An urn model from learning theory

Stéphane Boucheron

Laboratoire de Recherche en Informatique

CNRS URA 410 and Université Paris-Sud, 91405 Orsay (France)

Danièle Gardy

Laboratoire PRISM,

CNRS EP 0083 and Université de Versailles Saint-Quentin, 78035 Versailles (France)

Abstract

We present an urn model that appears in the study of the performance of a learning process of symmetric functions. This model is a variation on the classical occupancy model, in which the balls are of two types (good and bad). We study the cost of the learning process in the static and dynamic cases; we find gaussian limiting distributions and processes.

1 Introduction

From learning theory to urn problems

The original motivation of this investigation came from Computational Learning Theory [14]. During recent years, learning theory has paid a renewed attention to *learning curves*. Those curves monitor the improvement of the performance of the learner as she gets more information from her environment. Investigations based on Statistical Physics techniques [19] have portrayed a variety of classes of behaviors. Though those investigations were concerned with rather simple systems like perceptrons, they had to resort to advanced methods like replica calculus that still require foundational elaboration. Other analysis [12] used approximations to provide rigorous upper bounds. The results presented in those papers are some kind of laws of large numbers, they provide information on the average behavior of large systems, but they disregard the fluctuations around the average behavior. Our intention was to focus our effort on highly simplified problems and to carry out the analysis of the fluctuations around that behavior after appropriate normalization. It turns out that even for outrageously simplified learning problems this is a non trivial task.

The learning-theoretic problem we investigated is PAC-learning symmetric functions. Rather than defining PAC-learning problem in general, we will describe precisely the issues raised by the determination of the learning curves defined by symmetric functions under various conditions. Boolean symmetric functions map bitvectors from $\{0,1\}^n$ onto $\{0,1\}$ according to their Hamming weight (AND, OR, MAJORITY, PARITY are symmetric functions). The set of 2^{n+1} symmetric functions on n variables is denoted $\mathcal{G}_n$. Symmetric functions partition the elements of $\{0,1\}^n$ according to their Hamming weight in $n+1$ classes (the Hamming weight of an element of $\{0,1\}^n$ is its number of components equal to 1). $\{0,1\}^n$ is provided with a probability law D. Learning is a game between two parties : the learner $\mathcal{L}$ and an adversary $\mathcal{A}$. The adversary chooses the target symmetric function f^*. In the most primitive version, $\mathcal{L}$ draws elements according to D and asks $\mathcal{A}$ about the value of f on those elements. After k random drawings, $\mathcal{L}$ has the following couples :

$$\{(x_1, f^*(x_1)), \ldots (x_k, f^*(x_k)\} \quad ,$$

defining the sample $\mathbf{S}$. Using that sample, $\mathcal{L}$ may formulate some hypothesis on f^*.

$\mathcal{L}$ draws a new example x_{k+1}, and proposes a value $\mathcal{L}[\mathbf{S}](x_{k+1}) \in \{0,1\}$. Let us stress here on the fact that $\mathcal{L}$ does not have to propose a symmetric function $\hat{f}$ before having seen x_{k+1} but simply a possible label for x_{k+1}; $\mathcal{L}$ is allowed to toss random coins, to combine different expert advices, his unique goal is to predict as well as possible the value of $f^*(x_{k+1})$.

To assess the strategy of $\mathcal{L}$, its average performance called *generalization error* is evaluated as

$$\epsilon_g(\mathcal{L}[\mathbf{S}]) = \mathbf{E}_X \left(|\mathcal{L}[\mathbf{S}](X) - f^*(X)| \right) = \Pr_X \left\{ \mathcal{L}[\mathbf{S}](X) \neq f^*(X) \right\} \quad . \tag{1}$$

As the sample size k increases, the sequence $\epsilon_g(\mathcal{L}[\mathbf{S}^k])$ is a sequence of random variables that defines a sample path, that is a *process* indexed by sample size k. The best strategy for $\mathcal{A}$ is to choose f^* uniformly at random and the best strategy for $\mathcal{L}$ knowing $\mathbf{S}$ is to guess in the following way : if the sample contains a pair $x_i, f^*(x_i)$ such that x_i and x_{k+1} have equal weight predict $f^*(x_i)$, otherwise predict 0 or 1 with probability $1/2$. $\mathcal{L}$ will be wrong with probability $1/2$ on classes that are not represented in the sample.

The reader should notice at this point that drawing an example with specific weight is very much like throwing a ball in an urn with some specific label, the label being equal to the weight of the example. The analysis of learning symmetric functions in this simple setting is reducible to the analysis of random allocation in the classical urn model [15]. If we assume that all weights have the same probability (i.e. all urns have the same probability to receive a ball), the generalization error is governed by the number of unrepresented classes i.e. by the number of empty urns. Its expectation (over random choices of the target function and of the sample) is:

$$\mathbf{E}\epsilon_g = \frac{1}{2}\left(1 - \frac{1}{n+1}\right)^k \tag{2}$$

The variance can be computed without difficulty.

The above-described problem merely served to introduce the topic. The learning problem which actually interested us is a *noisy* variant of the preceding one. With probability $\mu < \frac{1}{2}$, $\mathcal{A}$'s answer concerning the current example is flipped (a 0 becomes a 1, and a 1 becomes a 0). In learning theory, this is called random classification noise. Labels of examples representing a given class in the sample are not any more necessarily identical, $\mathcal{L}$ may nevertheless hope that if he has many representatives of a given class a fraction $1 - \mu$ will be correctly labeled. He will thus make a probably correct inference using majority voting.

The preceding remark shows that the relation between the number of classes that are not represented in the sample and the generalization error does not work anymore. To derive expressions like (2), one has to analyze the occupancy scheme, to determine how many representatives there are in each class, and to analyze the influence of classification noise on that occupancy scheme. This has prompted us to investigate a modified urn problem.

Revisited urn occupancy models

Urn models are frequently used in discrete probability theory; see for example the book by Johnson and Kotz [13]. Among those models, so-called occupancy models have received considerable attention. In these models, balls are allocated at random amongst a sequence of urns, and the parameter under study is the number of urns satisfying some property, for example containing a given number of balls (most often, the number of empty urns). See again the book by Johnson and Kotz for a survey of results in this area, and the book by Kolchin et al. [15] for a detailed study of the asymptotic results that can be obtained.

We present a variation on this model, which allows for two types of balls, and induces different types of urns. Our urn model is as follows : We throw a specified number of balls in a set of urns; each ball is thrown independently of the others (for example, there is no limit on the number of balls that an urn can receive); each urn has the same probability to receive a given ball. The balls are of two types ("good" and "bad"), which gives three possibilities for an urn : either the "good" balls predominate, or the "bad" balls, or there is an equal number of each type (the urn may be empty). Each type of urn has a specific cost; the global cost is assumed to be a linear function of the numbers of urns of each type. We want to study this cost as a function of the number of balls k and the number of urns n.

In the classical model, when the number of urns n and the number of balls k are proportional (the central domain in [15]), the distribution of the number of empty urns follows asymptotically a normal distribution after adequate normalization and centering [13]. The mean and variance of the number of urns with a fixed number of balls are proportional to n [13]. There are also results on the stochastic process, when the balls are thrown one at a time (see [15, Ch. IV]) : the normalized and centered process is asymptotically gaussian and markovian [15], it can be regarded as a rescaled Brownian Motion [1].

Some urn occupancy models with two types of balls have already been proposed in the

literature. For example, Nishimura and Sibuya [16] and Selivanov [18] study the waiting time until at least an urn contains balls of two types; this is an extension of the birthday problem. Closer to our problem, Popova investigated in [17] the distribution of the vector whose components are number of urns with balls of the first type only, the second type only, and without any balls, and showed that the asymptotic distribution is either Poisson or normal, according to the relative values of the numbers of balls of each type and the number of urns.

The classical model as well as the extensions with two types of balls prompt us to search results of a similar flavor : mean value, variance and limiting distribution when the number of urns n, the number of balls k, and the relationship between them are known, and when $n, k \to +\infty$; when the balls are added sequentially, limiting process for a large time. However, the noisy urn model is significantly more complicated than the empty urn problem. For finite n, the process is no more Markovian. Hence the proof technique used in [1] breaks down. Though Markov property was not explicitly invoked in [15], the markovian character of the empty urn process is responsible for the relative tractability of the generating functions manipulated in [15].

We shall show that we can indeed obtain a full asymptotic description of the phenomenon when k and n are proportional.

The plan of our paper is the following : in the next section, we recall the results pertaining to the learning of symmetric functions without noise. Then we define precisely our model and show how we can associate generating functions to the parameters of interest in Section 3. We study the *static* case (the number of balls is fixed) in Section 4 and the *dynamic* model (the balls are thrown at regular intervals) in Section 5. We show that, when the number of balls is of the same order as the number of urns, the cost function behaves asymptotically as a gaussian distribution (in the static case) or as a gaussian, non-Markov process (in the dynamic case). The conclusion gives suggestions for possible extensions of our work; we give here some indications as to how the phenomenon of majority (classification of an urn in one of three types, according to the behavior of the majority of balls) can be extended to so-called decomposable structures. Finally, an appendix (Section 8) summarizes some mathematical notions that we need during our analysis.

2 The classical problem : background on symmetric functions

To get a better insight on the learning process and identify the fundamental trends and the role of random perturbations due to random sampling, we will carry out some asymptotic analysis, letting n go to infinity, while setting a scaling law between sample size k and problem dimension n; $\alpha \stackrel{\triangle}{=} \frac{k}{n}$. This turns out to analyze learning on a time scale that is proportional to problem dimension.

For each system size, one defines the RCLL process indexed by $\mathbf{R}^+$:

$$\tilde{\epsilon}_g^n(\alpha) \stackrel{\triangle}{=} \epsilon_{\mathbf{g}}(\lfloor \alpha.\mathbf{n} \rfloor). \tag{3}$$

An involved analysis [15, 1], shows that the limiting law of sample paths tends to be concentrated around the average curve

$$\epsilon_{\mathbf{g}}(\alpha) = \frac{\mathbf{1}}{\mathbf{2}} \mathrm{e}^{-\alpha}. \tag{4}$$

The limiting *learning curve* is thus an exponential; this exemplifies a common pattern when the class of target functions is finite. It represents the fundamental trend of the learning process. During an experiment, the actual curve will be a perturbation of this mean curve, the perturbation is due to randomness and finiteness.

Thus to get a correct interpretation, it is desirable to analyze the way disorder vanishes as system size increases. There are two possible views on fluctuations. The central limit (resp. large deviation) viewpoint searches tight results concerning the fluctuations of magnitude $\Theta(\sqrt{n})$ (resp. $\Theta(n)$).

The mean deviation approach was explored by Renyi and exposed in Kolchin et al. [15] using complex analysis techniques or in Barraez [1] using diffusion approximation techniques. This analysis shows that the limiting normalized centered process is Gaussian with covariance :

$$\mathrm{cov}(s,t) = \exp^{-t}\left(1 - s\exp^{-s}\right) \quad .(s \leq t) \tag{5}$$

This provides the characterization of the convergence of learning curves towards a limiting trajectory. It shows that the latter is not only an average trajectory but also a *typical* trajectory.

3 The new urn model and its generating functions

We shall make heavy use of the technic of generating functions for combinatorial enumeration to analyze the model in the static case; see for example [11] for a general presentation and [6] for a basic presentation applied to the analysis of algorithms. We shall give in this section several generating functions, each describing the problem from a slightly different point of view. We give below a brief summary of the basic facts that we shall use; see [6] for a more formal presentation.

- Our generating functions are exponential in the variable (y) marking the number of balls; this comes from the fact that the balls are indistinguishable : the result depends on the final configuration of balls, not on the order in which they were allocated (this is no longer true when we consider the dynamic case).
- Let $f(\mathbf{x})$ be the generating function relative to a set of variables $\mathbf{x}$ marking some parameters for one urn; the generating function describing the sequence of n distinguishable urns for the same set of parameters is $f(\mathbf{x})^n$.
- As a consequence, the generating function describing the allocation of balls in a single

urn is e^y : There is only one way to allocate k identical balls into one urn. The function describing the allocation of balls into the n urns is e^{ny}.
- If the situation in an urn can be partitioned in two situations, with associated generating functions respectively $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$, then the function describing the complete situation is $f_1(\mathbf{x}) + f_2(\mathbf{x})$.

3.1 What happens in an urn?

We consider a finite sequence of n distinguishable urns, in which we throw balls independently. The number of balls that an urn can receive is not bounded, and the balls are assumed to be indistinguishable. The exponential generating function describing the allocation of balls in an urn is then simply e^y, with the variable y marking the number of balls.

Now, assume that the balls are of two types ("good" and "bad"), with respective probabilities $1-\mu$ (good balls) and μ (bad balls) $(0 < \mu < 1/2)$. The generating function describing the allocation of balls in a given urn can be written as $e^y = e^{\mu y + (1-\mu)y}$.

The two types of balls lead to three possibilities for the urns : An urn is "good" if there are more good balls than bad balls; it is "bad" if there is a majority of bad balls, and "neutral" if there is an equal number of good and bad balls (possibly none, this case includes empty urns).

This translates into generating functions as follows : "In a good urn, the exponent of $(1-\mu)y$ is greater than the exponent of μy". To capture this idea, we shall introduce a new variable z, and substitute $(1-\mu)yz$ for $(1-\mu)y$ and $\mu y z^{-1}$ for μy : We get a function of y and z, which we write as a series on z :

$$e^{y((1-\mu)z+\mu/z)} = \sum_{p\in Z} a_p(y) z^p.$$

The *positive* powers of z indicate a good urn; the *negative* powers indicate a bad urn, and the *constant* term (without z) a neutral urn.

The next step is to get an expression of the term $a_p(y)$. We shall use the relation

$$e^{y(x+1/x)} = \sum_{p\in Z} I_p(2y)\, x^p,$$

with $I_p(y)$ denoting a (modified) Bessel function : $I_p(y) = \sum_r (y/2)^{p+2r}/r!(r+p)!$ (see the Appendix, Section 7.3, for a few properties of Bessel functions). Define $\sigma^2 := \mu(1-\mu)$; we have that

$$exp\left(y\left[(1-\mu)z + \frac{\mu}{z}\right]\right) = exp\left(\sigma y\left[\frac{\sigma\, z}{\mu} + \frac{\mu}{\sigma\, z}\right]\right) = \sum_{p\in Z} I_p(2\sigma\, y)\left(\frac{\sigma}{\mu}\right)^p z^p;$$

hence $a_p(y) = (\sigma/\mu)^p I_p(2\sigma\, y)$ and

$$e^{y[(1-\mu)z+\mu/z]} = \sum_{p\in Z} I_p(2\sigma\, y)\left(\frac{\sigma}{\mu}\right)^p z^p. \tag{6}$$

For $z = 1$, we get $e^y == \sum_{p\in Z} I_p(2\sigma y)(\sigma/\mu)^p$, which we shall use to simplify some computations in the sequel.

3.2 Balance of an urn

The *balance* of an urn is defined as the difference between the number of good balls and the number of bad balls; this is the exponent of the variable z in Equation (6). As the global cost is linear function of the costs of individual urns, the i^{th} moment of the global cost can be determined thanks to the joint law of the cost of i fixed urns. Hence, it is relevant to determine those joint laws. Moreover it is easy; a straightforward derivation or an approach by generating functions are equally successful.

Proposition 3.1 *The vector of balances in urns $1..i$ after throwing k balls in n urns is distributed according to a law Q_i that is with variational distance $2i/n$ from the law P_i of a vector of i independent random variables that are distributed as the difference between two independent Poisson random variables with means $\mu\, k/n$ and $(1-\mu)\, k/n$.*

Recall that the difference of two Poisson random variables with mean $y\mu$ and $y(1-\mu)$ has the same law as the random variable defined by first drawing an integer ℓ according to the Poisson law of mean y, then drawing ℓ i.i.d. $\{-1, 1\}$-valued random variable with mean $1 - 2\mu$.

Recall also that conditionally on the number of balls allocated in urns $1 \ldots i$, the balances of the i urns are independent. In the language of [4], this means that the σ-field generated by the number of balls allocated to the first i urns is sufficient to compute the variation distance between P_i and Q_i ([4, lemma 2.4]). Hence the variation distance between P_i and Q_i is equal to the variation distance between the law of the vector of balls allocated to urns $1 \ldots i$ and a vector of i independent Poisson random variables with mean k/n. By theorem (5.1) in [4] which relies on previous results by Kersten, this is smaller than $2i/n$. Notice that the latter quantity does not depend on k and μ.

For large n and k, the law of the balance is asymptotically equal to

$$\Pr(\text{balance } = p \text{ /a total of} k \text{ balls}) \sim e^{-\alpha} \left(\frac{\sigma}{\mu}\right)^p I_p(2\sigma\,\alpha).$$

It is also possible to derive this expression from the generating function marking the balance of an urn by z, and "forgetting" the state of the other urns : this function is equal to $f_1(y,z)\; e^{(n-1)y}$, with $f_1(y,z) := \sum_{p\in Z}(\sigma/\mu)^p I_p(2\sigma\, y) z^p$ describing what happens in the urn under inspection, and e^y describing each of the $n-1$ other urns. The desired probability is then

$$\frac{[y^k z^p]\{f_1(y,z)\; e^{(n-1)y}\}}{[y^k]\{f_1(y,1)\; e^{(n-1)y}\}} = \frac{k!}{n^k} \left(\frac{\sigma}{\mu}\right)^p [y^k]\{e^{(n-1)y} I_p(2\sigma\, y)\}.$$

This last coefficient can be estimated, for $n \to +\infty$ and $k = \alpha\, n$, as (see Section 7.1)

$$\frac{e^{(n-1)\alpha} I_p(2\sigma\,\alpha)}{\alpha^{\alpha n}\sqrt{2\pi\alpha n}}$$

The speed of convergence is of order $1/n$; this can be proved, either by a result of Diaconis and Friedman [4], or by the generating function approach : Each individual probability is $e^{-\alpha}(\sigma/\mu)^p I_p(2\sigma\,\alpha)(1+O(1/n))$ (the saddle-point approximation can be improved to give a full asymptotic expansion, in the vein of Good's extension of Daniels's result [10]), and summing these probabilities gives the result.

The same approach gives the joint distribution of the balance in a fixed number r of urns : The probability that the first urn has a balance p_1, the second urn a balance p_2, ..., the r^{th} urn a balance p_r, when there is a total of $k=\alpha n$ balls, is

$$e^{-r\,\alpha}\left(\frac{\sigma}{\mu}\right)^{p_1+\ldots+p_r} I_{p_1}(2\sigma\,\alpha)\ldots I_{p_r}(2\sigma\,\alpha).$$

3.3 The different states of an urn

Now we give the generating function describing the system of n urns, when we are interested no more in the balance of the urns, but simply in their states (good, bad or neutral). Define

$$\begin{aligned}\phi(y) &:= \sum_{p<0}\left(\frac{\sigma}{\mu}\right)^p I_p(2\sigma y) = \sum_{p>0}\left(\frac{\mu}{\sigma}\right)^p I_p(2\sigma y);\\ \psi(y) &:= \sum_{p>0}\left(\frac{\sigma}{\mu}\right)^p I_p(2\sigma y) = e^y - I_0(2\,\sigma y) - \phi(y).\end{aligned}$$

Now we introduce a new variable for each possible state of an urn : u is associated to a bad urn, v to a neutral urn, and w to a good urn. We obtain the generating function describing what happens in an urn, with y indicating the number of balls in the urn, by substituting u to z^p for $p<0$, v to z^0 and w to z^p for $p>0$:

$$u\sum_{p<0} a_p(y) + v\,a_0(y) + w\sum_{p>0} a_p(y).$$

The generating function describing the possible states of an urn is

$$f(u,v,w,y) := u\,\phi(y) + v\,I_0(2\,\sigma y) + w\,\psi(y).$$

and the generating function describing the behavior of the system of n urns is (recall that the generating function for a sequence of n urns is the n^{th} power of the function for one urn)

$$F(u,v,w,y) = f(u,v,w,y)^n = \left(u\,\phi(y) + v\,I_0(2\,\sigma y) + w\,\psi(y)\right)^n. \tag{7}$$

3.4 The cost function

Now that we have described the basic model of allocation of balls in urns, and that we have obtained in (7) the generating function marking the different possibilities for the n

urns, the next step is to associate a cost with each urn : We assume that the cost of an experiment (throwing k balls into n urns) is a linear function of the numbers of urns of each type:

$$Cost = C_0 \cdot Number\ of\ neutral\ urns + C_1 \cdot Number\ of\ bad\ urns + C_2 \cdot Number\ of\ good\ urns.$$

We shall use the following notations :

- The cost of a neutral or empty urn is C_0;
- The cost of a bad urn is C_1;
- The cost of a good urn is C_2.

Examples of costs relevant to learning theoretic applications are $C_0 = 1$, $C_1 = 2$ and $C_2 = 0$, or $C_0 = 1/2$, $C_1 = 1 - \mu$ and $C_2 = \mu$ (we usually have $C_2 \leq C_0 \leq C_1$).

The cost of a realization of the scheme of allocations is defined as the sum of the costs of each urn. To obtain its generating function, we start from the function $F(u, v, w, y)$ given by Equation(7), describing the state of the system of urns, and introduce the variable x marking the global cost as follows : We substitute x^{C_0} for v, x^{C_1} for u and x^{C_2} for w. We get

$$G(x, y) := g^n(x, y) := \left(x^{C_0} I_0(2\sigma\ y) + x^{C_1} \phi(y) + x^{C_2} \psi(y) \right)^n. \tag{8}$$

Up to a normalization, the coefficient $[x^p y^k]G(x, y)$ is equal to the probability that, after throwing k balls, the global cost is equal to p :

$$\text{Proba}(cost = p/k\ balls) = \frac{[x^p y^k]G(x, y)}{[y^k]G(1, y)} = \frac{k!}{n^k}[x^p y^k]G(x, y).$$

In the sequel, we shall assume that p is an integer, which allows us to use Cauchy's formula and complex integration technics to get approximations of generating function coefficients.

4 The static case

4.1 The average cost

The average cost is equal to $[y^k]G'_x(1, y)/[y^k]G(1, y)$. As $G'_x = n g'_x g^{n-1}$, we have that

$$g'_x(x, y) = C_0\ x^{C_0 - 1} I_0(2\sigma\ y) + C_1\ x^{C_1 - 1} \phi(y) + C_2\ x^{C_2 - 1} \psi(y).$$

Hence, with $g(1, y) = e^y$,

$$\begin{aligned} G'_x(1, y) &= n e^{(n-1)y} (C_0\ I_0(2\sigma\ y) + C_1\ \phi(y) + C_2\ (e^y - I_0(2\sigma\ y) - \phi(y)) \\ &= n[C_2\ e^{ny} + (C_0 - C_2) e^{(n-1)y} I_0(2\sigma\ y) + (C_1 - C_2) e^{(n-1)y} \phi(y)]. \end{aligned}$$

As $[y^k]G(1,y) = [y^k]e^{ny} = n^k/k!$, we obtain the average cost as

$$\frac{k!}{n^k}\, n\, \left(C_2 \frac{n^k}{k!} + (C_0 - C_2)[y^k]\{e^{(n-1)y} I_0(2\sigma y)\} + (C_1 - C_2)[y^k]\{e^{(n-1)y}\phi(y)\} \right).$$

To obtain an asymptotic value for $k = \alpha\, n$, we apply the saddle-point method to the evaluation of the coefficients $[y^k]\{e^{(n-1)y} I_0(2\sigma\, y)\}$ and $[y^k]\{e^{(n-1)y}\phi(y)\}$ (see again the Appendix; Section 7.1); we get

$$Average\ \ cost \sim n \left[C_2 \ + (C_0 - C_2)e^{-\alpha} I_0(2\sigma\ \alpha) + (C_1 - C_2)e^{-\alpha}\phi(\alpha) \right].$$

4.2 Variance and limiting distribution

When the number of urns and the number of balls are proportional, the distribution of the cost is asymptotically normal. There are several ways to prove it; for example we can apply results by Bender and Richmond [2]; however, in order to use these results, we have first to prove that the variance is of exact order n[1]. We shall use here another theorem, given in [7, p. 278], that gives us directly the asymptotic value of the variance together with the asymptotic normality; we recall it below.

Theorem 4.1 *Let $g(x,y) = \sum_{n,k} a_{n,k} x^k y^n$ be a function with positive coefficients $a_{n,k}$ and entire w.r.t. y. Assume that the parameters k and n grow to infinity in such a way that $k/n \to \alpha > 0$, and that α satisfies*

$$\lim_{y \to +\infty} y g_y'(1,y)/g(1,y) > \alpha.$$

Let ρ be the unique real positive solution of the equation $y g_y'(1,y)/g(1,y) = \alpha$. Define $\kappa(x,y) = x g_x'(x,y)/g(x,y)$ and $\lambda(x,y) = y g_y'(x,y)/g(x,y)$, and assume that, at $(1,\rho)$, $\kappa_x'\lambda_y' - \kappa_y'\lambda_x' \neq 0$. Then the function $f(x) = [y^k]\{g(x,y)^n\}/[y^k]\{g(1,y)^n\}$ is the generating function of a probability distribution that is asymptotically normal; its asymptotic mean and variance are

$$\mu = n\ \kappa(1,\rho); \quad \sigma^2 = n\ \frac{\kappa_x'\lambda_y' - \kappa_y'\lambda_x'}{\lambda_y'}(1,\rho).$$

In our problem, $\kappa(1,y) = e^{-y}((C_0 - C_2)I_0(2\sigma\, y) + (C_1 - C_2)\phi(y) + C_2 e^y)$ and we get back the expression of the average cost when $k = n\alpha$.

Computation of the asymptotic variance

Applying Theorem 3.1, we get that the asymptotic variance is $n\sigma_0^2$, with ($\rho = \alpha$ and $\lambda'(1,\alpha) = 1$)

$$\sigma_0^2 = e^{-\alpha} \left[g_x' + g_{x^2}'' - e^{-\alpha} \left((g_x')^2 - \alpha(g_{xy}'' - g_x') \right) \right],$$

[1]Proposition (3.1) immediately reveals that the variance of the normalized cost process should not exceed $O(1)$; thus the variance of the global process should be $O(n)$. It is straightforward to check that the normalized variance is the sum of one term corresponding to the variance of the cost of urn 1 -which is obviously $O(1)$- and of one term corresponding to the covariance of the cost of urns 1 and 2, multiplied by $n-1$. By proposition (3.1), the covariance should be $O(1/n)$.

where the derivatives of g are at $(1, \alpha)$. After some computations, and with the notations

$$\begin{aligned} d_0 &:= C_1 - C_0 &= Cost(bad\ urn) - Cost(neutral\ urn); \\ d_1 &:= C_0 - C_2 &= Cost(neutral\ urn) - Cost(good\ urn), \end{aligned}$$

we get

$$\begin{aligned} \sigma_0^2 &= (d_0+d_1)^2 e^{-\alpha}\phi(\alpha) + d_1^2 e^{-\alpha} I_0(2\sigma\,\alpha) \\ &\quad - \left((d_0+d_1)e^{-\alpha}\phi(\alpha) + d_1 e^{-\alpha} I_0(2\sigma\,\alpha)\right)^2 \\ &\quad + \alpha\left((\mu(d_0+d_1) - d_1)e^{-\alpha} I_0(2\sigma\,\alpha) - \sigma\,(d_0-d_1)e^{-\alpha} I_1(2\sigma\,\alpha)\right)^2. \end{aligned}$$

We sum up our results so far in the following theorem :

Theorem 4.2 *Assume that the number n of urns and the number k of balls grow to infinity in such a way that $k = \alpha n$ for some constant α. The cost (generalization error) is asymptotically normally distributed, with asymptotic mean and variance*

$$\begin{aligned} E &\sim n\left[C_2 + d_1 e^{-\alpha} I_0(2\sigma\,\alpha) + (d_0+d_1)e^{-\alpha}\phi(\alpha)\right]; \\ \sigma^2 &\sim n\left[(d_0+d_1)^2 e^{-\alpha}\phi(\alpha) + d_1^2 e^{-\alpha} I_0(2\sigma\,\alpha)\right. \\ &\quad - \left((d_0+d_1)e^{-\alpha}\phi(\alpha) + d_1 e^{-\alpha} I_0(2\sigma\,\alpha)\right)^2 \\ &\quad \left. + \alpha\left((\mu(d_0+d_1) - d_1)e^{-\alpha} I_0(2\sigma\,\alpha) - \sigma\,(d_0-d_1)e^{-\alpha} I_1(2\sigma\,\alpha)\right)^2\right]. \end{aligned}$$

4.3 The influence of the ratio $\alpha = k/n$

For our two examples of cost, we get

- When $C_0 = 1$, $C_1 = 2$ and $C_2 = 0$,

  $$E[Cost_1] \sim n e^{-\alpha}(I_0(2\sigma\,\alpha) + 2\phi(\alpha))$$

  and

  $$\sigma^2(Cost_1) = n\left[e^{-\alpha}\left(I_0(2\sigma\,\alpha) + 4\phi(\alpha)\right) + e^{-2\alpha}\left(\alpha(1-2\mu)^2 - (I_0(2\sigma\,\alpha) + 2\phi(\alpha))^2\right)\right].$$

- When $C_0 = 1/2$, $C_1 = 1-\mu$ and $C_2 = \mu$,

  $$E[Cost_2] \sim n(\mu + \frac{1-2\mu}{2} e^{-\alpha}(I_0(2\sigma\,\alpha) + 2\phi(\alpha)))$$

  and

  $$\begin{aligned} \sigma^2(Cost_2) = \; & n\left(\tfrac{1-2\mu}{2}\right)^2 [\; e^{-\alpha}\left(I_0(2\sigma\,\alpha) + 4\phi(\alpha)\right) \\ & + e^{-2\alpha}\left(\alpha(1-2\mu)^2 - (I_0(2\sigma\,\alpha) + 2\phi(\alpha))^2\right)]. \end{aligned}$$

Note that $E[Cost_2] = n\mu + ((1-2\mu)/2)\,E[Cost_1]$; the variances are similarly related : $\sigma^2(Cost_2) = (1/2-\mu)^2\,\sigma^2(Cost_1)$.

In the two examples of costs we considered, the asymptotic average cost is of the form $nh(\alpha)$ and the function h decreases when α grows. This is often satisfied : assume that $d_0 \geq d_1 \geq 0$ (which is often the case in our learning problem). The average cost is asymptotically equal to $nh(\alpha)$ with

$$h(\alpha) := C_2 \; + d_1 e^{-\alpha} I_0(2\sigma\,\alpha) + (d_0+d_1)e^{-\alpha}\phi(\alpha).$$

To study the variations of $h(\alpha)$, we compute its derivative $h'(\alpha)$; we use the fact that $\phi'(y) = \phi(y) + \mu I_0(2\sigma\,y) - \sigma I_1(2\sigma\,y)$ (see the appendix, eq. 25) to get rid of the derivative of ϕ; we get

$$h'(\alpha) = -\sigma e^{-\alpha}[(d_0-d_1)I_1(2\sigma\alpha) + (d_1\sigma/\mu - d_0\mu/\sigma)I_0(2\sigma\alpha)].$$

Hence, for $d_1 \leq d_0 \leq d_1(1-\mu)/\mu$, $h'(\alpha) < 0$ and $h(\alpha)$ is a decreasing function of α. Now, for $d_0 > d_1(1-\mu)/\mu (> d_1)$, h is first increasing, then decreasing; it has a maximum for α such that

$$\frac{I_1(2\sigma\alpha)}{I_0(2\sigma\alpha)} = \frac{\mu(d_0+d_1)-d_1}{\sigma(d_0-d_1)}.$$

The function $t \mapsto I_1(t)/I_0(t)$ is increasing on $[0,+\infty[$; hence the uniqueness of the solution.

The variance as a function of α:
We now consider the asymptotic variance as a function of α; we use the asymptotic equivalents

$$I_p(2\sigma\,\alpha) = \frac{e^{2\sigma\,\alpha}}{\sqrt{4\pi\sigma\,\alpha}}(1+O(1/t))$$

$(p = 0, 1)$ and (see Sections 7.2 and 7.3 of the Appendix)

$$\phi(\alpha) = \frac{\mu}{\sigma-\mu}\,\frac{e^{2\sigma\,\alpha}}{\sqrt{4\pi\sigma\,\alpha}}(1+O(1/t));$$

as $\sigma \leq 1/2$, $(e^{(1-2\sigma)\alpha}/\sqrt{4\pi\sigma\,\alpha})^2 = o(e^{(1-2\sigma)\alpha}/(\alpha\sqrt{4\pi\sigma\,\alpha}))$ and we get

$$\sigma_0^2 = \left(d_1^2 + \frac{\mu}{\sigma-\mu}(d_0+d_1)^2\right)\frac{e^{-(1-2\sigma)\alpha}}{\sqrt{4\pi\sigma\,\alpha}}(1+O(1/\alpha)). \tag{9}$$

The equation (9) shows that the asymptotic variance (for large n and k) is exponentially decreasing in α (modulated by $1/\sqrt{\alpha}$).

5 The dynamic case

Our aim in this part is to prove that the process describing the cost can be precisely characterized. The first step is to introduce a notion of time : We consider a discrete

time, and add a ball at each moment. The number of balls k is then equal to the time t; we still assume that the number of balls is proportional to the number of urns : $k = t = \alpha n$. The number of bad balls follows a binomial distribution of parameter μ; this distribution is asymptotically normal, with mean $k\mu = \alpha n t$, variance $k\mu(1-\mu) = \alpha\sigma^2 n$ and covariance $n(t_2 - t_1)\mu(1-\mu)$.

The cost is now a function of the time, i.e. a stochastic process : $Cost(t)$. We shall first compute the covariance at different times : $Cov(Cost(t_1), Cost(t_2))$. It is possible to derive such an expression by elementary means (see Section 5.1); we shall also present (in Section 5.3) a computation using the generating function describing what happens in the urns at two different times t_1 and t_2. This generating function is then used to prove the gaussian behavior of the asymptotic bivariate distribution (Section 5.4). Such an approach is inspired from what was done for the classical occupancy model [15, Ch. IV]; it is probably easier to generalize to higher dimensions that the direct approach.

5.1 The covariance

The fact that the balance in a fixed numbers of urns lends itself to an easy analysis via the Poisson approximation (cf proposition 3.1) provides a safe way to determine the asymptotic covariance of the cost process.

Let us adopt the following conventions: $Z_i(\alpha)$ = number of balls in urn i at time αn , and in the sequel $\alpha_1 \leq \alpha_2$, $N(i,\alpha) \stackrel{\Delta}{=} 1$ (resp. $M(i,\alpha) \stackrel{\Delta}{=} 1$) if urn i is neutral (resp. bad) at time $\alpha\, n$, and 0 otherwise.

The global cost $Cost$ at time αn is:

$$C_2 + \sum_{i=1}^{n} \left((C_0 - C_2)\, N(i,\alpha) + (C_1 - C_2)\, M_i(\alpha) \right) \quad . \tag{10}$$

If the centered cost process is normalized by $1/\sqrt{n}$, then using the exchangeability of the M_i and N_i, the normalized covariance can be written down as:

$$\begin{aligned}
&(C_0 - C_2)^2 \left[\mathbf{E}N_1(\alpha_1)\, N_1(\alpha_2) - \mathbf{E}N_1(\alpha_1)\mathbf{E}N_1(\alpha_2)\right] \\
&+(C_1 - C_2)^2 \left[\mathbf{E}M_1(\alpha_1)\, M_1(\alpha_2) - \mathbf{E}M_1(\alpha_1)\mathbf{E}M_1(\alpha_2)\right] \\
&+(C_0 - C_2)(C_1 - C_2)\Big[\mathbf{E}M_1(\alpha_1)\, N_1(\alpha_2) - \mathbf{E}M_1(\alpha_1)\mathbf{E}N_1(\alpha_2) \\
&\qquad\qquad + \mathbf{E}N_1(\alpha_1)\, M_1(\alpha_2) - \mathbf{E}N_1(\alpha_1)\mathbf{E}M_1(\alpha_2)\Big] \\
&+(n-1)\Big[(C_0 - C_2)^2 \left[\mathbf{E}N_1(\alpha_1)\, N_2(\alpha_2) - \mathbf{E}N_1(\alpha_1)\mathbf{E}N_2(\alpha_2)\right] \\
&\qquad + (C_1 - C_2)^2 \left[\mathbf{E}M_1(\alpha_1)\, M_2(\alpha_2) - \mathbf{E}M_1(\alpha_1)\mathbf{E}M_2(\alpha_2)\right)] \\
&\qquad + (C_0 - C_2)(C_1 - C_2)[\mathbf{E}M_1(\alpha_1)\, N_2(\alpha_2) - \mathbf{E}M_1(\alpha_1)\mathbf{E}N_2(\alpha_2) \\
&\qquad\qquad + \mathbf{E}N(1,\alpha_1)M(2,\alpha_2) - \mathbf{E}N(1,\alpha_1)\,\mathbf{E}M(2,\alpha_2)]\Big]
\end{aligned} \tag{11}$$

The various expectations involved in the preceding sum can be identified with probabilities of events occuring at different instants in possibly different urns. The computation of the joint probabilities is facilitated by noticing that if we condition on the number of balls that have been thrown in two distinct urns, the fates of those two urns are independent. One should also notice that $\mathbf{E}\left(N(1,\alpha_1)|Z_1\right)$ and $\mathbf{E}\left(M(1,\alpha_1)|Z_1\right)$ do not depend on n.

Let us see how to compute the difference between the joint probabilities and the product probabilities of two events concerning two different urns (say urn 1 and 2)).

$$\begin{aligned}
&\mathbf{E}N(1,\alpha_1)M(2,\alpha_2) - \mathbf{E}N(1,\alpha_1)\,\mathbf{E}M(2,\alpha_2)\\
&\quad = \textstyle\sum_{k_1,k_2} \mathbf{E}\left(N(1,\alpha_1)|Z_1\right)\mathbf{E}\left(M(2,\alpha_2)|Z_2\right)\\
&\qquad\qquad \left(\Pr\left\{Z_1(\alpha_1)=k_1 \wedge Z_2(\alpha_2)=k_2\right\} - \Pr\left\{Z_1(\alpha_1)=k_1\right\}\Pr\left\{Z_2(\alpha_2)=k_2\right\}\right)
\end{aligned} \tag{12}$$

The fact that the variational distance between the occupancy law of two different urns and the joint law of two independent Poisson random variables is $O(1/n)$ already warrants that the second summand in (11) $(+(n-1)[\ldots]$ is $O(1)$.

As all terms involving two distinct urns have this form, it is highly beneficial to compute the first two terms in the development of

$$I(k_1,k_2) \triangleq \left(\Pr\left\{Z_1(\alpha_1)=k_1 \wedge Z_2(\alpha_2)=k_2\right\} - \Pr\left\{Z_1(\alpha_1)=k_1\right\}\Pr\left\{Z_2(\alpha_2)=k_2\right\}\right)$$

when $n \to \infty$.

$$\begin{aligned}
I(k_1,k_2) &= \textstyle\sum_{h\le k_2} \binom{\alpha_1 n}{k_1+h}\binom{k_1+h}{h}\binom{(\alpha_2-\alpha_1)n}{k_2-h}(1-\frac{2}{n})^{\alpha_1 n-(k_1+h)}(1-\frac{1}{n})^{\alpha_2 n-(k_2-h)}\frac{1}{n^{k_1+k_2}}\\
&\qquad - \binom{\alpha_1 n}{k_1}\binom{(\alpha_2 n}{k_2}\frac{1}{n^{k_1+k_2}}(1-\frac{1}{n})^{\alpha_1 n+\alpha_2 n-(k_1+k_2)}\\
&\qquad \text{using developments (29) in section 7.5}\\
&= \mathrm{e}^{-\alpha_1-\alpha_2}\ \frac{\alpha_2^{k_2}\alpha_1^{k_1}}{k_2!\,k_1!}\ \left(\frac{(k_1-\alpha_1)(\alpha_2-k_2)}{\alpha_2 n}+o(1/n)\right)
\end{aligned} \tag{13}$$

The conditional probabilities are:

$$\begin{aligned}
\mathbf{E}\left(N(1,\alpha_1)\middle|Z_1(\alpha_1)=2k_1\right) &= \binom{2k_1}{k_1}\mu^{k_1}(-1\mu)^{k_1}\\
\mathbf{E}\left(M(1,\alpha_1)\middle|Z_1(\alpha_1)=k_1\right) &= \sum_{h\ :\ k_1<2h\le 2k_1}\binom{k_1}{h}\mu^h(1-\mu)^{k_1-h}
\end{aligned} \tag{14}$$

Then straightforward computations using identities (28) in section (7.5) lead to the determination of the second summand in (11). The computation of the joint probabilities in the first summand has to be carried out on a case by case basis. Only the first term in the development is required. Let us illustrate the problem on two situations computation of $\mathbf{E}(N(1,\alpha_1)\,N(1,\alpha_2)$ and $\mathbf{E}(M(1,\alpha_1)\,N(1,\alpha_2)$.

In the first case, there is no need to compute anything, since if $N(1,\alpha_1)=1$, the probability that $N(1,\alpha_2)=1$ is just the probability that there are as many good as bad

examples from class 1 among the last $(\alpha_2 - \alpha_1)n$ examples. The latter quantity turns to be $\Pr\{N(1, \alpha_2 - \alpha_1)\}$. Hence

$$\begin{aligned}
&\mathbf{E}\Big(N(1,\alpha_1)\,N(1,\alpha_2)\Big) - \mathbf{E}\Big(N(1,\alpha_1)\Big\}\,\mathbf{E}\Big(N(1,\alpha_2)\Big) \\
&\quad \to_{n\to\infty} \quad \mathrm{e}^{-\alpha_1-\alpha_2}\,\mathrm{I}_0(2\sigma\alpha_1)\Big(\mathrm{e}^{\alpha_1}\mathrm{I}_0(2\sigma(\alpha_2-\alpha_1)) - \mathrm{I}_0(2\sigma\alpha_2)\Big) \qquad (15)
\end{aligned}$$

In the second case, thanks to a reflection trick, there is no need to compute anything as well. We will first show that:

$$\mathbf{E}\Big(M(1,\alpha_1)\,N(1,\alpha_2)\Big) = \Pr\Big\{\text{urn 1 is good at } \alpha_1 \wedge N(1,\alpha_2) = 1\Big\}. \qquad (16)$$

Let us assume that on some sample path $\mathbf{S}$ of the random allocation process, $N(1,\alpha_2) = 1$ is realized. Now consider the sample $\mathbf{S}$ where the labels of balls allocated to urn 1 in $\mathbf{S}$ have been flipped. Since $N(1,\alpha_2)[\mathbf{S}] = 1$, $N(1,\alpha_2)[\mathbf{S}] = 1$, and moreover the two samples have the same probability, and urn 1 is good at time α_1 on $\mathbf{S}$ iff urn 2 is bad at time α_2 on $\mathbf{S}$. Hence when $N(1,\alpha_2)$ is realized, there is a probability-preserving 1-1 correspondence between the samples that are good at time α_1 and those that are bad at time α_1 Finally:

$$\begin{aligned}
&\mathbf{E}\Big(M(1,\alpha_1)\,N(1,\alpha_2)\Big) - \mathbf{E}\Big(M(1,\alpha_1)\Big)\,\mathbf{E}\Big(N(1,\alpha_2)\Big) \\
&= \quad \mathrm{e}^{-\alpha_1-\alpha_2}\Big(\frac{\mathrm{e}^{\alpha_1}}{2}(\mathrm{I}_0(2\sigma\alpha_2) - \mathrm{I}_0(2\sigma\alpha_1)\mathrm{I}_0(2\sigma(\alpha_2-\alpha_1))) - \phi(\alpha_1)\mathrm{I}_0(2\sigma\alpha_2)\Big) \qquad (17)
\end{aligned}$$

If we adopt the choice for $C_0 = 1/2$, $C_1 = 1-\mu$ and $C_2 = \mu$, the asymptotic covariance of the normalized centered cost process between (normalized) times α_1 and α_2 is equal to

$$\begin{aligned}
\frac{\mathrm{e}^{-\alpha_2}}{4}\Bigg(&\Big(\mathrm{I}_0(2\sigma\alpha_2) + 2\mathrm{I}_0(2\sigma\alpha_1)\phi(\alpha_2-\alpha_1) + \textstyle\sum_{i>0,j>0}\Big(\frac{\mu}{\sigma}\Big)^j \mathrm{I}_i(2\sigma\alpha_1)\,\mathrm{I}_{j-i}(2\sigma(\alpha_2-\alpha_1))\Big) \\
&-\mathrm{e}^{-\alpha_1}\Big(\mathrm{I}_0(2\sigma\alpha_1)\,\mathrm{I}_0(2\sigma\alpha_2) + 2\mathrm{I}_0(2\sigma\alpha_2)\phi(\alpha_1) + 2\mathrm{I}_0(2\sigma\alpha_1)\phi(\alpha_2) + \phi(\alpha_1)\,\phi(\alpha_2) \\
&\qquad + \alpha_1[(1-2\mu)\mathrm{I}_0(2\sigma\alpha_1) + \sigma\mathrm{I}_1(2\sigma\alpha_1)][(1-2\mu)\mathrm{I}_0(2\sigma\alpha_2) + \sigma\mathrm{I}_1(2\sigma\alpha_2)]\Big)\Bigg) \qquad (18)
\end{aligned}$$

5.2 Multivariate generating function

We assume here that the balls are thrown in two groups. The first group is thrown at the time t_1 (or during the interval $[1, t_1]$) and its balls are marked by y_1; we shall use the variable z_1 to help in distinguishing good balls from bad balls (see Section 3.1) : a good ball is marked as $(1-\mu)y_1z_1$ and a bad ball as $\mu y_1/z_1$. Similarly, the second group is thrown at the time t_2 (or during the interval $[t_1+1, t_2]$) and we use the variables y_2 and z_2 to mark its balls. The generating function describing the allocation of balls of the two groups in a single urn is

$$e^{y_1(\mu/z_1+(1-\mu)z_1)+y_2(\mu/z_2+(1-\mu)z_2)}. \qquad (19)$$

We shall use the variables u_i, v_i and w_i to indicate the state of the urn after throwing the first group ($i = 1$) and the second group ($i = 2$) : an urn which is bad at time t_i is marked by u_i; if it is neutral it is marked by v_i, and by w_i if it is good. We rewrite the function (19) as

$$e^{y_2(\mu/z_2+(1-\mu)z_2)} \cdot \sum_{n\in Z} I_n(2\sigma\, y_1) \left(\frac{\sigma\, z_1}{\mu}\right)^n$$

and we mark the state of the urn after throwing the first group of balls; we get

$$e^{y_2(\mu/z_2+(1-\mu)z_2)} \cdot \left[w_1 \sum_{n>0} I_n(2\sigma\, y_1) \left(\frac{\sigma\, z_1}{\mu}\right)^n + v_1 I_0(2\sigma\, y_1) + u_1 \sum_{n<0} I_n(2\sigma\, y_1) \left(\frac{\sigma\, z_1}{\mu}\right)^n\right].$$

Now we consider the second group of balls : we expand the term $e^{y_2(\mu/z_2+(1-\mu)z_2)}$, substitute z for z_1 and for z_2, and get

$$\begin{aligned} & w_1 \sum_{n>0,\ p\in Z} I_n(2\sigma\, y_1) I_p(2\sigma\, y_2) \left(\frac{\sigma\, z}{\mu}\right)^{n+p} \\ & + v_1 I_0(2\sigma\, y_1) \sum_{p\in Z} I_p(2\sigma\, y_2) \left(\frac{\sigma\, z}{\mu}\right)^p \\ & + u_1 \sum_{n<0,\ p\in Z} I_n(2\sigma\, y_1) I_p(2\sigma\, y_2) \left(\frac{\sigma\, z}{\mu}\right)^{n+p}. \end{aligned}$$

The sign of the exponent of z, $n + p$, determines the type of the urn at the time t_2. We first consider the case where the urn is neutral at the time t_1 : The factor of v_1, $\sum_{p\in Z} I_p(2\sigma\, y_2)\,(\sigma\, z/\mu)^p$, becomes

$$u_2 \sum_{p<0} I_p(2\sigma\, y_2) \left(\frac{\sigma}{\mu}\right)^p + v_2 I_0(2\sigma\, y_2) + w_2 \sum_{p>0} I_p(2\sigma\, y_2) \left(\frac{\sigma}{\mu}\right)^p.$$

Using the definitions of $\phi(y)$ and $\psi(y)$, we get the terms relative to the variable v_1 :

$$v_1 u_2 I_0(2\sigma\, y_1)\phi(y_2) + v_1 v_2 I_0(2\sigma\, y_1) I_0(2\sigma\, y_2) + v_1 w_2 I_0(2\sigma\, y_1)\psi(y_2).$$

We now consider the case where the urn is good at the time t_1 : this corresponds to the terms in w_1. The coefficient of w_1, $\sum_{n>0,\ p\in Z} I_n(2\sigma\, y_1) I_p(2\sigma\, y_2)\,(\sigma\, z/\mu)^{n+p}$, becomes

$$\begin{aligned} w_2 \sum_{n>0,\ n+p>0} I_n(2\sigma\, y_1) I_p(2\sigma\, y_2) \left(\frac{\sigma}{\mu}\right)^{n+p} & + v_2 \sum_{n>0} I_n(2\sigma\, y_1) I_{-n}(2\sigma\, y_2) \\ & + u_2 \sum_{n>0,\ n+p<0} I_n(2\sigma\, y_1) I_p(2\sigma\, y_2) \left(\frac{\sigma}{\mu}\right)^{n+p}. \end{aligned}$$

Now the coefficient of $w_1 v_2$ can be expressed simply in terms of the Bessel function I_0 (see the addition formula (23) in the Appendix) :

$$\sum_{n>0} I_n(2\sigma\, y_1) I_{-n}(2\sigma\, y_2) = [I_0(2\sigma\,(y_1 + y_2) - I_0(2\sigma\, y_1)\, I_0(2\sigma\, y_2)]/2 =: \Delta I(y_1, y_2).$$

Define

$$S(y_1,y_2) := \sum_{n>0,\ n+p>0} I_n(2\sigma\, y_1) I_p(2\sigma\, y_2) \left(\frac{\sigma}{\mu}\right)^{n+p} = \sum_{n>0} I_n(2\sigma\, y_1) \left(\frac{\sigma}{\mu}\right)^{n} \sum_{p>-n} I_p(2\sigma\, y_2) \left(\frac{\sigma}{\mu}\right)^{p};$$

then the coefficient of $w_1 w_2$ is equal to $S(y_1,y_2)$, and the coefficient of $w_1 u_2$ can be expressed in terms of I_0, ϕ and S. To do this, we simplify the term $\psi(y_1)\ e^{y_2}$ with the help of the addition formula (22) :

$$\begin{aligned}
\psi(y_1)e^{y_2} &= \sum_{n>0,\, q\in Z} I_n(2\sigma\, y_1) I_{q-n}(2\sigma\, y_2) \left(\frac{\sigma}{\mu}\right)^{q} \\
&= \sum_{n,q>0} I_n(2\sigma\, y_1) I_{q-n}(2\sigma\, y_2) \left(\frac{\sigma}{\mu}\right)^{q} + \sum_{n>0} I_n(2\sigma\, y_1) I_{-n}(2\sigma\, y_2) \\
&\quad + \sum_{n>0,\, q<0} I_n(2\sigma\, y_1) I_{q-n}(2\sigma\, y_2) \left(\frac{\sigma}{\mu}\right)^{q}.
\end{aligned}$$

The first sum of the right-hand side is equal to $S(y_1,y_2)$, and the second sum to $\Delta I(y_1,y_2)$; the third sum is the coefficient of $w_1 u_2$. Hence the terms relative to the variable w_1 are

$$w_1 w_2 S(y_1,y_2) + w_1 v_2 \Delta I(y_1,y_2) + w_1 u_2 (\psi(y_1)e^{y_2} - S(y_1,y_2) - \Delta I(y_1,y_2)).$$

The contribution of the terms in u_1 (the urn is bad at the time t_1) is similarly computed : Define

$$T(y_1,y_2) := \psi(y_1+y_2) - S(y_1,y_2) - I_0(y_1)\psi(y_2);$$

then the terms including the variable u_1 can be simplified and we get

$$u_1 w_2 T(y_1,y_2) + u_1 v_2 \Delta I(y_1,y_2) + u_1 u_2 (\phi(y_1)e^{y_2} - T(y_1,y_2) - \Delta I(y_1,y_2)).$$

The multivariate generating function describing the behavior of a single urn at the times t_1 and t_2 is thus

$$\begin{aligned}
&w_1 w_2 S(y_1,y_2) + w_1 v_2 \Delta I(y_1,y_2) + w_1 u_2 (\psi(y_1)e^{y_2} - S(y_1,y_2) - \Delta I(y_1,y_2)) \\
&\quad + v_1 w_2 I_0(2\sigma\, y_1)\psi(y_2) + v_1 v_2 I_0(2\sigma\, y_1) I_0(2\sigma\, y_2) + v_1 u_2 I_0(2\sigma\, y_1)\phi(y_2) \\
&+ u_1 w_2 T(y_1,y_2) + u_1 v_2 \Delta I(y_1,y_2) + u_1 u_2 (\phi(y_1)e^{y_2} - T(y_1,y_2) - \Delta I(y_1,y_2)).
\end{aligned}$$

For $u_i = v_i = w_i = 1$ (we "forget" the state of the urn at the times t_1 and t_2), we get back the generating function describing the allocation of the two types of balls, which is simply $e^{y_1+y_2}$. If we consider the urn at the time t_1 ($u_2 = v_2 = w_2$), we get

$$e^{y_2}\ (u_1\phi(y_1) + v_1 I_0(2\sigma\, y_1) + w_1\psi(y_1)) = e^{y_2} f(u_1, v_1, w_1; y_1).$$

If we consider the urn at the time t_2 and forget its state at the time t_1 ($u_1 = v_1 = w_1 = 1$), we get

$$u_2 \left(e^{y_1+y_2} - I_0(2\sigma\,(y_1+y_2)) - \psi(y_1+y_2)\right) + v_2 I_0(2\sigma\,(y_1+y_2)) + w_2\psi(y_1+y_2)$$

These formulæ could also be derived directly, by marking directly the parameters of interest in suitable generating functions.

5.3 Another derivation of the covariance

The bivariate generating function that we have just computed describes what happens in the urns at two different times, and we can use it to compute the covariance. First, we get the multivariate generating function of the cost at two different times : We use the variables x_1 and x_2 to mark the cost at the times t_1 and t_2; the variables y_1 and y_2 are used to mark respectively the number of balls at the time t_1 and the number of balls added between t_1 and t_2. The function $H(x_1, x_2, y_1, y_2)$ is, as before, equal to $h(x_1, x_2, y_1, y_2)^n$, with h describing what happens in an urn : h is obtained from the multivariate function in u_i, v_i and w_i by substituting $x_i^{C_1}$ for u_i, $x_i^{C_0}$ for v_i and $x_i^{C_2}$ for w_i $(i = 1, 2)$; we get

$$
\begin{aligned}
&h(x_1, x_2, y_1, y_2) = \\
&x_1^{C_2} x_2^{C_2} S(y_1, y_2) + x_1^{C_2} x_2^{C_0} \Delta I(y_1, y_2) + x_1^{C_2} x_2^{C_1} (\psi(y_1) e^{y_2} - S(y_1, y_2) - \Delta I(y_1, y_2)) \\
&+ x_1^{C_0} x_2^{C_2} I_0(2\sigma\, y_1) \psi(y_2) + x_1^{C_0} x_2^{C_0} I_0(2\sigma\, y_1) I_0(2\sigma\, y_2) + x_1^{C_0} x_2^{C_1} I_0(2\sigma\, y_1) \phi(y_2) \\
&+ x_1^{C_1} x_2^{C_2} T(y_1, y_2) + x_1^{C_1} x_2^{C_0} \Delta I(y_1, y_2) + x_1^{C_1} x_2^{C_1} (\phi(y_1) e^{y_2} - T(y_1, y_2) - \Delta I(y_1, y_2)).
\end{aligned}
$$

We then use this function as follows : Let $Cost_1$ and $Cost_2$ be the costs at two different times t_1 and t_2 $(t_1 < t_2)$. The covariance is defined as $E[Cost_1 \cdot Cost_2] - E[Cost_1] \cdot E[Cost_2]$. The expectation of the cost at a time t_1, knowing that we throw k_1 balls up to the time t_1, is easily computed from the generating function $G(x, y)$ for the cost, given by the equation (8), as $[y^{k_1}] G'_x(1, y) / [y^{k_1}] G(1, y)$ (see Section 3.1). Similarly, the expectation of the cost at a time t_2, knowing that we throw k_2 balls in the interval $]t_1, t_2]$, and a total of $k_1 + k_2$ balls, is $[y^{k_1+k_2}] G'_x(1, y) / [y^{k_1+k_2}] G(1, y)$. Now the expectation of the product, given that we throw k_1 balls in the interval $]0, t_1]$ and k_2 balls in the interval $]t_1, t_2]$, can be obtained as

$$
E[Cost_1 \cdot Cost_2] = \frac{[y^{k_1} y^{k_2}] H''_{x_1 x_2}(1, 1, y_1, y_2)}{[y^{k_1} y^{k_2}] H(1, 1, y_1, y_2)}.
$$

Of course, $[y^{k_1} y^{k_2}] H(1, 1, y_1, y_2) = 1/(k_1! k_2!)$. Now $H''_{x_1 x_2} = n h''_{x_1 x_2} h^{n-1} + n(n-1) h'_{x_1} h'_{x_2} h^{n-2}$ and we get an expression for $H''_{x_1 x_2}(1, 1, y_1, y_2)$. We can obtain the covariance by taking the derivatives of H for $x_1 = x_2 = 1$, then extracting the coefficients (we shall need a second-order approximation for the ones that have a multiplicative factor $n(n-1)$; a first-order approximation suffices for those terms that have a multiplicative factor of order n), then injecting these approximations in the expression for the covariance; this approach requires also that we get a second-order approximation of the expectations (the first-order terms have a multiplicative factor of order n^2 and are canceled by the terms with a similar weight in $E[Cost_1 \cdot Cost_2]$).

5.4 Bidimensional distribution

We show now that the bi-dimensional distribution is asymptotically normal.

To do this, we shall show that the characteristic function of the normalized costs converges towards the characteristic function of a bivariate normal distribution. Define

$$
\xi_1 := \frac{Cost_1 - E[Cost_1]}{\sqrt{n}}; \qquad \xi_2 := \frac{Cost_2 - E[Cost_2]}{\sqrt{n}}.
$$

The bivariate characteristic function of ξ_1 and ξ_2 is obtained from the generating function of the costs $H(x_1, x_2, y_1, y_2)$ (see Section 5.3) as

$$F_{\xi_1,\xi_2}(t_1,t_2) = e^{-\frac{i}{\sqrt{n}}(t_1 E[Cost_1]+t_2 E[Cost_2])} \left[\frac{y_1^{k_1}}{k_1!}\,\frac{y_2^{k_2}}{k_2!}\right] \left\{H\left(e^{it_1/\sqrt{n}}, e^{it_2/\sqrt{n}}, y_1, y_2\right)\right\}.$$

We have thus to evaluate the coefficient $[y_1^{k_1}\ y_2^{k_2}]\left\{h\left(e^{it_1/\sqrt{n}}, e^{it_2/\sqrt{n}}, y_1, y_2\right)^n\right\}$. To do this, we shall use, as we did before when we met coefficients of the n^{th} power of a function, a saddle-point approximation. We write the coefficient as

$$\left(\frac{1}{2i\pi}\right)^2 \oint\oint h\left(e^{it_1/\sqrt{n}}, e^{it_2/\sqrt{n}}, y_1, y_2\right)^n \frac{dy_1}{y_1^{k_1+1}}\,\frac{dy_2}{y_2^{k_2+1}},$$

and we use for integration contours two circles centered at the origin, and passing through the saddle points, i.e. whose radii r_1 and r_2 are the respective solutions of the equations $y_1 h'_{y_1}/h = (k_1+1)/n$ and $y_2 h'_{y_2}/h = (k_2+1)/n$. We do not need to solve exactly these equations : To show the convergence of the characteristic function towards the characteristic function of a bidimensional distribution, we shall let $n \to +\infty$; hence $e^{it_1/\sqrt{n}} \to 1$ and $e^{it_2/\sqrt{n}} \to 1$, and we can choose for approximate saddle points the solutions of the equations $y_1 h'_{y_1}(1,1,y_1,y_2)/h(1,1,y_1,y_2) = k_1/n$ and $y_2 h'_{y_2}(1,1,y_1,y_2)/h(1,1,y_1,y_2) = k_2/n$. These solutions are $r_1 = \alpha_1$ and $r_2 = \alpha_2$, and we integrate on the contours $\{y_1 = \alpha_1 e^{i\theta_1}, -\pi \le \theta_1 \le \pi\}$ and $\{y_2 = \alpha_2 e^{i\theta_2}, -\pi \le \theta_2 \le \pi\}$. Now the integral for $|\theta_1| \le \pi, |\theta_2| \le \pi$ can be broken into two parts : the central part is for $|\theta_1|, |\theta_2| \le \log n/\sqrt{n}$, and gives the main contribution to the integral; the remainder gives error terms. The main point is that the derivatives of second or third order of H, at or around the point $x_1 = x_2 = 0$, are of order n. We do not give here the detailed computations, which are rather cumbersome; the interested reader can go back to [15, p. 167-170] or to [8, p. 402-408], where similar bivariate methods are applied to the same kind of problem, namely to study limiting distributions through their generating function.

A similar approach can probably be used to prove that the finite-dimensional distributions are asymptotically normal; however we have first to get the generating function for a finite number of costs. Although this poses no theoretical difficulty, the number of terms of the generating function for p costs is 3^p, which makes it difficult to write the function in a pleasant form.

6 Conclusion and extensions

We have presented a new urn model to study the generalization error in learning symmetric functions with noise.

From the initial learning-theoretic viewpoint, this detailed analysis reveals that a typically exponential learning curve can undergo subtle distortions when random classification noise is introduced : the generalization error is no more a simple exponential but the product of an exponential with series of Bessel functions. This would have been difficult to

characterize in numerical simulations, though the fact that noisy learning curves are no more simple exponentials was already apparent.

We have shown the gaussian behavior of the limiting distribution and process, when the number of balls k and the number of urns n are proportional. It should be noted that the relation is not strict : our results can probably be extended to k/n belonging to a closed interval of $]0, +\infty[$ (the *central domain* of [15]). However, when n and k no longer have the same growth rate, we can expect a different behavior. The analogy with the empty urns model suggests that, for example, we might get Poisson results for $k = n \log n$. Possible extensions also include the waiting time until some cost is reached, i.e. until the error of the learning process becomes smaller than some bound.

The generality of that pattern of fluctuations in learning problems remains to be assessed.

We believe that another contribution of our paper is the presentation of a new kind of admissible construction : the majority phenomenon that comes from building a combinatorial structure on two types of objects (good and bad in this paper), then deciding on the type of the structure according to the type of the majority of the basic objects. For example, we can have two types of basic objects, build cycles on theses objects and combine these cycles into a set, then ask for the number of cycles of the set that have a majority of elements of one type, or an equal number of elements of each type. It should be possible to extend the distribution results on the number of components presented by Flajolet and Soria [5] to study the number of components of a given type (good, bad or neutral) for various combinatorial constructs.

Acknowledgments
We thank P. Flajolet for information on Bessel functions and G. Louchard for information on gaussian processes.

7 Appendix

We give in this part some mathematical results that we need for our analysis : asymptotic expansions at order 2 for coefficients of functions of the type $e^{ny} f(y)$, basic facts on Bessel functions, and some properties of the function ϕ defined in Section 3.3.

7.1 Asymptotic expansions

We need in several places of our computations the first terms of the asymptotic expansion of a coefficient of the type $[y^k]\{e^{ny} f(y)\}$. This is basically a special case (for $g(y) = e^y$) of a variation on a coefficient of the type $[y^k]\{g(y)^n\}$. Such coefficients were studied, when n and k grow to infinity while staying (roughly) proportional, by Daniels [3], who gave the asymptotic equivalent, and by Good [10], who extended the results of Daniels to get a full asymptotic expansion. In [9], we presented an extension of Daniels's result to allow

for a factor $f(y)$ of slower growth rate, which is the case if $f(y)$ does not depend on n at all. All these papers use a saddle point approximation, which we can adapt to deal with coefficients $[y^k]\{e^{(n-a)y}f(y)\}$. We get

$$[y^k]\{e^{(n-a)y}f(y)\} = \frac{e^{(n-a)\alpha}f(\alpha)}{\alpha^{k+1}\sqrt{2\pi r(\alpha)}}\left(1+\frac{\epsilon^2+2\epsilon}{2k}-\frac{1}{12k}+o(\frac{1}{k})\right),$$

with $\alpha = k/n$ and $\epsilon = 1+a\alpha-\delta f(\alpha)$; $\delta f(y) = yf'(y)/f(y)$ and $r(\alpha) = f''(\alpha)/f(\alpha) - (f'/f)^2(\alpha) + (k+1)/\alpha^2$.

7.2 Bessel functions

We refer the reader to the book by Whittaker and Watson [21, Ch. 17] or the treatise by Watson [20] for detailed information.

Definition

$$I_n(t) = \sum_r \frac{1}{r!(r+n)!}\left(\frac{t}{2}\right)^{2r+n}.$$

The summation is for $r \geq 0$ if $n \geq 0$, and for $r \geq -n$ if $n < 0$. Note that $I_{-n} = I_n$. We use mostly the functions

$$\begin{aligned} I_0(t) &= \sum_{r\geq 0}\frac{1}{r!^2}\left(\frac{t}{2}\right)^{2r}; \\ I_1(t) &= \sum_{r\geq 0}\frac{1}{r!(r+1)!}\left(\frac{t}{2}\right)^{2r+1}. \end{aligned}$$

Derivatives of Bessel functions

$$I_n'(t) = I_{n+1}(t)+\frac{n}{t}I_n(t) = \frac{I_{n-1}(t)+I_{n+1}(t)}{2}; \tag{20}$$

$$I_n''(t)+\frac{1}{t}I_n'(t) = (1+\frac{n^2}{t^2})I_n(t). \tag{21}$$

In particular, $I_0' = I_1$ and $I_1'(t) = I_2(t)+(1/t)I_1(t)$.

Addition formulæ

$$\sum_{n\in Z} I_n(x)I_{q-n}(y) = I_q(x+y). \tag{22}$$

For $q = 0$,

$$\sum_{n\in Z} I_n(x)I_n(y) = I_0(x+y). \tag{23}$$

Asymptotic behavior
For real $t \to +\infty$,

$$I_n(t) = \frac{e^t}{\sqrt{2\pi t}}\left(1-\frac{4n^2-1}{8t}+O\left(\frac{1}{t^2}\right)\right).$$

7.3 The function $\phi(y)$

We recall that the function ϕ is defined as a (weighted) sum of Bessel functions :

$$\phi(y) = \sum_{p \geq 1} \left(\frac{\mu}{\sigma}\right)^p I_p(2\sigma\, y). \tag{24}$$

An alternative definition uses the Lommel functions (see [20, p. 537])

$$U_n(w,z) = \sum_{m \geq 0} (-1)^m (w/z)^{n+2m} J_{n+2m}(z),$$

with $J_p(z)$ the classical Bessel function ($I_p(y) = (-i)^p J_p(iy)$) :

$$\phi(y) + I_0(2\sigma\, y) = U_0(-2i\mu\, y, 2i\sigma\, y) + iU_1(-2i\mu\, y, 2i\sigma\, y).$$

The function ϕ is increasing.

Asymptotic behavior
For $y \to +\infty$:

$$\phi(y) \sim K\, \frac{e^{2\sigma\, y}}{\sqrt{4\pi\, \sigma\, y}} \qquad \text{with} \qquad K = \frac{\mu}{\sigma - \mu} = \frac{1}{\sqrt{\frac{1-\mu}{\mu}} - 1}.$$

Remark : the factor K is equal to 0 for $\mu = 0$, is increasing with μ, and becomes infinite when $\mu \to 1/2$.

Sketch of proof :
The proof begins with the equality

$$e^{\frac{z}{2}(u+\frac{1}{u})} = \sum_{p=-\infty}^{+\infty} u^p I_p(z).$$

Hence $I_p(z) = [u^p]\{e^{\frac{z}{2}(u+\frac{1}{u})}\} = (1/2i\pi) \oint e^{(z/2)(u+1/u)} u^{-p-1} du$, and we obtain an integral representation of ϕ for $t = \mu/\sigma \in]0,1[$:

$$\phi(y) = \frac{1}{2i\pi} \oint e^{\sigma\, y(u+\frac{1}{u})} \sum_{p \geq 1} \frac{t^p}{u^{p+1}}\, du = \oint e^{h(y,u)} du$$

with $h(y,u) = \sigma y(u + 1/u) + \log(t/(u(u-t))$. The integration contour circles around t; the saddle point heuristic suggests that we choose as contour a circle of radius the value of u that cancels $h'_u(y,u)$. For large y, this value is close to 1 and we choose for integration contour the circle $\{u = e^{i\theta}, -\pi \leq \theta \leq \pi\}$. The details can be worked out without any major difficulty and we obtain the asymptotic value of $\phi(y)$. If desired, the saddle point method can give more terms of the expansion.

Differential equation

The function ϕ satisfies a linear differential equation, which can be used to give an expression of the derivative $\phi^{'}(y)$. We simply derive the relation (24) and use the equality (20); we obtain

$$\phi^{'}(y) = \phi(y) + \mu\ I_0(2\sigma\ y) - \sigma\ I_1(2\sigma\ y). \tag{25}$$

Define $z = 2\sigma\ y$ and $\phi_1(z) = \phi(z/2\sigma)$; ϕ_1 is a solution of the differential equation

$$2\sigma\ \phi_1^{'}(z) = \phi_1(z) + \mu I_0(z) - \sigma I_1(z). \tag{26}$$

We seek a solution of the type $\phi_1(z) = e^{z/2\sigma}\theta(z)$, with

$$\theta^{'}(z) = \frac{\mu}{2\sigma} e^{-z/2\sigma} I_0(z) - \frac{1}{2} e^{-z/2\sigma} I_1(z).$$

Now $I_1(z) = I_0^{'}(z)$ and $\int e^{-z/2\sigma} I_1(z) = e^{-z/2\sigma} I_0(z) + (1/2\sigma) \int e^{-z/2\sigma} I_0(z)$. This gives the general solution

$$\phi_1(z) = -\frac{1-2\mu}{4\sigma} e^{z/2\sigma} \int_0^z e^{-t/2\sigma} I_0(t)\, dt - \frac{1}{2} I_0(z) + C e^{z/2\sigma}.$$

The constant C is chosen such that $\phi(0) = 0$; hence $C = 1/2$ and we get an expression for ϕ :

$$\phi(y) = \frac{1}{2}\,(e^y - I_0(2\sigma\, y)) - \frac{1}{2}(1-2\mu)\, e^y\, \theta_0(2\sigma\, y) \quad \text{with} \quad \theta_0(z) = \int_0^z e^{-t} I_0(2\sigma\, t)\, dt. \tag{27}$$

The asymptotic expression of $\phi(y)$ together with the expression of $\phi^{'}(y)$ in terms of $\phi(y)$ give, for $y \to +\infty$,

$$\phi^{'}(y) \sim \frac{2\sigma\,\mu}{\sigma - \mu}\ \frac{e^{2\sigma\, y}}{\sqrt{4\pi\ \sigma\ y}}.$$

The second derivative of ϕ satisfies the following relation, which helps to simplify some expressions in the computation of the variance :

$$\phi^{''}(y) = \phi(y) + \mu^2 I_0(2\,\sigma y) + (2\mu - 1)\sigma\, I_1(2\,\sigma y) - \sigma^2 I_2(2\,\sigma y).$$

7.4 Identities

The following identities prove useful during the derivation of the covariance.

$$\begin{array}{l}
\sum_{0\le h} \frac{(\sigma\ y)^{2h}}{h!h!}\,(2h) = 2\,\sigma\, y\, \mathrm{I}_1(2\,\sigma y) \\
\sum_{0\le h} \frac{(\sigma\ y)^{2}}{h!h!}\,((2\,h)^2 - (2\,h)) = (2\,\sigma\ y)\Big(2\,\sigma\, y\, \mathrm{I}_0(2\,\sigma y) - \mathrm{I}_1(2\,\sigma y)\Big) \\
\sum_{0\le h<l} \frac{\mu^l\,(1-\mu)^h\ y^{h+l}}{h!l!}\,(h+l) = y\ \frac{\partial \phi_\mu(y)}{\partial y} \\
\sum_{0\le h<l} \frac{\mu^l\,(1-\mu)^h\ y^{h+l}}{h!l!}\,((h+l)^2 - (h+l)) = y^2\ \frac{\partial^2 \phi_\mu(y)}{\partial y^2}
\end{array} \tag{28}$$

7.5 Applications of Stirling's formula

Development to the second order of

$$\frac{1}{n^k}\frac{y\,n!}{(y\,n-k)!}\,(1-\frac{r}{n})^{y\,n-k}$$

when $n \to \infty$:

$$\begin{aligned} y^k \mathrm{e}^{-r\,y}\Big(1 &- \frac{y\,r^2-2\,kr}{2\,n} - \frac{k^2-k}{2\,ny} \\ &+ \frac{r^2}{n^2}(\frac{k}{2} - \frac{yr}{3} + \frac{(yr-2k)^2}{8}) \\ &+ \frac{k^2}{n^2y^2}(\frac{3}{8} - \frac{k}{6} - \frac{k^2}{8} + (yr-k)^2) + o(1/n^3)\Big) \end{aligned} \tag{29}$$

References

[1] D. Barraez. *Diametro de Transmission, Ciclos Dominantes y el Problema Classica de Colocaciones*. PhD thesis, Universidad de Caracas, 1994.

[2] E.A. BENDER and L.B. RICHMOND. Central and local limit theorems applied to asymptotic enumeration ii: Multivariate generating functions. *Journal of Combinatorial Theory, Series A*, 34:255–265, 1983.

[3] H.E. DANIELS. Saddlepoint approximations in statistics. *Annals of Mathematical Statistics*, 25:631–650, 1954.

[4] P. Diaconis and D. Freedman. A dozen de finetti-style results in search of a theory. *Ann. Inst. Henri Poncaré*, 23(2):397–423, 1987.

[5] P. FLAJOLET and M. SORIA. Gaussian limiting distributions for the number of components in combinatorial structures. *Journal of Combinatorial Theory (A)*, 53(2):165–182, March 1990.

[6] P. FLAJOLET and J. VITTER. *Handbook of Theoretical Computer Science*, chapter Average-Case Analysis of Algorithms and Data Structures. North Holland, 1989.

[7] D. GARDY. Méthode de col et lois limites en analyse combinatoire. *Theoretical Computer Science, Part A*, 94(2):261–280, March 1992.

[8] D. GARDY. Join sizes, urn models and normal limiting distributions. *Theoretical Computer Science (A)*, 131:375–414, August 1994.

[9] D. GARDY. Some results on the asymptotic behaviour of coefficients of large powers of functions. *Discrete Mathematics*, 139:189–217, 1995.

[10] I.J. GOOD. Saddle-point methods for the multinomial distribution. *Annals of Mathematical Statistics*, 28:861–881, 1957.

[11] I.P. GOULDEN and D.M. JACKSON. *Combinatorial enumeration.* Wiley & Sons, 1983.

[12] D. Haussler, M. Kearns, H. S. Seung, and N. Tishby. Rigorous learning curve bounds from statistical mechanics. In *Proc. 7th Annu. ACM Workshop on Comput. Learning Theory*, pages 76–87. ACM Press, New York, NY, 1994.

[13] N.L. JOHNSON and S. KOTZ. *Urn models and their application.* Wiley & Sons, 1977.

[14] M. Kearns and U. Vazirani. *Topics in Learning Theory.* MIT Press, 1994.

[15] V. KOLCHIN, B. SEVAST'YANOV, and V. CHISTYAKOV. *Random Allocations.* Wiley & Sons, 1978.

[16] K. NISHIMURA and M. SIBUYA. Occupancy with two types of balls. *Ann. Inst. Statist. Math*, 40(1):77–91, 1998.

[17] T. Yu. POPOVA. Limit theorems in a model of distribution of particles of two types. *SIAM Journal on Theory of Probability and its application*, pages 511–516, 1967.

[18] B. I. SELIVANOV. On waiting time in the schema of random allocation of coloured particles. *Discrete Math. Appl.*, 5(1):73–82, 1995.

[19] H.S. Seung, H. Sompolinsky, and N. Tishby. Statistical mechanics of learning from examples. *Physical review A*, 45(8):6056–6091, april 1992.

[20] G.N. WATSON. *Theory of Bessel functions.* Cambridge University Press, 1922. Second edition 1941.

[21] E.T. WHITTAKER and G.N. WATSON. *A Course of Modern Analysis.* Cambridge University Press, 1927 (Fourth edition).