RANDOM VARIATION AND CONCENTRATION EFFECTS IN PCR

PETER JAGERS AND FIMA KLEBANER

Abstract. Even though the efficiency of the PCR reaction decreases, analyses are made in terms of Galton-Watson processes, or simple deterministic models with constant replication probability (efficiency). Recently Schnell and Mendoza have suggested that the form of the efficiency, can be derived from enzyme kinetics. This results in the sequence of molecules numbers forming a stochastic process with the properties of a branching process with population size dependence, which is supercritical, but has a mean reproduction number that approaches one. Such processes display ultimate linear growth, after an initial exponential phase, as is the case in PCR. It is also shown that the resulting stochastic process for a large Michaelis Menten constant behaves like the deterministic sequence x_n arising by iterations of the function $f(x) = x + x/(1 + x)$.

1. The problem

Since its first presentation (Saiki *et al.*, 1985), the polymerase chain reaction (PCR) has emerged as the established technology for creating multiple copies of a given nucleotide sequence. The method has a wide use and is now even commercially accessible. Easy-read descriptions of the elementary chemistry of it are available, *e.g.* in http://us.labsystems.roche.com/products/pcr/pcrintro.shtml. Raeymaekers (2000) overviews the mathematics of the method on a similar basic level.

PCR is a stepwise procedure where in each step a molecule either remains or is replaced by two copies. More careful descriptions discern between different types of molecules, like long or short, *cf.* Nedelman *et al.* (1992). In the copying procedure errors may also occur, giving rise to new types, mutants (Krawczak *et al.*, 1989, Sun, 1995, Olofsson and Shaw, 2002). Disregarding those aspects, we concentrate upon the growth process itself in order to obtain a more refined model than usual of quantitative PCR, *i.e.* PCR where the successive molecule numbers are recorded.

This leads to viewing the reaction as a (single-type) Galton-Watson branching process, each individual having one or two offspring in the

Work supported by the Australian Research Council, Adlerbertska Forskningsfonden and the Knut och Alice Wallenberg Foundation.

next generation. Of course, molecules do not divide like cells, nor do they give birth to offspring like birds or plants. They replicate in a catalytic process. Nevertheless, the mathematical pattern is the same as that of population growth in discrete time, an individual being replaced by one or two individuals in the next round (generation).

The probability of the latter event is usually termed the *efficiency* in the present connection. It is natural from the experimental setup that the efficiency of the reaction should decrease, and indeed data exhibit an exponential growth phase ultimately turning into linear increase of molecule numbers. Still, basically all analysis methods presume a constant efficiency, possibly after a perfunctory reference to the limitations of this assumption. (Notable exceptions are provided by the series of papers by Jacob and Peccoud, who introduced a special form of branching process in varying environments, and Nedelman *et al.* (1992), considering random environments.)

As a consequence, traditional approaches have to confine themselves to the initial period of the reaction, during which exponential growth seems to prevail, and the efficiency has not decreased substantially. But already Saiki *et al.*, pointed out that the exponential growth eventually turns linear. Our purpose is to describe the whole process, including this change, in terms of branching processes with population-size dependence. The possibility of efficiency influenced by molecule numbers has been mentioned but not pursued in literature, *e.g.* by Weiss and von Haeseler (1997).

We shall use an explicit approach, starting from the recent analyses of reaction efficiency, as determined by the free substrate concentration, being essentially the same as the amount of target DNA, i.e. the number of what we simply called molecules above. To the extent that Michaelis-Menten kinetics can be assumed to hold, it follows (Schnell and Mendoza, 1997) that, when the number of molecules is z, the probability of successful division is given by

$$p(z) = \frac{K}{K+z},$$

where K is the Michaelis-Menten constant of the reaction (usually quite large, so that the efficiency is close to one during the initial cycles). Indeed, by Michaelis-Menten kinetics the reaction rate is

$$\frac{V_{\max} z}{K+z}$$

when the free substrate amount is z, $V_{\max}$ denoting the so called maximum velocity, as usual. During one reaction step, of fixed duration, starting from z molecules the increase is proportional to this rate. On the other hand, if $p(z)$ denotes the efficiency, as a function of z, the increase during one cycle starting from z molecules is $zp(z)$. Equating

the two expressions, we obtain

$$p(z) = \frac{C}{K+z},$$

for some constant C. Since this should decrease from one, as z increases from zero, $C = K$.

In branching process terminology, we thus consider a population-size dependent binary splitting Galton-Watson process, where the alternative to splitting is remaining into the next generation (experiment cycle), or equivalently giving birth to one offspring. Since the expected number of offspring per individual is

$$m(z) = 1 - p(z) + 2p(z) = 1 + K/(K+z),$$

population size still assumed to be z, this is a *near-critical* process, as studied by Klebaner (1984).

The following section shows that for this model the slope of the linear increase of molecules during the saturation phase should equal the Michaelis-Menten constant K. The biological reader may have difficulties in following the detailed arguments, but the conclusion has an obvious significance.

2. Population-size dependent branching processes

Write $Z_0, Z_1, Z_2, \ldots$ for the number of molecules in the successive cycles. The initial number may be known or unknown, fixed or random. In PCR it would often be natural to assume that it is a Poisson random variable, with a mean to be estimated. Initially, for starting numbers of the usual size, Z_n would be much smaller than Michaelis-Menten constant, and the conditional expectation of the the molecule number in the n-th cycle, given the number Z_{n-1} in the preceding round would satisfy

$$\mathbb{E}[Z_n|Z_{n-1}] = Z_{n-1} + \frac{KZ_{n-1}}{K+Z_{n-1}} \approx 2Z_{n-1},$$

giving the well-known exponential style growth of $\{Z_n\}$. But as $n \to \infty$ so does Z_n and therefore rather

$$\mathbb{E}[Z_n|Z_{n-1}] \approx Z_{n-1} + K.$$

By dominated convergence indeed

$$\mathbb{E}[Z_n] = \mathbb{E}[Z_{n-1}] + \mathbb{E}[\frac{KZ_{n-1}}{K+Z_{n-1}}] =$$

$$\ldots = \mathbb{E}[Z_0] + \sum_{k=0}^{n-1} \mathbb{E}[\frac{KZ_k}{K+Z_k}] \sim Kn,$$

as $n \to \infty$. This shows that we should expect linear growth at the rate K ultimately. (We use $a_n \sim b_n$ to mean that $\lim_{n\to\infty} a_n/b_n = 1$.) But even more holds true.

Indeed, when the total molecule number is z, the variance of the offspring distribution of any single individual is

$$\sigma^2(z) = 4p(z) + 1 - p(z) - m^2(z) = p(z)(1-p(z)) = \frac{Kz}{(K+z)^2}.$$

Thus, the variance of the number Z_n of molecules at the n-th cycle can be decomposed into

$$\mathrm{Var}[Z_n] = \mathbb{E}[\mathrm{Var}[Z_n|Z_{n-1}]] + \mathrm{Var}[\mathbb{E}[Z_n|Z_{n-1}]] =$$

$$= \mathbb{E}[\frac{KZ_{n-1}^2}{(K+Z_{n-1})^2}] + \mathrm{Var}[Z_{n-1} + \frac{KZ_{n-1}}{K+Z_{n-1}}].$$

In this the first term converges towards K from below, as n passes. The second expression is $\sim \mathrm{Var}[Z_{n-1}]$, for basically the same reason,

$$\frac{KZ_{n-1}}{K+Z_{n-1}} \to K$$

in a bounded fashion. It follows that

$$\mathrm{Var}[Z_n] \sim Kn, n \to \infty.$$

This yields the mean square convergence of Z_n/n towards K:

$$\mathbb{E}[(Z_n/n - K)^2] = \mathrm{Var}[Z_n]/n^2 + (\mathbb{E}[Z_n]/n - K)^2 \to 0.$$

However, there is also almost sure convergence, *i.e.* with probability one the sequence $Z_n/n \to K$. This is important, as it means that we can follow the successive molecule numbers of any one experiment, and rest assured about the convergence.

To prove it, first make the Doob decomposition

$$Z_n = Z_0 + A_n + M_n,$$

where

$$A_n = \sum_{k=0}^{n-1} \frac{KZ_k}{K+Z_k}$$

is an increasing sequence and $A_n/n \to K$ almost surely, as $n \to \infty$, and

$$M_n = \sum_{k=1}^{n} Z_k - \mathbb{E}[Z_k|Z_{k-1}]$$

constitutes a martingale with respect to $\mathcal{F}_n$, the sigma-algebra generated by $Z_0, Z_1, \dots Z_n$. It satisfies

$$\mathbb{E}[(M_n - M_{n-1})^2|\mathcal{F}_{n-1}] = \mathrm{Var}[Z_n|\mathcal{F}_{n-1}] = \mathrm{Var}[Z_n|Z_{n-1}] =$$

$$= \frac{KZ_{n-1}^2}{(K+Z_{n-1})^2} \le K.$$

Hence,

$$\sum_{n=1}^{\infty} \mathbb{E}[(M_n - M_{n-1})^2]/n^2 < \infty,$$

and so by the Law of Large Numbers for martingales (*e.g.* Chow and Teicher, 1997, p. 415), $M_n/n \to 0$ a. s. This completes the proof that $Z_n/n \to K$ also with probability one.

3. Deterministic approximations for large Michaelis-Menten constants

In this section we analyse the process by using the fact that the Michaelis-Menten constant K is large and study the normalised process $X_n^K = Z_n/K$, making use of the observation that the efficiency is a function of $x = z/K$ only, as pointed out by Schnell and Mendoza (1997b), $K/(K+z) = 1/(1+x)$. We shall show that as $K \to \infty$, X_n^K behaves as the deterministic sequence x_n arising by iterations of the function $f(x) = x + x/(1+x)$,

$$x_{n+1} = f(x_n), \tag{1}$$

a discrete time analogue of what Schnell and Mendoza referred to as the dimensionless reduced concentration.

Details may seem cumbersome for readers not so well versed in stochastic processes. But again the outcome has a clear biological significance, allowing us to define exponential growth at or above a certain rate in terms of the reduced concentration.

If we write $\xi_1, \xi_2 \ldots$ for the offspring of the various individuals in the n:th cycle then, given Z_n these are independent random variables all with

$$\mathbb{P}(\xi_i = 2 | X_n^K = x) = 1 - \mathbb{P}(\xi_i = 1 | X_n^K = x) = \frac{1}{1+x}$$

for $i = 1, 2, 3, \ldots$. Further,

$$X_{n+1}^K = \frac{1}{K} \sum_{i=1}^{KX_n^K} \xi_i = f(X_n^K) + \eta_{n+1}^K, \tag{2}$$

where

$$f(x) = \mathbb{E}[X_{n+1}^K | X_n^K = x] = x + \frac{x}{1+x},$$

and

$$\eta_{n+1}^K \to 0,$$

as $K \to \infty$, by the law of large numbers, and hence

$$X_n^K \to x_n. \tag{3}$$

For more details, see Klebaner (1993).

To see the behaviour of x_n for large n, first notice that it is an increasing sequence, and hence either has a limit or tends to infinity. Since f is continuous and has no fixed points, the former possibility is ruled out. Next we write

$$x_{n+1} = x_n + 1 - \frac{1}{1+x_n},$$

and iterate to obtain

$$x_{n+1} = x_0 + n - \sum_{i=1}^{n} \frac{1}{1+x_i}.$$

As $1/(1+x_n) \to 0$ as $n \to \infty$, the Kronecker lemma gives that as $n \to \infty$,

$$x_n/n \to 1.$$

The preceding section showed that for any given K,

$$X_n^K/n \to 1,$$

as $n \to \infty$. Hence X_n^K behaves very much like its deterministic analogue x_n. This allows us to define "the region of exponential growth" of the process, by taking it to be the same as for the deterministic sequence x_n. We include plots of the latter, exhibiting the changeover from exponential into linear growth (Figures 1 and 2), using the same starting values as Schnell and Mendoza (1997b).

The deviation of X_n^K from x_n is of order $1/\sqrt{K}$, as given by the following Central Limit Theorem, see Theorem 2 of Klebaner (1993): as $K \to \infty$

$$\sqrt{K}(X_n^K - x_n) \Rightarrow N(0, D_n^2), \tag{4}$$

where the arrow $\Rightarrow$ denotes convergence in distribution,

$$D_n^2 = \sum_{j=0}^{n-1} v^2(x_j) \prod_{i=j}^{n-1} (f'(x_i))^2, n = 1, 2, 3, \ldots, \tag{5}$$

and

$$v^2(x) = x\sigma^2(x) = \frac{x^2}{(x+1)^2}.$$

Thus, countings of molecules, normed by the Michaelis Menten constant, are approximately normally distributed, around x_n, with a variance whose size is of the order $1/K$,

$$X_n^K \approx N(x_n, \frac{D_n^2}{K}).$$

The theory underlying these results has been generalized to a multivariate setting by Watkins (2000), who referred to it as consistency and fluctuations theorems in biology. In a more accessible form, the univariate case was also given in Klebaner (1997).

We now give an asymptotic bound for the variances D_n^2 in (5) as n grows, so that

$$X_n^K \approx N(x_n, \frac{n}{K}). \tag{6}$$

Since K is large this distribution is pretty much concentrated about x_n. Using $f'(x) = 1 + 1/(1+x)^2$ and $x_i \sim i$, we can see that the infinite product $\prod_{i=1}^{\infty}(f'(x_i))^2$ converges to some number $C_1 > 0$. Let $C_j^n =$

$\prod_{i=j}^{n-1}(f'(x_i))^2$. Then since $v^2(x_j) \to 1$, $C_j^n = C_1^n/C_1^j$, and $C_1^n \to C_1$, we have

$$D_n^2 = \sum_{j=0}^{n-1} v^2(x_j)C_j^n = C_1^n \sum_{j=0}^{n-1} v^2(x_j)/C_1^j \sim \sum_{j=0}^{n-1} v^2(x_j) \sim n. \tag{7}$$

We have now managed to describe the basic properties of the random process of molecule growth in terms of the deterministic sequence $\{x_n\}$. For most purposes this should suffice. However, in order to obtain an explicit description it is enticing to proceed to a differential equations and continuous time description.

Since the basic difference equation is

$$x_{n+1} - x_n = x_n/(1 + x_n),$$

it would seem natural to consider

$$\frac{dx_t}{dt} = \frac{x_t}{1 + x_t}. \tag{8}$$

This can be solved by separation of variables to yield

$$t = \ln x_t + x_t - x_0 - \ln x_0. \tag{9}$$

Write $g(x) = x + \ln x$, and denote its inverse by $G(x)$. Then

$$x_t = G(t + g(x_0)).$$

If $x_0 = 1/K$, we have

$$x_t = G(t - \ln K).$$

In terms of the Wessel (or omega) function W, $G(\ln x) = W(x)$.

Whereas this approximation is accurate for large x, it overestimates the correct x_n while little. For this first period the equation

$$\frac{dx_t}{dt} = \ln 2 \frac{x_t}{1 + x_t}, \tag{10}$$

considered by Schnell and Mendoza, would come closer. However during the ultimate linear part the solution of this equation has too little a slope, $\ln 2$ instead of 1. In conclusion thus it seems better not to replace the difference equation by any differential counterpart.

4. Concluding Remarks

First of all, our discussion shows that the period of exponential growth at or above a certain rate r is determined by

$$e^r \leq \frac{2K + Z_n}{K + Z_n} = \frac{2 + X_n^K}{1 + X_n^K} \approx \frac{2 + x_n}{1 + x_n}.$$

In other words, as long as the dimensionless reduced concentration does not exceed

$$\frac{2 - e^r}{e^r - 1}, \quad 0 < r < \ln 2,$$

growth remains exponential at rate $\geq r$.

Second, the Michaelis-Menten constant can be estimated from the slope of the linear increase function, since $Z_n/n \to K$. It follows that an estimate of $p(z)$ for different z is given by

$$\frac{Z_n}{Z_n + nz}, z \geq 0,$$

where the cycle number n is large enough for the molecule amount to be growing in a linear fashion.

Based on the normal approximation in (6),

$$\frac{Z_n - nK}{\sqrt{nK}} \approx N(0,1). \tag{11}$$

we obtain 95% confidence limits for K based on two standard deviations (by solving a quadratic inequality) as

$$\frac{Z_n + 2}{n} \pm 2\frac{\sqrt{Z_n + 1}}{n}, \tag{12}$$

showing that the estimation error is of a smaller order of magnitude than the estimate.

In summary, provided our model depicts actual quantitative PCR well enough, it paves the way for including data from the relatively stable late phase of linear growth into the analysis of quantitative PCR.

5. References

(1) Chow, Y. S. and Teicher, H. (1997) *Probability Theory*, Third ed. Springer, New York *etc.*

(2) Cornish-Bowden, A. (1979) *Fundamentals of Enzyme Kinetics.* Butterworths, London.

(3) Jacob, C. and Peccoud, J. (1996) Estimation of the offspring mean for a supercritical branching process from migrating binomial observations. *C. R. Acad. Sci. Paris*, Série I, 322, 763-768.

(4) Jacob, C. and Peccoud, J. (1996) Inference on the initial size of a supercritical branching process from migrating binomial observations. *C. R. Acad. Sci. Paris*, Série I, 322, 875-880.

(5) Jacob, C. and Peccoud, J. (1998) Estimation of the parameters of a branching process from migrating binomial observations. *Adv. Appl. Prob.* 30 948-967.

(6) Jagers, P. (1975) *Branching Processes with Biological Applications.* Wiley, Chichester and New York.

(7) Klebaner, F.C. (1984) On population-size-dependent branching processes. *Adv. Appl. Prob.* 16, 30-55

(8) Klebaner, F.C. (1993) Population-dependent branching processes with a threshold. *Stoch. Proc. Appl.* 46, 115-127.

(9) Klebaner, F.C. (1997) Population and density dependent branching processes. In *Classical and Modern Branching Processes*,

Athreya K.B. and Jagers P. eds. Mathematics and Its Applications, IMA 84, 165-170.

(10) Krawczak, M., Reiss, J., Schmidtke, J., and Rösler, U., (1989) Polymerase chain reaction: replication errors and reliability of gene diagnosis. *Nucl. Acids Res.* 17, 2197-2201.

(11) Nedelman, J., Heagert, P., and Lawrence, C. (1992) Quantitative PCR: procedures and precisions. *Bull. Math. Biol.* 54, 477-502.

(12) Olofsson, P. and Shaw, C. (2002) Exact sampling formulas for multi-type Galton-Watson processes. *J. Math Biol.* 45, 279-293.

(13) Peccoud, J. and Jacob, C. (1996) Theoretical uncertainty of measurements using quantitativ polymerase chain reaction. *Biophys. J.* 71, 101-108.

(14) Peccoud, J. and Jacob, C. (1998) Statistical estimation of PCR amplification rates. In: Ferré, F. (ed.), *Gene Quantification.* Birkhäuser, Boston.

(15) Raeymaekers, L. (2000) Basic principles of quantitative PCR. *Mol. Biotechnol.* 15, 115-122.

(16) Saiki, R., Scharf, S.,Faloona, F., Mullis, K., Horn, G.T., Ehrlich, H.A., and Arnheim, M. (1985) Enzymatic amplification of β-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* 239, 487-491.

(17) Schnell, S. and Mendoza C. (1997a) Enzymological considerations for a theoretical description of the quantitative competitive polymerase chain reaction (QC-PCR). *J. Theor. Biol.* 184, 433-440.

(18) Schnell, S. and Mendoza C. (1997b) Theoretical description of the polymerase chain reaction (QC-PCR). *J. Theor. Biol.* 184, 433-440

(19) Sun, F. (1995) The polymerase chain reaction and branching processes. *J. Comp. Bio.* 2, 63-86.

(20) Watkins, J. (2000) Consistency and fluctuations theorems for discrete time structured population models having demographic stochasticity. *J. Math. Biol.* 41, 253-271.

(21) Weiss, G. and von Haeseler, A. (1997) A coalescent approach to the polymerase chain reaction. *Nucleic Acids Research* 25, 3082-3087.

School of Mathematical Sciences, Chalmers University of Technology, Gothenburg, Sweden

School of Mathematical Sciences, Monash University, Clayton, Victoria 3800, Australia

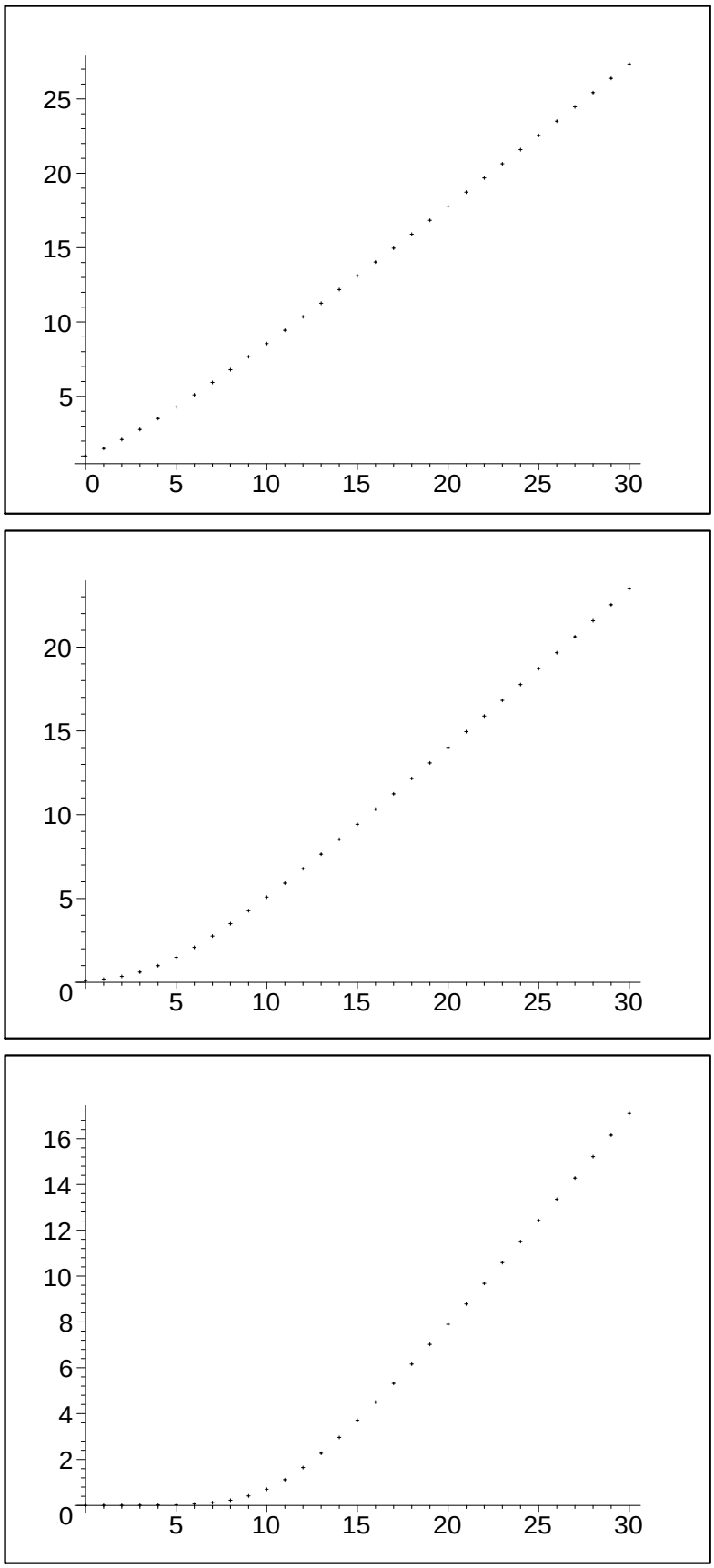

FIGURE 1. y-coordinate: n-th iterate of f, x-coordinate: iteration number n, for $x_0 = 1, 0.1, 0.001$

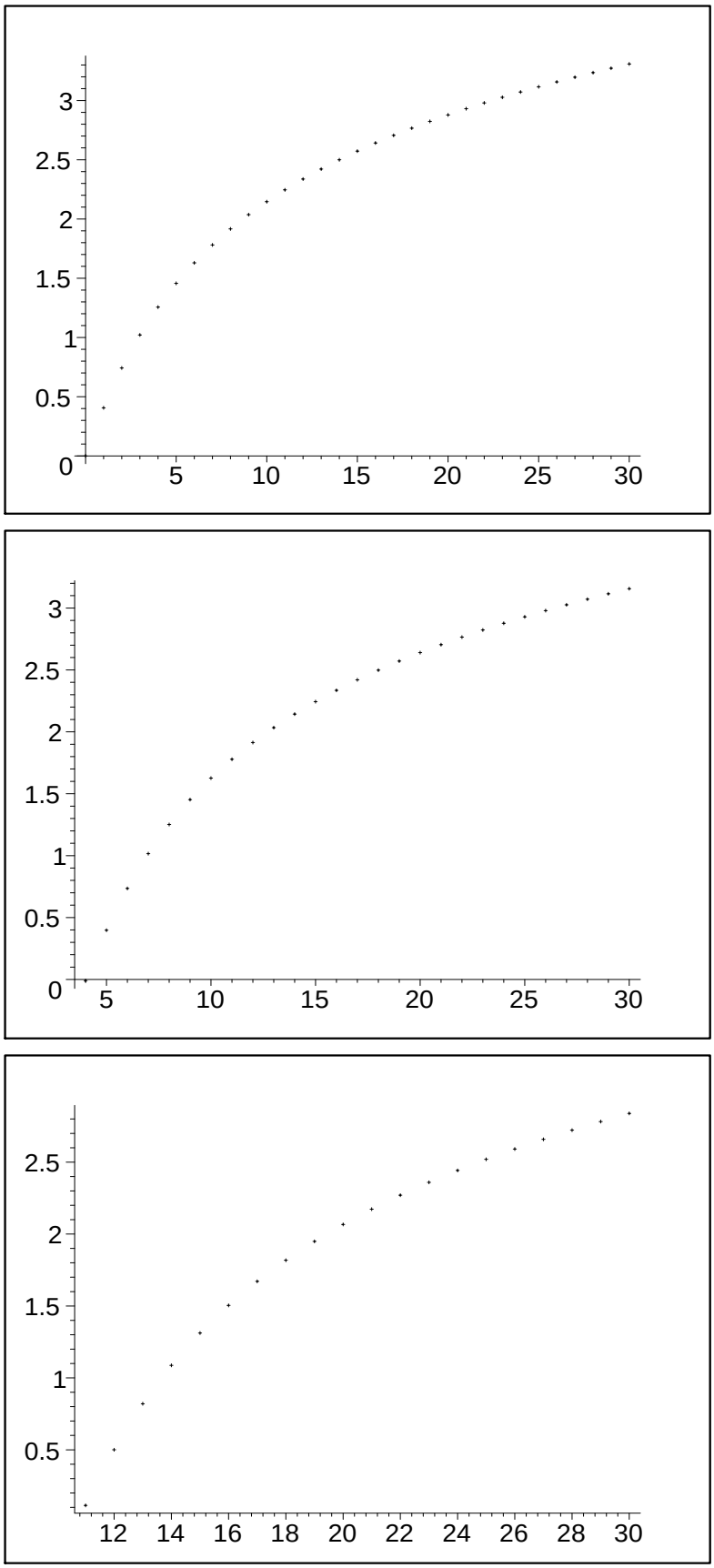

FIGURE 2. y-coordinate: logarithm of n-th iterate of f, x-coordinate: iteration number n, for $x_0 = 1, 0.1, 0.001$