z}^{(\nu+2)} = \varnothing.
\end{gathered}
\tag{34}
$$

Схема (34) практически совпадает (с точностью до нюансов, обсужденных выше) с алгоритмом жесткой Брегмановской кластеризации (Bregman Hard Clustering) [23]. Интересно отметить, что там же приводится замечание о том, что при конкретных выборах дивергенции Брегмана $B_{A^*}(\vec{\xi}, \vec{\zeta})$ получаются популярные методы кластеризации. Именно, классический алгоритм K-средних (K-means), алгоритм LBG [24] и алгоритм теоретико-информационной кластеризации [50] являются частными случаями жесткой кластеризации, когда $B_{A^*}(\vec{\xi}, \vec{\zeta})$ имеет вид квадрата евклидова расстояния, расстояния Итакуры–Сайто или дивергенции Кульбака–Лейблера [39].

С целью продемонстрировать возможности кодирования и восстановления предложенными в работе генеративными автокодировщиками заданных выборочными представлениями изображений, был использован пример простых рецептивных полей гауссовского типа [42] (см. (24)):

$$
\begin{gathered}
\eta(\vec{\xi} \mid \vec{\zeta}) = \frac{1}{2\pi D^2} \exp\left\{ -(\vec{\xi} - \vec{\zeta})^2 / 2D^2 \right\} = \exp\left\{ \vec{\xi}^{\,T} \vec{\nu} - A(\vec{\nu}) \right\} \rho_0(\vec{\xi}), \\
\rho_0(\vec{\xi}) = \frac{1}{2\pi D^2} \exp\left\{ -\vec{\xi}^{\,2} / 2D^2 \right\}, \quad A(\vec{\nu}) = D^2 \vec{\nu}^{\,2} / 2, \quad \vec{\zeta} = D^2 \vec{\nu}.
\end{gathered}
\tag{35}
$$

На рис. 4 приведены результаты кодирования-восстановления изображения “Cameraman”, содержащего 1 000 000 отсчетов (рис. 2). Кодирование осуществлялось по схеме жесткой аппроксимации вычислений **I** (32)—**II** (33) с разбиением выборочного представления $X_k = (\vec{x}_1, \ldots, \vec{x}_k)$ на $K = l \times l$ непересекающихся сеточных кластеров $\left\{ X_{k_z}^{(\nu)} \subset \Delta_z \right\}$, соответствующих узлам прямоугольной сетки, размерами l узлов по вертикали и l узлов по горизонтали изображения Ω (рис. 3). Число итераций схемы полагалось равным $\nu_{\max} = 10$. По вычисленным в результате параметрам $\vec{\theta}^* = \left\{ \{w_z^*\}, \left\{ \vec{\zeta}_z^* \right\} \right\}$ восстановление плотности распределения вероятностей отсчета $\vec{x} \in \Omega$ (совпадающей с нормированной интенсивностью $I(\vec{x})$ излучения на Ω в данном случае

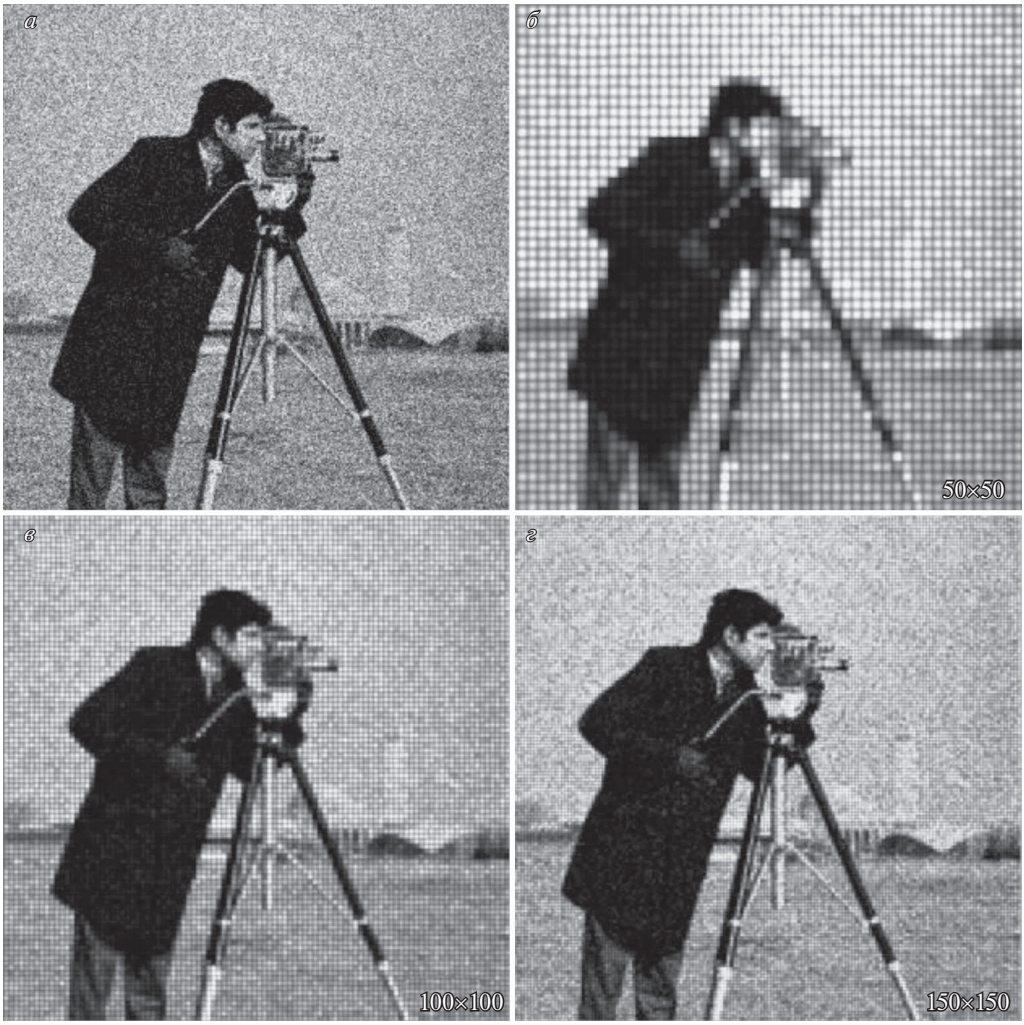

Рис. 4. Результаты кодирования–восстановления выборочного представления изображения “Cameraman”: *а* — выборочное представление размером 1 000 000 отсчетов как на рис. 2; *б*, *в*, *г* — восстановленные изображения, соответствующие сеткам в 50×50, 100×100 и 150×150 узлов.

с нормированной битмап-картой исходного изображения “Cameraman”, см. рис. 2), осуществлялось в соответствии с (16), (24)

$$\rho(\vec{x},\vec{\theta}^{*}) = \sum_{i\in\delta_{\vec{x}}} \frac{w_i^{*}}{2\pi D^2} \exp\left\{-\frac{(\vec{x}-\vec{\mu}_i-\vec{\zeta}_z^{*})^2}{2D^2}\right\}. \tag{36}$$

6. Заключение

Предложенная в работе концепция автокодировщиков, предназначенных для автоматической генерации сжатых изображений, оказалась содержатель-

ной в отношении открывающихся с ее помощью новых возможностей по синтезу реальных алгоритмов кодирования-восстановления изображений. Разработанное для этих целей специальное представление изображений с помощью выборок отсчетов (выборочных представлений) позволяет, с одной стороны, избежать связанных с идеальными изображениями проблем вычислительных ресурсов, а с другой стороны, открывает широкие возможности по адаптации методов машинного обучения к задачам, подобным рассмотренным в работе.

Основываясь на специфике выборочных представлений, удалось сформулировать генеративную (порождающую) модель автокодировщиков, которая почти автоматически выводит на такие разделы машинного обучения, как наивный байесов подход, методы максимального правдоподобия, итеративные алгоритмы типа EM-алгоритма поиска оценок максимально-правдоподобных оценок для смесей, алгоритмы кластеризации типа K-means, LBG алгоритмы векторного квантования и т.д. В этой связи в работе представлено несколько различающихся по сложности примеров итеративных алгоритмических схем кодирования–восстановления изображений.

Особенностью предложенных схем является активно используемая в них концепция рецептивных полей. Она позволяет эффективно обходить известные трудности итеративных алгоритмов, обрабатывающих смеси с большим числом компонент, например порядка 10^4–10^6, что по порядку соответствует числу рецептивных полей в сетчатке глаза. Апелляция к зрительному восприятию здесь не случайна, поскольку, как подчеркнуто в работе, качеству восприятия изображений и адаптации механизмов зрительного восприятия человека к задачам цифровой обработке изображений сегодня уделяется повышенное внимание.

В связи с изложенным выше отметим, что практически все приведенные в работе алгоритмы были численно апробированы, проведенные эксперименты подтвердили их эффективность по отношению к памяти и времени вычисления. К примеру, выбранные для иллюстрации отсчетные изображения “Cameraman” (см. рис. 2) потребовали на вычисление восстановленных изображений времени менее секунды даже в случае наиболее плотной сетки в 150×150 узлов (22 500 компонент).

В целом, основываясь на полученных результатах, хотелось бы выразить надежду, что предложенные в работе подходы найдут в ближайшее время как дальнейшее теоретическое развитие, так и плодотворное использование в прикладных задачах.

ПРИЛОЖЕНИЕ

Асимптотический вид апостериорного распределения параметров для выборки k независимых, одинаково распределенных (iid) отсчетов

Пусть дана выборка из k независимых одинаково распределенных (iid) случайных отсчетов $X_k = (\vec{x}_1, \ldots, \vec{x}_k)$, плотность распределения вероятностей

каждого из которых определяется некоторым элементом параметрического семейства $G = \left\{\rho(\vec{x} \mid \vec{\theta})\right\}$, $\vec{x} \in \Omega$, $\vec{\theta} \in \Theta \subset \mathbb{R}^p$. Пусть существует точка $\vec{\theta}_{ML} \in \in \Theta$, в которой имеет место максимум по $\vec{\theta}$ логарифма функции правдоподобия совместной плотности распределения отсчетов X_k

$$\nabla_{\vec{\theta}} L(\vec{\theta}_{ML}; X_k) = \vec{0},$$

$$L(\vec{\theta}; X_k) = \ln\bigl(\rho(X_k \mid \vec{\theta})\bigr) = \ln\left(\prod_{j=1}^{k} \rho(\vec{x}_j \mid \vec{\theta})\right). \tag{П.1}$$

Запишем в окрестности $\vec{\theta}_{ML}$ тейлоровское разложение логарифма функции правдоподобия (П.1):

$$L(\vec{\theta}; X_k) = L(\vec{\theta}_{ML}; X_k) + \frac{1}{2}(\vec{\theta} - \vec{\theta}_{ML})^T \left(\frac{\partial^2 L(\vec{\theta}_{ML}; X_k)}{\partial\vec{\theta}_p \partial\vec{\theta}_q}\right)(\vec{\theta} - \vec{\theta}_{ML}) + \ldots. \tag{П.2}$$

На основании (П.1) для матрицы вторых частных производных в (П.2) можно записать следующее представление

$$\left(\frac{\partial^2 L(\vec{\theta}_{ML}; X_k)}{\partial\vec{\theta}_p \partial\vec{\theta}_q}\right) = k\left(\frac{1}{k}\sum_{j=1}^{k} \frac{\partial^2 \ln \rho(\vec{x}_j \mid \vec{\theta}_{ML})}{\partial\vec{\theta}_p \partial\vec{\theta}_q}\right). \tag{П.3}$$

Рассматривая вторые производные $\partial^2 \ln \rho(\vec{x}_j \mid \vec{\theta}_{ML})/\partial\vec{\theta}_p \partial\vec{\theta}_q$ в (П.3) как случайные величины — функции от случайных $\vec{x}_j$, можно асимптотически при $k \gg 1$ заменить выборочное среднее в правой части (П.3) соответствующим средним по распределению $\rho(\vec{x} \mid \vec{\theta})$

$$\frac{1}{k}\sum_{j=1}^{k} \frac{\partial^2 \ln \rho(\vec{x}_j \mid \vec{\theta}_{ML})}{\partial\vec{\theta}_p \partial\vec{\theta}_q} \simeq \int \rho(\vec{x} \mid \vec{\theta}) \frac{\partial^2 \ln \rho(\vec{x} \mid \vec{\theta}_{ML})}{\partial\vec{\theta}_p \partial\vec{\theta}_q} d\vec{x}. \tag{П.4}$$

Если учесть асимптотическую несмещенность $\vec{\theta}_{ML} \sim \vec{\theta}$, то нетрудно заметить, что правая часть (П.4) представляет собой с точностью до знака информационную матрицу Фишера $\Im(\vec{\theta})$ распределения $\rho(\vec{x} \mid \vec{\theta})$. Подставляя (П.4) в (П.3), а затем в (П.2), получаем следующую форму разложения логарифмической функции правдоподобия

$$L(\vec{\theta}; X_k) \simeq L(\vec{\theta}_{ML}; X_k) - \frac{k}{2}(\vec{\theta} - \vec{\theta}_{ML})^T \Im(\vec{\theta})(\vec{\theta} - \vec{\theta}_{ML}). \tag{П.5}$$

Экспонирование (П.5) дает асимптотику самой функции правдоподобия

$$\rho(X_k \mid \vec{\theta}) \simeq \rho(X_k \mid \vec{\theta}_{ML}) \exp\left\{-\frac{k}{2}(\vec{\theta} - \vec{\theta}_{ML})^T \Im(\vec{\theta})(\vec{\theta} - \vec{\theta}_{ML})\right\}. \tag{П.6}$$

От $\vec{\theta}$ в правой части (П.6) зависит только экспоненциальный сомножитель, поскольку $\vec{\theta}_{ML}$ в первом сомножителе является функцией от $X_k = (\vec{x}_1, \ldots, \vec{x}_k)$ и от $\vec{\theta}$ не зависит. Следовательно, поведение $\rho(X_k \mid \vec{\theta})$ в зависимости от $\vec{\theta}$ определяется только квадратичной формой в экспоненте с большим коэффициентом $k \gg 1$. Отсюда асимптотически $\rho(X_k \mid \vec{\theta})$ имеет острый пик в точке $\vec{\theta}_{ML}$, что может быть использовано для вычисления других распределений. Например, предполагая некоторое априорное распределение параметров $P(\vec{\theta})$, более широкое, чем $\rho(X_k \mid \vec{\theta})$, можно последовательно найти

$$\begin{aligned} \rho(X_k) &= \int_{\Theta} \rho(X_k \mid \vec{\theta}) P(\vec{\theta}) d\vec{\theta} \approx \rho(X_k \mid \vec{\theta}_{ML}) C(\vec{\theta}_{ML}) P(\vec{\theta}_{ML}), \\ C(\vec{\theta}_{ML}) &= \int_{\Theta} \exp\left\{ -\frac{k}{2} (\vec{\theta} - \vec{\theta}_{ML})^T \Im(\vec{\theta}) (\vec{\theta} - \vec{\theta}_{ML}) \right\} d\vec{\theta}, \end{aligned} \tag{П.7}$$

а затем

$$\begin{aligned} \rho(\vec{\theta} \mid X_k) &= \frac{\rho(X_k \mid \vec{\theta}) P(\vec{\theta})}{\rho(X_k)} \approx \\ \approx \frac{1}{C(\vec{\theta}_{ML})} \frac{P(\vec{\theta})}{P(\vec{\theta}_{ML})} &\exp\left\{ -\frac{k}{2} (\vec{\theta} - \vec{\theta}_{ML})^T \Im(\vec{\theta}) (\vec{\theta} - \vec{\theta}_{ML}) \right\} \approx \\ \approx \frac{1}{C(\vec{\theta}_{ML})} &\exp\left\{ -\frac{k}{2} (\vec{\theta} - \vec{\theta}_{ML})^T \Im(\vec{\theta}) (\vec{\theta} - \vec{\theta}_{ML}) \right\}, \end{aligned} \tag{П.8}$$

что представляет собой узкое, гауссова типа распределение, стремящееся к $\delta(\vec{\theta} - \vec{\theta}_{ML})$-функции Дирака при $k \to \infty$.

Приближение седловой точки для аппроксимации плотности распределения вероятностей случайной векторной величины

Приближение седловой точки является хорошо известным инструментом в статистике и достаточно широко используется при асимптотической аппроксимации поведения выборочного среднего большого числа k независимых, одинаково распределенных случайных величин [51]. Для анализа распределений отдельных случайных величин данное приближение используется существенно реже [43]. Ниже, с целью адаптации к рассматриваемым задачам метода седловой точки в случае одной (векторной) случайной величины, приводятся основные шаги обоснования этого приближения.

Пусть случайный вектор $\vec{\xi}$ задан на некоторой области $\Delta \subset \mathbb{R}^2$, являющейся носителем плотности распределения вероятностей $\rho_0(\vec{\xi})$. Для простоты будем полагать плотность симметричной, финитной функцией, так что носитель $\Delta = \left\{ \vec{\xi} \mid \rho_0(\vec{\xi}) > 0 \right\}$ является открытым, ограниченным, симметричным относительно начала координат множеством в $\mathbb{R}^2$. В этом случае для любых

(в том числе комплексных) векторов $\vec{\nu} \in \mathbb{C}^2$ существует двумерное преобразование Лапласа плотности $\rho_0(\vec{\xi})$ [52], которое может быть представлено в виде

$$F(\vec{\nu}) = \int\limits_{\Delta} \exp(-\vec{\xi}^{\,T}\vec{\nu})\rho_0(\vec{\xi})d\vec{\xi} = \int\limits_{\Delta} \exp(\vec{\xi}^{\,T}\vec{\nu})\rho_0(\vec{\xi})d\vec{\xi}, \tag{П.9}$$

где использован тот факт, что для симметричных $\rho_0(\vec{\xi})$ и Δ преобразование Лапласа $F(\vec{\nu})$ также симметрично.

Плотность $\rho_0(\vec{\xi})$ в свою очередь может быть выражена с помощью обратного преобразования Лапласа [52] от $F(\vec{\nu})$ (П.9)

$$\begin{aligned}\rho_0(\vec{\xi}) &= \frac{1}{(2\pi\hat{i})^2} \int\limits_{-\hat{i}\infty}^{\hat{i}\infty} \int\limits_{-\hat{i}\infty}^{\hat{i}\infty} \exp(\vec{\xi}^{\,T}\vec{\nu})F(\vec{\nu})d\vec{\nu} = \\ &= \frac{1}{(2\pi\hat{i})^2} \int\limits_{-\hat{i}\infty}^{\hat{i}\infty} \int\limits_{-\hat{i}\infty}^{\hat{i}\infty} \exp(A(\vec{\nu}) - \vec{\xi}^{\,T}\vec{\nu})d\vec{\nu},\end{aligned} \tag{П.10}$$

где $\hat{i}$ — мнимая единица, $A(\vec{\nu}) = \ln F(\vec{\nu})$ — производящая функция кумулянтов. В (П.10) также использован факт симметричности $\rho_0(\vec{\xi})$.

Интеграл в (П.10) можно приближенно найти с помощью метода седловой точки, если деформировать путь интегрирования так, чтобы он прошел через (седловую) точку $\vec{\nu}_{\vec{\xi}}$, обращающую градиент $(A(\vec{\nu}) - \vec{\xi}^{\,T}\vec{\nu})$ в ноль: $\nabla A(\vec{\nu}_{\vec{\xi}}) = \vec{\xi}$. При этом, разлагая показатель экспоненты в (П.10) с точностью до членов второго порядка по $(\vec{\nu} - \vec{\nu}_{\vec{\xi}})$, можно получить следующий приближенный результат:

$$\begin{aligned}\rho_0(\vec{\xi}) &\approx \frac{\exp(A(\vec{\nu}_{\vec{\xi}}) - \vec{\xi}^{\,T}\vec{\nu}_{\vec{\xi}})}{(2\pi\hat{i})^2} \int\limits_{-\hat{i}\infty}^{\hat{i}\infty} \int\limits_{-\hat{i}\infty}^{\hat{i}\infty} \exp\left\{\frac{1}{2}(\vec{\nu} - \vec{\nu}_{\vec{\xi}})^T \left[\nabla\nabla^T A(\vec{\nu}_{\vec{\xi}})\right](\vec{\nu} - \vec{\nu}_{\vec{\xi}})\right\} d\vec{\nu} = \\ &= \frac{\exp(A(\vec{\nu}_{\vec{\xi}}) - \vec{\xi}^{\,T}\vec{\nu}_{\vec{\xi}})}{(2\pi)^2} \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}(\vec{\eta} + \hat{i}\vec{\nu}_{\vec{\xi}})^T \left[\nabla\nabla^T A(\vec{\nu}_{\vec{\xi}})\right](\vec{\eta} + \hat{i}\vec{\nu}_{\vec{\xi}})\right\} d\vec{\eta} = \\ &= \frac{\exp\left\{A(\vec{\nu}_{\vec{\xi}}) - \vec{\xi}^{\,T}\vec{\nu}_{\vec{\xi}}\right\}}{2\pi\sqrt{\det \nabla\nabla^T A(\vec{\nu}_{\vec{\xi}})}},\end{aligned} \tag{П.11}$$

где переход к гауссовому интегралу осуществлен в результате замены переменных интегрирования $\vec{\nu} = \hat{i}\vec{\eta}$.

Учитывая, что выражение в числителе (П.11) с точностью до знака совпадает со значением сопряженной по Лежандру к $A(\vec{\nu})$ функции $A^*(\vec{\xi})$ [43]

$$A^*(\vec{\xi}) = \vec{\xi}^{\,T}\vec{\nu}_{\vec{\xi}} - A(\vec{\nu}_{\vec{\xi}}), \quad \nabla A(\vec{\nu}_{\vec{\xi}}) = \vec{\xi}, \tag{П.12}$$

и тот факт, что $A(\vec{\nu})$ и $A^*(\vec{\xi})$ имеют взаимно обратные матрицы Гесса: $\nabla\nabla^T A^*(\vec{\xi}) = [\nabla\nabla^T A(\vec{\nu}_{\vec{\xi}})]^{-1}$ получим приближение седловой точки для плотности $\rho_0(\vec{\xi})$ [43]

$$\rho_0(\vec{\xi}) \approx \frac{\sqrt{\det \nabla\nabla^T A^*(\vec{\xi})}}{2\pi} \exp\left\{-A^*(\vec{\xi})\right\}. \tag{П.13}$$

Соотношение (П.13) можно переписать в эквивалентном виде

$$\rho_0(\vec{\xi}) \exp\left\{A^*(\vec{\xi})\right\} \approx \frac{1}{2\pi}\sqrt{\det \nabla\nabla^T A^*(\vec{\xi})}, \tag{П.14}$$

который позволяет записать приближенное выражение для левой части через квадратный корень из определителя матрицы Гесса $\nabla\nabla^T A^*(\vec{\xi})$.

Дальнейшее приближение может быть получено, если вспомнить, что $A(\vec{\nu})$ является производящей функцией кумулянтов для $\rho_0(\vec{\xi})$. Это значит, что $\nabla\nabla^T A(\vec{0}) = R_0$ — корреляционная функция $\rho_0(\vec{\xi})$, определитель которой приближенно равен кварату площади Δ. Тем самым $\left[\det \nabla\nabla^T A^*(\vec{0})\right]^{-1} = \det \nabla\nabla^T A(\vec{0}) = \det R_0 \approx |\Delta|^2 \approx D^4$, где D — характерный размер Δ. Полагая $\det \nabla\nabla^T A^*(\vec{\xi}) \approx \det \nabla\nabla^T A^*(\vec{0}) \approx D^{-4}$, можно приблизить правую часть (П.14) следующим выражением:

$$\rho_0(\vec{\xi}) \exp\left\{A^*(\vec{\xi})\right\} \approx \frac{1}{2\pi D^2}. \tag{П.15}$$

СПИСОК ЛИТЕРАТУРЫ

1. *Ezhilraman V., Srinivasan S.* State of the art in image processing & big data analytics: issues and challenges // International J. of Engineering & Technology. 2018. V. 7. P. 195–199. https://doi.org/10.14419/ijet.v7i2.33.13885
2. *Bull D.R., Zhang F.* Intelligent image and video compression: communicating pictures. 2nd ed. London: Academic Press, 2021.
3. *Zeyu Y., Fei W., Rendong Y., et al.* On Perceptual Lossy Compression: The Cost of Perceptual Reconstruction and An Optimal Training Framework // Proc. of the 38-th International Conference on Machine Learn., PMLR. 2021.
   https://doi.org/10.48550/arXiv.2106.02782
4. *Shannon C.E.* Coding Theorems for a Discrete Source with a Fidelity Criterion – Institute of Radio Engineers, International Convention Record. V. 7, 1959 // in C.E. Shannon: Collected Papers. 1993. P. 325–350.
   https://doi.org/10.1109/9780470544242.ch21

5. *Tschannen M., Agustsson E., Lucic M.* Deep Generative Models for Distribution-Preserving Lossy Compression // Proc. of the 32nd International Conference on Neural Inform. Processing Systems (NIPS). 2018. P. 5933–5944.
6. *Blau Y., Michaeli T.* Rethinking Lossy Compression: The Rate–Distortion–Perception Tradeoff // Proc. of the 36th International Conference on Machine Learn., PMLR. 2019. V. 97. P. 675–685.
7. *Matsumoto R.* Introducing the perception–distortion tradeoff into the rate–distortion theory of general information sources // IEICE Commun. Express. 2018. V. 7. No. 11. P. 427–431. https://doi.org/10.1587/comex.2018XBL0109
8. *Wang Z., Bovik A., Sheikh H., Simoncelli E.* Image quality assessment: from error visibility to structural similarity // IEEE Trans. Image Process. 2004. V. 13. No. 4. P. 600–612. https://doi.org/10.1109/TIP.2003.819861
9. *Wang Z., Bovik A.* Video quality assessment based on structural distortion measurement // Signal processing. Image commun. 2004. V. 19. No. 2. P. 121–132. https://doi.org/10.1016/S0923–5965(03)00076-6
10. *Sheikh H., Bovik A., de Veciana G.* An information fidelity criterion for image quality assessment using natural scene statistics // IEEE Trans. on image processing. 2005. V. 14. No. 12. P. 2117–2128. https://doi.org/10.1109/TIP.2005.859389
11. *Larson E.C., Chandler D.M.* Most apparent distortion: full–reference image quality assessment and the role of strategy // J. Electron. Imaging. 2010. V. 19. No. 1. P. 011006–011006. https://doi.org/10.1117/1.3267105
12. *Bishop C.M., Lasserre J.* Generative or Discriminative? Getting the Best of Both Worlds // Bayes. Statist. 2007. V. 8. P. 3–24.
13. *Goodfellow I., Pouget–Abadie J., Mirza M., et al.* Generative Adversarial Networks // Commun. ACM. 2020. V. 63. No. 11. P. 139–144. https://doi.org/10.1145/3422622
14. *Kingma D.P., Welling M.* Auto–Encoding Variational Bayes // arXiv:1312.6114. arxiv.org. 2013.
15. *Hinton G.E., Osindero S., The Y.-W.* A Fast-Learning Algorithm for Deep Belief Nets // Neural Computation. 2006. V. 18. No. 7. P. 1527–1554. https://doi.org/10.1162/neco.2006.18.7.1527
16. *Hassabis D., Kumaran D., Summerfield C., Botvinick M.* Neuroscience-Inspired Artificial Intelligence // Neuron. 2017. V. 95. No. 2. P. 245–258. https://doi.org/10.1016/j.neuron.2017.06.011
17. *Antsiperov V.E.* Representation of Images by the Optimal Lattice Partitions of Random Counts // Patt. Recogn. and Image Anal. 2021. V. 31. No. 3. P. 381–393. https://doi.org/10.1134/S1054661821030044
18. *Antsiperov V.E., Kershner V.A.* Image Coding by Count Sample, Motivated by the Mechanisms of Light Perception in the Visual System // Commun. Comput. Inform. Sci. 2022. V. 1534. P. 715–729. https://doi.org/10.1007/978-3-030-96040-7_54
19. *Scott D.W.* Multivariate Density Estimation. Hoboken: John Wiley & Sons, Inc. 1992. https://doi.org/10.1002/9780470316849
20. *Rufo M.J., Martin J., Perez C.J.* Bayesian analysis of finite mixture models of distributions from exponential families // Comput. Statist. 2006. V. 21. No. 3–4. P. 621–637. https://doi.org/10.1007/s00180-006-0018-8

21. *McLachlan G.J., Krishnan T.* The EM Algorithm and Extensions. 2nd ed. Hoboken: John Wiley & Sons, Inc. 2007.
22. *Tzikas D., Likas A., Galatsanos N.* The variational approximation for Bayesian inference // IEEE Signal Proc. Magazine. 2008. V. 25. No. 6. P. 131–146. https://doi.org/https://doi.org/10.1109/msp.2008.929620
23. *Banerjee A., Merugu S., Dhillon I.S., Ghosh J.* Clustering with Bregman Divergences // J. Machine Learn. Res. 2005. V. 6. P. 1705–1749.
https://doi.org/10.1137/1.9781611972740.22
24. *Linde Y., Buzo A., Gray R.M.* An algorithm for vector quantizer design // IEEE Trans. Commun. 1980. V. 28. No. 1. P. 84–95.
https://doi.org/10.1109/TCOM.1980.1094577
25. *Lloyd S.* Least squares quantization in PCM // IEEE Trans Inform. Theory. 1982. V. 28. No. 2. P. 129–137. https://doi.org/10.1109/TIT.1982.1056489
26. *Kohonen T.* Self-Organized Formation of Topologically Correct Feature Maps // Biolog. Cybernet. 1982. V. 43. No. 1. P. 59–69. https://doi.org/10.1007/bf00337288
27. *Barrett H.H., Myers K.J.* Foundations of image science. Hoboken: John Wiley & Sons, Inc. 2004.
28. *Fossum E.* The Invention of CMOS Image Sensors: A Camera in Every Pocket // 2020 Pan Pacific Microelectronics Symposium (Pan Pacific). 2020. P. 1–6.
https://doi.org/10.23919/PanPacific48324.2020.9059308
29. *Gabriel C.G., Perrinet L., Keil M.* Biologically Inspired Computer Vision: Fundamentals and Applications. Weinheim: Wiley-VCH. 2015.
30. *Fox M.* Quantum Optics: An Introduction. Oxford, New York: Oxford University Press, 2006.
31. *Streit R.L.* Poisson Point Processes. Imaging, Tracking and Sensing. New York: Springer. 2010.
32. *Bertero M., Boccacci P., Desidera G., Vicidomini G.* Image deblurring with Poisson data: from cells to galaxies // Inverse Problems. 2009. V. 25. No. 12. P. 123006.
https://doi.org/10.1088/0266-5611/25/12/123006
33. *Robert C.P., Casella G.* Monte Carlo Statistical Methods. 2nd ed. New York: Springer–Verlag. 2004. https://doi.org/10.1007/978-1-4757-4145-2
34. *Hinton G.E., Zemel R.S.* Autoencoders, minimum description length and Helmholtz free energy // Proc. of the 6th International Conference on Neural Inform. Processing Systems (NIPS'93). 1993. P. 3–10.
35. *Goodfellow I., Bengio Y., Courville A.* Autoencoders // Deep Learning. MIT Press. 2016.
36. *Baldi P.* Autoencoders, unsupervised learning and deep architectures // JMLR: Workshop and Conference Proceedings. 2012. V. 27. P. 37–49.
37. *Alain G., Bengio Y., Yao L., et.al.* GSNs: Generative Stochastic Networks // arXiv:1503.05571. arXiv.org. 2015.
38. *Aldrich J.* R.A. Fisher and the Making of Maximum Likelihood 1912–1922 // Statistical Science. 1997. V. 12. No. 3. P. 162–176.
39. *Van Erven T.T., Harremoes P.* Renyi Divergence and Kullback-Leibler Divergence // IEEE Trans. on Inform. Theory. 2014. V. 60. No. 7. P. 3797–3820.
https://doi.org/10.1109/TIT.2014.2320500

40. *Schiller P.H., Tehovnik E.J.* Vision and the Visual System // Oxford: Oxford University Press. 2015. https://doi.org/10.1093/acprof:oso/9780199936533.001.0001

41. *Cooler S., Schwartz G.W.* An offset ON-OFF receptive field is created by gap junctions between distinct types of retinal ganglion cells // Nat Neurosci. 2021. V. 24. P. 105–115. https://doi.org/10.1038/s41593-020-00747-8

42. *Young R.A.* Oh say, can you see? The physiology of vision // Proc. of SPIE. 1991. V. 1453. No. 1. P. 92–123. https://doi.org/10.1117/12.44348

43. *McCullagh P.* Tensor methods in statistics. London, New York: Chapman and Hall/CRC. 1987. https://doi.org/10.1201/9781351077118

44. *Brown L.D.* Fundamentals of Statistical Exponential Families // Hayward IMS. 1986.

45. *Frigyik A.B., Srivastava S., Gupta M.R.* Functional Bregman divergence // IEEE Int. Symposium on Inform. Theory. 2008. P. 1681–1685.
https://doi.org/10.1109/ISIT.2008.4595274

46. *Rheinbold W.C.* Methods for Solving Systems of Nonlinear Equations. 2nd ed. Society for Industrial and Applied Mathematics. 1998.

47. *Ortega J.M., Rheinboldt W.C.* Iterative solution of nonlinear equations in several variables. Society for Industrial and Applied Mathematics. 2000.

48. *Redner R.A., Walker H.F.* Mixture Densities, Maximum Likelihood and the EM Algorithm // SIAM Review. 1984. V. 26. No. 2. P. 195–239.
https://doi.org/10.1137/1026034

49. *Bridle J.S.* Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition // Neurocomputing. 1990. V. 68. P. 227–236. https://doi.org/10.1007/978-3-642-76153-9_28

50. *Dhillon I., Mallel S., Kumar R.* A divisive information–theoretic feature clustering algorithm for text classification // J. of Machine Learn. Res. 2003. V. 3. No. 4. P. 1265–1287. https://doi.org/10.1162/153244303322753661

51. *Reid N.* Saddlepoint Methods and Statistical Inference // Statistical Science. 1988. V. 3. No. 2. P. 213–227. https://doi.org/10.1214/ss/1177012906

52. *Ditkin V.A., Prudnikov A.P.* Operational calculus in two variables and its applications. New York: Dover Publications, Inc. 2017.

Статья представлена к публикации членом редколлегии А.А. Лазаревым.

Поступила в редакцию 02.02.2022

После доработки 24.06.2022

Принята к публикации 29.06.2022