

# Generative Model for Autoencoders Learning by Image Sampling Representations

V. E. Antsiperov <sup>a</sup>

*Kotelnikov Institute of Radioengineering and Electronics of RAS, Mokhovaya 11-7, Moscow, Russian Federation*

**Keywords:** Ideal Image, Counting Statistics, Autoencoders, Generative Model, Machine Learning, Feature Extraction.

**Abstract:** The article substantiates a generative model for autoencoders, learning by the input image representation based on a sample of random counts. This representation is used instead of the ideal image model, which usually involves too cumbersome descriptions of the source data. So, the reduction of the ideal image concept to sampling representations of fixed (controlled) size is one of the main goals of the article. It is shown that the corresponding statistical description of the sampling representation can be factorized into the product of the distributions of individual counts, which fits well into the naive Bayesian approach and some other machine learning procedures. Guided by that association the analogue of the well-known EM algorithm – the iterative partition–maximization procedure for generative autoencoders is synthesized. So, the second main goal of the article is to substantiate the partition–maximization procedure basing on the relation between autoencoder image restoration criteria and statistical maximum likelihood parameters estimation. We succeed this by modelling the input count probability distribution by the parameterized mixtures, considering the hidden mixture variables as autoencoder’s internal (coding) data.

## 1 INTRODUCTION


Machine learning methods have been attracting the attention of researchers for more than half a century. The first methods and approaches were largely borrowed / adapted from the statistical theory, whose foundations were established about a hundred years ago, primarily by the works of R. Fisher. Discoveries in related fields in the middle of the XX-th century, primarily in neuropsychology, greatly influenced the development and originality of machine learning. We note in this connection the McCulloch and Pitts model of the neuron (1943) and the Hebb’s rule for the perceptron (1949). This was followed by a relatively long period of experience accumulation and analysis of the possibilities of implementing network learning methods on computers. A breakthrough in this direction was the invention in the mid-1980s by Rumelhart, Hinton, and Williams of the error backpropagation algorithm for training neural networks (1986).

Over the past 30-40 years, the evolution of neural networks has come a long way. Along with the supervised approaches, unsupervised ones began to

develop, deep learning is gaining more importance. Under the influence of these trends, several new classes of neural networks have appeared and developed. Note here DBNs (Deep Belief Networks), CNNs (convolutional neural networks), RNNs (recurrent neural networks), as well as LSTMs (Long Short-Term Memory) and AEs (autoencoders).

Without exaggeration, autoencoders are at the forefront of unsupervised learning. This is partly due to the symmetry of their architecture, which is a coupled codec pair. In non-AE approaches, where either an encoder or a decoder is absent, expensive optimization algorithms should be used to find the code or sampling techniques to achieve restoration. In contrast to them, AE contains both elements in its structure, moreover, encoder and decoder actively influence the solution of each other's problems.

In this work, we propose a new approach to learning AE by the images presented in a special way – by special sampling representations. The first half of the paper discusses in detail the relation of sampling representation with the ideal image model. The second part of the paper is devoted to learning AE in generative model using the sampling

<sup>a</sup> <https://orcid.org/0000-0002-6770-1317>

representations at the input. The questions of generative model formalization, encoding/decoding procedures optimization and their connection to the method of maximum likelihood estimation in statistical theory are considered in detail.

## 2 IMAGE SAMPLING REPRESENTATION

In several previous papers (Antsiperov, 2021 a, b) we proposed the representation of images by the samples of random counts (basing on counting statistics (Fox, 2006)). This approach was partially substantiated in (Seitz, 2011) by the physical mechanisms of the real image formation and detection.

Today's relevance of the proposed representation is due, on the one hand, to progress in the SPAD (single photon avalanche diodes) video matrixes, that register radiation in the form of a discrete set of photocounts (Fossum, 2020), (Morimoto, 2020). On the other hand, it is due to the ever-increasing trends in the adaptation of human visual perception mechanisms for digital image processing (Beghdadi 2013), (Rodieck, 1998).

Both of the tendencies mentioned incorporate several common features. The SPAD-matrixes, as well as human retina include some sensitive 2D-surface, which consists of a very large number of detectors/receptors. These detectors are so small, that each of them can detect the individual photon of the incident radiation. The detailed, comprehensive review of these trends in modern photon-counting sensors can be found in the book (Fossum, 2017). The use of visual perception mechanisms for digital image processing is widely discussed in (Gabriel, 2015). So, the listed features can be taken as the basis of the concept of the *ideal imaging device*, generalizing besides the photon-counting sensors mentioned also the photographic plates with gelatin-silver emulsion, etc.

Formally, the definition of an ideal imaging device is as follows. It is a two-dimensional surface  $\Omega$  with coordinates  $\vec{x} = (x_1, x_2)$ , on which identical point detectors are allocated close to each other. Point detectors, or in terms of (Fossum, 2020) "jots", have by a definition a vanishingly small area  $da$  of light-sensitive surfaces. Accordingly, if the total number of detectors is  $N$ , then the total area of surface  $\Omega$  is equal to  $A = Nda$ . Under the assumption that  $A$  is fixed and  $da \rightarrow 0$ , the number  $N$  is assumed to be arbitrarily large:  $N \rightarrow \infty$ .

When the ideal imaging device registers the stationary radiation with intensity  $I(\vec{x})$ ,  $\vec{x} \in \Omega$ , some of point detectors generate the counts – random events that in the case of SPAD matrixes are the releases of an electron from  $p-n$  junction of the photodiode, in the case of retina – activations of rhodopsin molecules in photoreceptors and in the case of photographic plates – appearing of metallic silver atom clusters in or on a silver halide crystal. Within the framework of the semiclassical theory of interaction between radiation and matter, the counts are associated with incident photons captured by atoms / molecules of the detector's material. At the limit  $da \rightarrow 0$  the probability of a count (in any interpretation) for a given point detector is the product  $\alpha TI(\vec{x})da$ , where  $\alpha = \eta(h\bar{\nu})^{-1}$ ,  $h\bar{\nu}$  – the average energy of the incident photon ( $h$  – Planck's constant,  $\bar{\nu}$  – characteristic frequency of radiation), the dimensionless coefficient  $\eta < 1$  is the quantum efficiency of the detector's material (Fox, 2006),  $T$  is exposure time. Thus, when the incident radiation of intensity  $I(\vec{x})$  is registered, the state of each point detector  $\vec{x} \in \Omega$  can be described by a binary random variable  $\sigma_{\vec{x}} \in \{0, 1\}$ , taking the values  $\sigma_{\vec{x}} = 1$  and  $\sigma_{\vec{x}} = 0$ , depending on whether has the detector generate a count. The conditional (at a given intensity  $I(\vec{x})$ ) probabilities of  $\sigma_{\vec{x}}$  have the form of Bernoulli distribution:

$$\begin{aligned} P(\sigma_{\vec{x}} = 1 | I(\vec{x})) &= \alpha TI(\vec{x})da, \\ P(\sigma_{\vec{x}} = 0 | I(\vec{x})) &= 1 - \alpha TI(\vec{x})da. \end{aligned} \quad (1)$$

Note that, according to (1), formally, the mean number of counts for given point detector  $\vec{x}$  is equal to  $\bar{\sigma}_{\vec{x}} = \alpha TI(\vec{x})da$  (it is assumed, that  $\bar{\sigma}_{\vec{x}} < 1$ ). Accordingly, the integral  $\bar{n} = \sum_{\vec{x} \in \Omega} \bar{\sigma}_{\vec{x}} = \alpha T \iint_{\Omega} I(\vec{x})da$  defines the mean number of all registered over time  $T$  counts.

Based on the concept of an ideal imaging device and considering the main features of its registration mechanism (1), it is possible to formulate a model of an *ideal image* as a resultant set of counts, generated during the registration process. Namely, under the ideal image we mean the (ordered) set  $X = (\vec{x}_1, \dots, \vec{x}_n)$ ,  $\vec{x}_i \in \Omega$  of  $n$  random counts registered ( $\sigma_{\vec{x}_i} = 1$ ) by the ideal imaging device during the time  $T$ . Thus, an ideal image is a kind of random object, a random set of count coordinates  $\vec{x}_i \in \Omega$ , which should be distinguished from any of its realization. We use the name "ideal image" for the proposed construction, following the authors of (Pal, 1991), in which they introduced this term for the first time in the early 90s. It is worth noting that the randomness of the ideal image is related not only to the random

nature of count coordinates  $\vec{x}_i$ , but also it is determined by the random value of  $n$  – the number of counts in the set.

A complete statistical description of ideal image  $X$  in the form of finite-dimensional probability distribution densities  $\{\rho(\vec{x}_1, \dots, \vec{x}_n, n|I(\vec{x}))\}$ ,  $\vec{x}_i \in \Omega, n = 0, 1, \dots$  can be obtained by assuming conditional independence of all counts  $\vec{x}_i$  (under the condition of the given  $I(\vec{x})$  and  $n$ ). It is well known (Poisson's theorem, (Gallager, 2013)), that under such assumptions, asymptotically, with  $N \rightarrow \infty$  the probability distribution of Bernoulli process (trial)  $\{\sigma_{\vec{x}}\}$ ,  $\vec{x} \in \Omega$  (1) converges to the Poisson process distributions (Streit, 2010) on  $\Omega$ :

$$\rho(\vec{x}_1, \dots, \vec{x}_n, n|I(\vec{x})) = \prod_{i=1}^n \rho(\vec{x}_i|I(\vec{x})) \times P_n(W) \tag{2}$$

where

$$P_n(W) = \frac{(\alpha TW)^n}{n!} \exp(-\alpha TW),$$

$$\rho(\vec{x}_i|I(\vec{x})) = \frac{I(\vec{x}_i)}{W}, W = \iint_{\Omega} I(\vec{x}) da$$

where  $P_n(W)$  is the Poisson probability distribution (Gallager, 2013), (Streit, 2010) of  $n$  – number of counts in ideal image  $X$ ,  $\bar{n} = \alpha TW$  is its mean,  $W$  is the overall registered radiation power. Conditional  $\rho(\vec{x}|I(\vec{x}))$  is the density of single count probability distribution. In connection with (2), it is interesting to note that the (statistical) intensity of 2D point Poisson process  $\bar{\sigma}_{\vec{x}}/da$  coincides with the physical intensity  $I(\vec{x})$  of the recorded radiation up to the constant  $\alpha T$ .

From the theoretical point of view the concept of an ideal image is a very attractive statistical object due to the simplicity of its statistical description (2) and its interpretation as an inhomogeneous point Poisson process that had been well studied for a long time. However, in practical tasks it is not always possible to use this concept directly in the form in which it is formulated. Namely, for common recorded radiation intensities  $I(\vec{x})$ , the direct use of the ideal image realization  $X = (\vec{x}_1, \dots, \vec{x}_n)$  would require enormous computational resources when the number of counts  $n$  is big enough. So, considering that on a clear day the flux of photons from the sun falling on a surface with  $A \sim 1 \text{ mm}^2$  per second is of order  $\sim 10^{15} - 10^{16}$  photons (Rodieck, 1998), the devices in a photon counting mode will generate the information flow of the value of  $\bar{n} \sim 10^{15}$  (1 Pbit/sec). Obviously, it is very problematic to process such information with the modern computing technique.

To avoid the "curse of dimension" of ideal image representation, we propose the following solution

(Antsiperov, 2021 a). Let us represent the image not by the complete sets of ideal image counts  $X = (\vec{x}_1, \dots, \vec{x}_n)$ , but only by some subset  $X_k = (\vec{x}_{j_1}, \dots, \vec{x}_{j_k})$ ,  $j_i \in \{1, \dots, n\}$  of acceptable fixed size  $k$ , where  $\vec{x}_{j_i}$  are randomly selected counts from  $X$ . Formally, considering  $X$  as a general population of counts, we propose to use only a random sample  $X_k$  of them to represent the image. Obviously, in full agreement with the classical statistical paradigm, such a "sample" representation will still represent the ideal image  $X$ . Let us name such a sample  $X_k$  "representation by a sample of random counts" or, in short, the *sampling representation*.

The statistical description of sampling representation follows easily from (2) by integrating density  $\rho(\vec{x}_1, \dots, \vec{x}_n, n|I(\vec{x}))$  over the not selected in  $X_k$  count coordinates and summing the result over the number  $l$  of not selected counts. In the actual case  $1 \ll k \ll \bar{n}$  the statistical description of sampling representation  $X_k$  is given with high accuracy by the probability distribution density of the form:

$$\rho(X_k|I(\vec{x})) = \prod_{j=1}^k \rho(\vec{x}_j|I(\vec{x})),$$

$$\rho(\vec{x}_j|I(\vec{x})) = \frac{I(\vec{x}_j)}{W}, W = \iint_{\Omega} I(\vec{x}) da. \tag{3}$$

Regarding description (3), it should be noted that it has very simple structure, depending only on the shape (normalized version) of the registered intensity  $I(\vec{x})/W$ . This immediately leads to some attractive representation properties. First, it reveals the conditional independence of all  $k$  counts  $\vec{x}_j$  and their identical conditional distribution (iid property). Second, the density of the individual count  $\rho(\vec{x}_j|I(\vec{x}))$  is very simply related to intensity  $I(\vec{x})$  of radiation – they are proportional to each other. Third, the description satisfies the following universality property: it does not depend either on the quantum efficiency of the detector material  $\eta$ , or on the incident radiation average frequency  $\bar{\nu}$ , or on frame time  $T$ . These sampling presentation properties provide a convenient, suitable input data for many well-developed statistical and machine learning approaches, including the naive Bayesian approach (Barber, 2012).

A consequence of the universality property is also the fact that statistical description of the sampling representation (3) does not depend on physical units of intensity  $I(\vec{x})$ . So, if, for example, the intensity  $I(\vec{x})$  is given by pixels of some bitmap image, obtained by digitization with a quantization parameter  $Q = \Delta I$ , then description (4) will not directly dependent on  $Q$ , but it will depend only on



the pixel color depth  $v$ . This remark allows to generate sampling representations not only for the source data of the photon counting sensors, but also for common digital images obtained by traditional camera. To do this, one should use the all-image pixels  $\{n_i\}$  to form an approximation  $n_i/\sum n_i \approx I(\vec{x}_i)/W$  of the original normalized intensity and then simulate the process of sampling independent counts from it according (3). Regarding the computational organization of sampling, fortunately, in the field of machine learning, there is a large arsenal of methods, collectively called Monte Carlo methods (Murphy, 2012), that can do this very efficiently.

It should be noted that for some methods even pixels  $\{n_i\}$  normalization is not required at all – it is sufficient that all pixels are bounded from above by the constant  $2^v$ , where  $v$  is the color depth of the image. For example, Figure 1 shows count representations of the picture “cameraman” (distributed by the MathWorks, Inc. with permission from the MIT), presented in subfigure A. The sampling representations with sizes 500 000, 1 000 000 and 5 000 000, 1 000 000 and 5 000 000 counts – B, C and D were carried out by one of the simplest sampling procedures – acceptance-rejection method using a uniform proposal distribution.

### 3 LEARNING AE BY SAMPLING REPRESENTATIONS

Usually, autoencoders (AE) are considered as a special class of artificial neural networks (ANNs) (Hinton, 1994), but for our purposes it is desirable to define them from a more general point of view. Namely, we will consider AEs as a special class of information systems, understood as an “integrated set of components for collecting, storing and processing data” (Information system. In Encyclopedia Britannica, 2020). In current context, the processing data means the images. As usual, AEs have a symmetric three-tier input-code-output structure, as shown in Figure 2, where the middle tier is for encoding the input data. Pairs of adjacent tiers make up two reciprocal components: input-code as an encoder and code-output as a decoder (Goodfellow, 2016). The goal of AE is to restore the input data to the output, while observing certain restrictions imposed on the internal encoding. Because of these restrictions, it is not allowed to simply copy data from input to output. Typical restrictions are related to a dimensionality reduction of intermediate (coding)

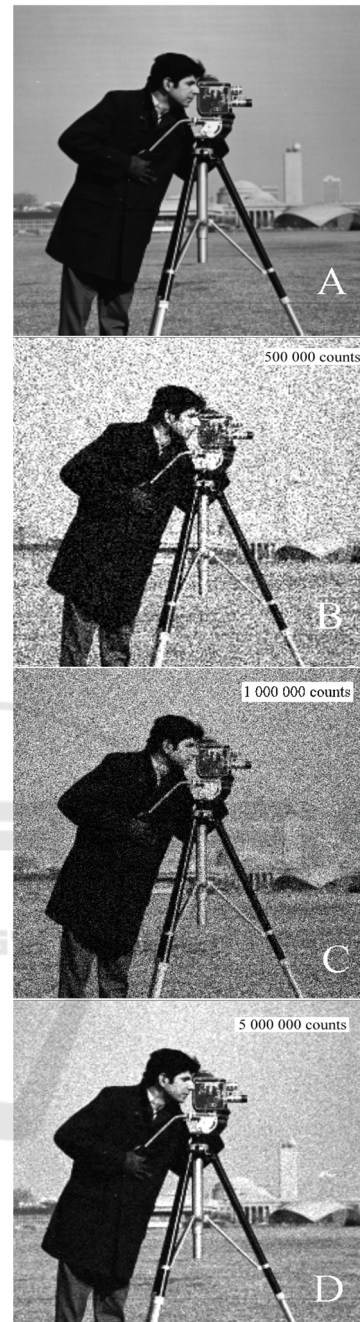


Figure 1: Representation of the “cameraman” image by samples of random counts: A – the original image in TIF format, B, C, D – sampling representations of the sizes, respectively, 500 000, 1 000 000 and 5 000 000 counts.

data, which excludes input-output bijection. The presence of such a bottleneck on the one hand and the main AE task on the other hand implies some optimal coding for intermediate data.

In the light of modern approaches, such a coding can be synthesized basing on unsupervised AE

learning. Note that although the term "autoencoder" is currently the most popular term, due to the very broad scope of given definition, it can also be used synonymously with auto-associative memory networks (Kohonen, 1989), replicator ANN (Hecht-Nielsen, 1995), etc.

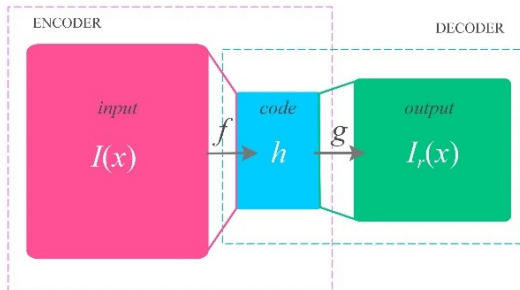


Figure 2: Schema of a basic Autoencoder.

To formalize the AE problem, let us present its general mathematical frame (Baldi, 2012). We assume the following. First, the sets  $\mathcal{G}$  of images  $I(\vec{x}), \vec{x} \in \Omega$  and  $\mathcal{F}$  of internal images representations (code)  $\vec{h}$  are given (see Figure 2). Second, the classes of operators  $f: \mathcal{G} \rightarrow \mathcal{F}$  (encoders) and  $g: \mathcal{F} \rightarrow \mathcal{G}$  (decoders), agreed in dimensions with  $\mathcal{F}$  and  $\mathcal{G}$  and with the given restrictions are specified. Third, a numerical measure of distortion between the image  $I(\vec{x})$  and some of its reconstruction  $I_r(\vec{x})$  is available – the so-called loss function  $L(I(\vec{x}), I_r(\vec{x}))$  (Goodfellow, 2016). Within this framework, the main problem of AE is to minimize the loss function with respect to the encoder  $f$  and decoder  $g$  operators:

$$\{f^*, g^*\} = \arg \min_{f, g} L(I(\vec{x}), g \circ f(I(\vec{x}))). \quad (4)$$

Any solution  $f^*$  (5) would be considered as the desired coding for the optimal restoration  $g^*$  of the image. Unfortunately, solving (4) in its most general form is an unrealistic task. Therefore, in the study of practical problems, it is necessary to specify the elements of the general AE framework. Different kinds of AEs can be derived depending on the choice of sets  $\mathcal{G}$  and  $\mathcal{F}$ , special classes of operators  $f$  and  $g$  and the explicit form of loss function  $L$ . If, for example,  $\mathcal{G}$  and  $\mathcal{F}$  are linear spaces of dimensions  $n$  and  $p$  respectively,  $f$  and  $g$  are appropriate linear operators ( $(n \times p)$  and  $(p \times n)$  matrixes) and  $L$  is a  $L^2$  norm  $\|I(\vec{x}) - I_r(\vec{x})\|_2^2$  in  $\mathcal{G}$ , we get a linear autoencoder. It is interesting to note that a linear autoencoder results in the same internal data representation  $\vec{h}$  as the principal component analysis (PCA) (Plaut, 2018). Moreover, it is easy to

generalize the PCA to nonlinear NLPCA, if weaken the linearity condition for the encoder  $f: \mathcal{G} \rightarrow \mathcal{F}$ . Such (nonlinear) AEs can learn a non-linear manifolds for coded data instead of finding a low dimensional approximating hyperplane.

In our case, the images are specified by sampling representations  $X_k = (\vec{x}_i), i = 1, \dots, k$ , generated according to the probability distribution density of the counts  $\rho(\vec{x}|I(\vec{x}))$ , which is uniquely related to the registered intensity  $I(\vec{x})$  (3). So, it is quite reasonable to consider the set  $\mathcal{G}$  as the set of probability densities  $\{\rho(\vec{x})\}$  on the image surface  $\vec{x} \in \Omega$ . This immediately brings us to the generative models for autoencoders (Goodfellow, 2016). In contrast to the traditional AE, which are most naturally interpreted as the regularization schemes, autoencoders in generative paradigm consider the internal encoded data  $\vec{h}$  as latent variables and the coding operation  $f: \mathcal{G} \rightarrow \mathcal{F}$  as inference procedure (computing latent representation for given  $X_k$ ). In this regard, generative models learn to maximize the likelihood of  $I(\vec{x})$  conditioned by input data (representation)  $X_k$ , rather than copying inputs to outputs. So, regularization issues don't matter much for generative AE. As an example, a couple of generative modeling approaches to autoencoders can be mentioned here – the variational autoencoder (VA) (Kingma, 2014) and the generative stochastic network (GSN) (Alain, 2015).

To formalize the generative model for sampling representations  $X_k = (\vec{x}_i)$ , let us consider the first set  $\mathcal{G}$  as some parametric family  $\mathcal{G} = \{\rho(\vec{x} | \vec{\theta})\}, \vec{x} \in \Omega, \vec{\theta} \in \Theta \subset \mathbb{R}^p$  of probability distribution densities of individual count  $\vec{x}$ . The parameters  $\vec{\theta} \in \Theta$  of representation are associated with the unknown normalized intensity  $I(\vec{x})$  and are intended for its parametric approximation. Parametrization of the distributions under study  $\{\rho(\vec{x}|I(\vec{x}))\}$  is a common technique that simplifies the problem of functional optimization to a problem of optimal parameters estimation. The unsupervised learning for generative model AE consists in fitting  $\rho(\vec{x} | \vec{\theta}^*) \in \mathcal{G}$  for some  $\vec{\theta}^*$ , considering as AE output, to a training data  $X_k$ . Note, that formally, sampling representation  $X_k$  is not an element of  $\mathcal{G}$  and it can't be considered as the input of AE. Instead of it, we should utilize conditional probability density  $\rho(\vec{x} | X_k = (\vec{x}_i))$ . Assuming, in the spirit of the Bayesian approach, that the parameters  $\vec{\theta}$  are random variables with any prior probability distribution density  $P(\vec{\theta})$ , similarly to how it was done in (Antsiperov, 2021b), we can write the following expression for the conditional density:

$$\begin{aligned}
 \rho(\vec{x} | X_k = (\vec{x}_i)) &= \frac{\rho(\vec{x}, \vec{x}_1, \dots, \vec{x}_k)}{\rho(\vec{x}_1, \dots, \vec{x}_k)} = \\
 &= \frac{\iint_{\Theta} \rho(\vec{x}, \vec{x}_1, \dots, \vec{x}_k | \vec{\theta}) P(\vec{\theta}) d\vec{\theta}}{\rho(\vec{x}_1, \dots, \vec{x}_k)} = \\
 &= \frac{\iint_{\Theta} \rho(\vec{x} | \vec{\theta}) \rho(\vec{x}_i | \vec{\theta}) P(\vec{\theta}) d\vec{\theta}}{\rho(\vec{x}_1, \dots, \vec{x}_n)} = \\
 &= \iint_{\Theta} \rho(\vec{x} | \vec{\theta}) \rho(\vec{\theta} | X_k = (\vec{x}_i)) d\vec{\theta}
 \end{aligned} \quad (5)$$

where we used the property of the conditional iid of all samples  $\vec{x}, \vec{x}_1, \dots, \vec{x}_k$  (for a given  $\vec{\theta}$  or, which is the same, for a given  $I(\vec{x})$ ). As it is known, density  $\rho(\vec{\theta} | X_k = (\vec{x}_i))$  is, at least asymptotically  $k \gg 1$ , a much narrower function than  $\rho(\vec{x} | \vec{\theta})$ . Since the maximum of  $\rho(\vec{\theta} | X_k = (\vec{x}_i))$  coincides with the maximum likelihood estimate  $\vec{\theta}_{ML}$ , the density  $\rho(\vec{x} | \vec{\theta})$  can be taken out from the integral in (5) as the independent of  $\vec{\theta}$  factor  $\rho(\vec{x} | \vec{\theta}_{ML})$ . This immediately leads to the following (see also (Antsiperov, 2021b)):

$$\rho(\vec{x} | X_k = (\vec{x}_i)) \cong \rho(\vec{x} | \vec{\theta}_{ML}). \quad (6)$$

Coming back, we can use at the input of AE the density  $\rho(\vec{x} | \vec{\theta}_{ML}) \in \mathcal{G}$ , which accumulates all the necessary information of the representation  $X_k = (\vec{x}_i)$  by means of a statistic  $\vec{\theta}_{ML}(\vec{x}_i)$ , that is a solution of R.A. Fisher's maximum likelihood equation (Aldrich, 1997):

$$\begin{aligned}
 \vec{\theta}_{ML} &= \arg \max_{\vec{\theta} \in \Theta} L(\vec{\theta}; X_k), \\
 L(\vec{\theta}; X_k) &= \rho(X_k | \vec{\theta}) = \prod_{i=1}^n \rho(\vec{x}_i | \vec{\theta}).
 \end{aligned} \quad (7)$$

Considering the developed generative model formalization, it seems, that conceptually the solution of the AE main problem becomes straightforward. Namely, this solution consists in forming the density  $\rho(\vec{x} | \vec{\theta}_{ML}) \in \mathcal{G}$  at the input of the autoencoder, using the sample representation  $X_k$  (by generating sample statistics  $\vec{\theta}_{ML}(X_k)$ ) and transmitting it in some coded form to the output. Obviously, formed in such a manner the input and output densities provide the minimum for any suitable loss function  $L(\rho(\vec{x} | \vec{\theta}_{ML}), \rho(\vec{x} | \vec{\theta}^*))$ , if the appropriate encoding-decoding procedures guarantee  $\vec{\theta}^* \sim \vec{\theta}_{ML}$ .

The seeming elegance of solving AE problem within the generative model framework is associated with the replacement of the input data coding problem by the problem of calculating maximum likelihood estimate (7). However, the problem (7), which has

been known for a hundred years, starting with Fisher's works (see (Aldrich, 1997)), in real applications turns out not much simpler, than the coding problem. Moreover, the development of such modern direction as machine learning (including autoencoders) was largely due to the needs of an approximate solution of the maximum likelihood problem. For this reason, to further refine the generative model, we will develop the "proper autoencoder" method for solving the main problem of the AE, considering it as a special class of methods for solving the maximum likelihood equation (7).

The main assumption for our special generative model is that the parametric family  $\mathcal{G} = \{\rho(\vec{x} | \vec{\theta})\}$  admits latent (hidden) variables. Let us consider the simplest case, where each count  $\vec{x}$  is associated with a single latent variable  $j$ , that takes only a finite set of discrete values:  $j \in \{1, \dots, K\}$ . Let us denote the density of joint probability distribution of  $\vec{x}$  and  $j$  by  $\rho(\vec{x}, j | \vec{\theta})$ . In what follows, we implicitly assume that  $\rho(\vec{x}, j | \vec{\theta})$  is more tractable than  $\rho(\vec{x} | \vec{\theta})$ , since the latter is a marginal distribution of the former:

$$\rho(\vec{x} | \vec{\theta}) = \sum_{j=1}^K \rho(\vec{x}, j | \vec{\theta}) \quad (8)$$

and the sum in (8) can contain a large number  $K$  of terms. The density (9) is generally called the finite mixture of components and traditionally components are written in the form  $\rho(\vec{x}, j | \vec{\theta}) = w_j \rho_j(\vec{x} | \vec{\theta})$ , where  $w_j = \rho(j | \vec{\theta})$  is the weight (probability) of the  $j$ -component, and  $\rho_j(\vec{x} | \vec{\theta})$  is the conditional distribution of the count coordinates  $\vec{x}$  for the component  $j$ . As a rule, component weights are considered as a subset of a parameters:  $\{w_j\} \subset \Theta$ .

In accordance with (8), the density of the joint distribution of the sampling representation  $X_k = (\vec{x}_i)$  (3) can be written in the form:

$$\begin{aligned}
 \rho(X_k | \vec{\theta}) &= \prod_{i=1}^k [\sum_{j_i=1}^K \rho(\vec{x}_i, j_i | \vec{\theta})] = \\
 \sum_{\vec{h}} [\prod_{i=1}^k \rho(\vec{x}_i, j_i | \vec{\theta})] &= \sum_{\vec{h}} \rho(X_k, \vec{h} | \vec{\theta}).
 \end{aligned} \quad (9)$$

where the ordered set (vector)  $\vec{h} = (j_1, \dots, j_k)$ ,  $j_i \in \{1, \dots, K\}$  represents the latent variables of autoencoder, i. e. the inner representation of  $\rho(X_k | \vec{\theta})$ , defined on the  $n$ -cube  $\mathcal{F} = \{1, \dots, K\}^k$ . It is easy to see that the density (6) is also a finite mixture of components with component weights  $W_{\vec{h}} = \prod_{i=1}^n w_{j_i}$  and conditional distributions  $\rho_{\vec{h}}(X_k | \vec{\theta}) = \prod_{i=1}^n \rho_{j_i}(\vec{x}_i | \vec{\theta})$  of  $X_k$  for given component with multi-index  $\vec{h}$ .



For densities representable in the form of mixtures (9), there is an important relation between their so-called score  $\vec{s}(\vec{\theta}, X_k) = \nabla_{\vec{\theta}} \ln \rho(X_k | \vec{\theta})$  and corresponding scores for a joint distributions  $\vec{s}_{\vec{h}}(\vec{\theta}, X_k) = \nabla_{\vec{\theta}} \ln \rho(X_k, \vec{h} | \vec{\theta})$ :

$$\begin{aligned} \vec{s}(\vec{\theta}, X_k) &= \frac{1}{\rho(X_k | \vec{\theta})} \nabla_{\vec{\theta}} \rho(X_k | \vec{\theta}) = \\ &= \frac{1}{\rho(X_k | \vec{\theta})} \sum_{\vec{h}} \nabla_{\vec{\theta}} \rho(X_k, \vec{h} | \vec{\theta}) = \\ \sum_{\vec{h}} \frac{\rho(X_k, \vec{h} | \vec{\theta})}{\rho(X_k | \vec{\theta})} \frac{1}{\rho(X_k, \vec{h} | \vec{\theta})} \nabla_{\vec{\theta}} \rho(X_k, \vec{h} | \vec{\theta}) &= \\ \sum_{\vec{h}} \rho(\vec{h} | X_k, \vec{\theta}) \vec{s}_{\vec{h}}(\vec{\theta}, X_k) & \end{aligned} \quad (10)$$

The importance of scores in the statistics is associated with the fact that sufficient conditions for solution  $\vec{\theta}_{ML}$  of the maximum likelihood problem (7) can be written in the form:  $\vec{s}(\vec{\theta}, X_k) = \vec{0}$ . Accordingly, the importance of the relation (10) lies in the fact that it allows one to express these conditions in an alternative form:

$$\sum_{\vec{h}} \rho(\vec{h} | X_k, \vec{\theta}) \vec{s}_{\vec{h}}(\vec{\theta}, X_k) = \vec{0}. \quad (11)$$

Insofar as

$$\vec{s}_{\vec{h}}(\vec{\theta}, X_k) = \nabla_{\vec{\theta}} \ln \rho(X_k, \vec{h} | \vec{\theta}) = \sum_{i=1}^k \nabla_{\vec{\theta}} \ln \rho(\vec{x}_i, j_i | \vec{\theta}) \quad (12)$$

and if  $\rho(\vec{x}, j | \vec{\theta})$  is more tractable than  $\rho(\vec{x} | \vec{\theta})$  (8), then the score  $\vec{s}_{\vec{h}}(\vec{\theta}, X_k)$  (12) will be more tractable than  $\vec{s}(\vec{\theta}, X_k)$  and it turns out that the solution of (11) is much easier to find than solution of  $\vec{s}(\vec{\theta}, X) = \vec{0}$ .

In addition, it is easy to see that (11) is very similar to the gradient optimization equations for the well-known *EM*-algorithm (Gupta, 2010), or its hard clustering variant, known as *K*-means segmentation. Indeed, approximate solution of (11) can be carried out by iterations containing two main steps. The first step consists in calculating the latent variables  $\vec{h}$  that maximize the posterior distribution  $\rho(\vec{h} | X_k, \vec{\theta}) = \rho(X_k, \vec{h} | \vec{\theta}) / \rho(X_k | \vec{\theta})$  for obtained in the previous iteration parameters  $\vec{\theta}$ . And the second step is to find a solution  $\vec{\theta}$  of the equation  $\vec{s}_{\vec{h}}(\vec{\theta}, X_k) = \vec{0}$  with  $\vec{h}$  found at the first step.



Figure 3: Reconstructions of “cameraman” image (1 000 000 counts) by various number of components *K* in intermediate representation: A – the original image in TIF format, B, C, D –reconstructions, corresponding to  $K = 100^2, 250^2, 300^2$  components.

Putting together all the conclusions obtained above, the final encoding  $f$  and decoding  $g$  operators of autoencoders in a generative model with latent variables  $\vec{h}$  (in a finite mixture model) can be formulated as follows:

$$\begin{aligned} \textbf{Encoding: } f: \mathcal{G} \rightarrow \mathcal{F} : \rho(X_k | \vec{\theta}) \rightarrow \vec{h}: \\ \vec{h}(X_k, \vec{\theta}) = \\ \arg \max_{j \in \{1, \dots, K\}} \left( \rho(\vec{x}_1, j | \vec{\theta}), \dots, \rho(\vec{x}_k, j | \vec{\theta}) \right). \end{aligned} \quad (13)$$

$$\begin{aligned} \textbf{Decoding: } g: \mathcal{F} \rightarrow \mathcal{G} : \vec{h} \rightarrow \rho(X_k | \vec{\theta}): \\ \vec{s}_{\vec{h}}(\vec{\theta}, X_k) = \vec{0}. \end{aligned} \quad (14)$$

The developed approach to learning generative autoencoders by image sampling representations can be naturally implemented in the form of a recurrent computational procedure. Some examples of reconstruction of sampling representation shown in Figure 1 C (1 000 000 counts) are shown in Figure 3.

## 4 CONCLUSIONS

In the framework of generative model, a new approach is proposed. It provides the synthesis of learning methods for autoencoders by images, presented as samples of random counts. The issues of simplicity of interpretation of the approach and the immediacy of its algorithmic implementation are the main content of the work. They make it attractive in both theoretical and practical terms, especially in the context of modern machine learning-oriented trends. In a sense, the proposed method is an adaptation of R. Fisher's maximum likelihood method for autoencoders, which is widely used in traditional statistics. The fruitful use of the latter has led to a huge number of important statistical results. In this regard, the author expresses the hope that the proposed approach will also be useful in solving a wide range of modern machine learning problems.

## REFERENCES

- Alain, G., Bengio, Y., et al. (2015). GSNs: Generative Stochastic Networks. arXiv:1503.05571.
- Aldrich, J. R.A. (1997). Fisher and the Making of Maximum Likelihood 1912-1922. In *Statistical Science*. V. 12(3). P. 162-176.
- Antsiperov, V.E. (2021, a). Representation of Images by the Optimal Lattice Partitions of Random Counts. In *Pat. Rec. and Image Analysis*, V. 31(3), P. 381–393.
- Antsiperov, V. (2021, b). New Maximum Similarity Method for Object Identification in Photon Counting Imaging. In *Proc. of the 10th Int. Conf. ICPRAM 2021*. V. 1: ICPRAM, P. 341-348.
- Baldi, P. (2012). Autoencoders, unsupervised learning, and deep architectures. In *Proc. of Machine Learning Research*, V. 27, PMLR, Bellevue, USA. P. 37–49.
- Barber, D. (2012). *Bayesian Reasoning and Machine Learning*. Cambridge Univ. Press, Cambridge.
- Beghdadi, A., Larabi, M.C. Bouzerdoum, A., Iftikharuddin, K.M. (2013). A survey of perceptual image processing methods. In *Signal Processing: Image Communication*, V. 28(8) P. 811-831.
- Fossum, E.R., Teranishi, N., et al. (eds.) (2017). *Photon-Counting Image Sensors*. MDPI.
- Fossum, E.R. (2020). The Invention of CMOS Image Sensors: A Camera in Every Pocket. In *Pan Pacific Microelectronics Symposium*. P. 1-6.
- Fox, M. (2006). *Quantum Optics: An Introduction*. Oxford University Press. Oxford, New York.
- Gabriel, C.G., Perrinet L., et al. (eds.) (2015). *Biologically Inspired Computer Vision: Fundamentals and Applications*. Wiley-VCH, Weinheim.
- Gallager, R. (2013). *Stochastic Processes: Theory for Applications*. Cambridge University Press. Cambridge.
- Goodfellow, I, Bengio, Y., Courville, A. (2016). Autoencoders. In *Deep Learning*, MIT Press. P. 499.
- Gupta, M. R. (2010). Theory and Use of the EM Algorithm. In *Foundations and Trends in Signal Processing*, V.1 (3). P. 223-296.
- Hecht-Nielsen, R. Replicator. (1995). Neural Networks for Universal Optimal Source Coding. In *SCIENCE*, V.269 (5232). P. 1860-1863.
- Hinton, G. E., Zemel, R. S. (1994). Autoencoders, minimum description length and Helmholtz free energy. In *Adv. in neural inform. process* V. 6. P. 3-10.
- Kingma, D.P. (2014). Welling M. Auto-Encoding Variational Bayes. arXiv:1312.6114.
- Kohonen, T. (1989). *Self-Organization and Associative Memory*, 3-d Edition. Springer.
- Morimoto, K., Ardelean, A., et al. (2020) Megapixel time-gated SPAD image sensor for 2D and 3D imaging applications. In *Optica* V. 7(4). P. 346-354.
- Murphy, K.P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.
- Pal, N.R. and Pal, S.K. (1991). Image model, Poisson distribution and object extraction. In *J. Pattern Recognit. Artif. Intell.* V. 5(3). P. 459–483.
- Plaut, E. (2018). From Principal Subspaces to Principal Components with Linear Autoencoders. arXiv: 1804.10253.
- Rodieck, R.W. (1998). *The First Steps in Seeing*. Sinauer. Sunderland, MA.
- Seitz, P., Theuwissen, A.J.P. (eds). (2011). *Single-photon imaging*. Springer. Berlin, New York.
- Streit, R.L. (2010). *Poisson Point Processes. Imaging, Tracking and Sensing*. Springer. New York.