Chapter 19

Fitting and Statistical Analysis of Single-Channel Records

DAVID COLQUHOUN and F. J. SIGWORTH

1. Introduction

The aims of analysis of single-channel records can be considered in two categories. The first is to allow one to observe results at leisure in order to determine their qualitative features. It may, for example, be found that the single-channel currents were not all of the same amplitude or that they showed obvious grouping into bursts or that artifacts appeared on the record that might be misleading. These effects are often not easy to see on the oscilloscope screen as an experiment proceeds. It is best to have a computer program that allows one, after the experiment, to scroll flexibly through the recorded data and zoom in on portions of the record to observe details at high time resolution.

The second aim is to perform quantitative analyses of measurable variables (e.g., the channel-open durations), in which these quantities are compared with theoretical distributions, and to try to infer a biological mechanism from the result. Although other measurable variables can be studied, in this chapter we consider only the analysis of channel current amplitudes and dwell times. The current through a single channel is assumed to consist of rectangular pulses having one or a few discrete current levels and infinitely short transition times. The analysis procedures we describe involve, first, the estimation of the amplitudes and times of transition in the measured currents and, second, the fitting of distributions to these estimates.

It is undoubtedly true that one of the disadvantages of recording from single ion channels is the length of time that it takes to analyze the results. One reason for this is that the quantities we measure, for example, the length of time for which a channel stays open, are random variables (as discussed in Chapter 18, this volume). In the simplest case of a quantity that has a simple exponential distribution with mean lifetime τ, the standard deviation of an observation should be simply τ (see, for example, Colquhoun, 1971; Chapter 18, this volume). Therefore, the standard deviation of the mean on n observations should be $\tau/\sqrt{n}$. (The usage of the terms standard deviation and standard error is discussed in section 6.7.1.) In order to find the mean lifetime with an accuracy of 10%, it is necessary to measure 100 or so individual lifetimes. In practice, it is advisable to measure many more events than this. The

Note to the reader: At the authors' request this chapter will use the abbreviations ms and μs instead of msec and μsec.

DAVID COLQUHOUN • Department of Pharmacology, University College London, London WC1E 6BT, England. F. J. SIGWORTH • Department of Cellular and Molecular Physiology, Yale University School of Medicine, New Haven, Connecticut 06510.

Single-Channel Recording, Second Edition, edited by Bert Sakmann and Erwin Neher. Plenum Press, New York, 1995.

main reason additional measurements are needed is that one can never be sure in advance of the shape of the distribution. It is very common for the distribution of observations not to be described by a single exponential distribution but by a mixture of two or three or more exponential terms. Indeed, under some circumstances, the distribution need not be described by a mixture of exponentials at all; for example, this is, strictly speaking, the case when the resolution of the observations is limited (see Section 12 of Chapter 18, this volume, and Section 6.11 below). It will rarely be satisfactory to measure fewer than 200 openings, and a few thousand openings will suffice for quite complex distributions *if* the time constants are well separated. For the evaluation of complex models, data sets with millions of events have been acquired and analyzed (e.g., McManus and Magleby, 1988).

2. Acquiring Data

2.1. Pulsed and Continuous Recordings

Some experiments rely on the application of a stimulus to open the channels. A pulse of applied neurotransmitter or a membrane depolarization is given, and the resulting channel currents are measured. In order to obtain a sufficiently large number of events, sometimes hundreds or thousands of pulsed stimuli are presented. Such experiments are best performed using a computer both to control the application of the stimulus and to acquire data directly during an interval (perhaps a few tens or hundreds of milliseconds) surrounding the time of each stimulus. The resulting recorded data then consist of "sweeps" having a precise timing relationship to the stimulus.

In other experiments the activity of channels is observed under steady-state conditions, for example, in the presence of a constant concentration of an agonist or a constant membrane potential. To obtain the maximum information from the experiment the data are best recorded continuously, for example, with an FM tape recorder, on digital audio tape, or with the combination of a PCM adapter and a videotape recorder. The decreasing costs of computer mass storage media (optical disks, digital tape drives) are making it practical to digitize the data and store it directly in the computer. This makes sense, since for analysis the data must be transferred to the computer eventually.

2.2. Filtering the Data

The filtering of the current-monitor signal from a patch-clamp amplifier is both unavoidable and necessary for practical data analysis. The design of the patch-clamp amplifier places a limit on its frequency response (typically up to 100 kHz or so), so that its output signal can be considered a filtered version of the "true" (infinite bandwidth) current signal. Some filtering is also a necessary part of the data-recording process. FM tape recorders use filters to remove the FM carrier frequencies from the output signal. For the analogue-to-digital converters of digital tape recorders and computer data-acquisition systems, the signal must be first be filtered to avoid aliasing; the DAT and PCM systems designed for audio recording typically incorporate sharp-rolloff elliptic filters for this purpose, which strongly attenuate frequency components above 20 kHz. Finally, some filtering is required anyway for data analysis in order to reduce the background noise sufficiently to allow single-channel events to be detected and characterized.

The question of the optimum degree of filtering is discussed below (Section 3.2). The events of interest are rectangular, so it is undesirable to use a filter with a very sharp rolloff, such as a Butterworth or elliptic filter, because this sort of filter distorts a step input to produce an overshoot and "ringing" appearance (although this sort of filter would be appropriate if the single-channel records are to be used for calculation of a noise spectrum). Most commonly, a Bessel filter (four poles or more) is used. On some commercial active filter instruments, this sort of filter characteristic is sometimes referred to as damped mode or low Q. The cutoff frequency labeled on the front panel of the active filter is sometimes the frequency at which the high- and low-frequency asymptotes of the log-attenuation versus log-frequency graph intersect. For a Bessel filter, however, the frequency at which the attenuation is -3 dB is about half of that value. This gives rise to an ambiguity in the specification of filtering that is used. It is desirable that the criterion used always be stated, and it is preferable that the cutoff frequency, f_c, always be specified as the -3 dB frequency, as we do in this chapter.

A useful theoretical model for a general-purpose filter is the Gaussian filter, which has a frequency response function $B(f)$ of the form

$$B(f) = e^{-kf^2} \tag{1}$$

where the constant k is chosen to give 3 dB of attenuation at f_c; i.e., $|B(f_c)|^2 = 1/2$, yielding $k = \ln(2)/2f_c^2$.

Some of the useful properties of the Gaussian filter arise from the fact that the Fourier transform of a Gaussian function is itself a Gaussian function. The inverse transform of equation 1 gives the filter's impulse response, which can be written in the same form as a Gaussian probability distribution:

$$h(t) = \frac{1}{(2\pi)^{1/2}\sigma_g} \exp\left(\frac{-t^2}{2\sigma_g^2}\right) \tag{2}$$

where the width of the impulse response is characterized by σ_g, which is analogous to the standard deviation of a probability distribution. Its value is inversely proportional to f_c

$$\sigma_g = \frac{(\ln 2)^{1/2}}{2\pi f_c} \tag{3}$$

Of special interest for single-channel analysis is the property that the frequency response of two Gaussian filters in cascade is itself Gaussian, with the effective cutoff frequency f_c given by

$$\frac{1}{f_c^2} = \frac{1}{f_1^2} + \frac{1}{f_2^2} \tag{4}$$

where f_1 and f_2 are the cutoff frequencies of the two filters. This property allows repeated filtering to be done on the signal with predictable results. Because Gaussian digital filters are simple to program (see Appendix 3), it is possible to refilter data even after it has been digitized and stored in the computer.

The response characteristic of a Bessel filter is well approximated by the Gaussian response, and the two actually become identical as the number of poles in the Bessel filter becomes large. Equation 4 is therefore useful for estimating the final bandwidth of an entire

recording system. A typical system might consist of a patch clamp with roughly Bessel response, a DAT recorder with sharp-cutoff elliptic filters in the recording and playback paths, and a Bessel filter to reduce the bandwidth before digitization by the computer. The contribution from the patch clamp and Bessel filter can be combined as in equation 4. To a first approximation, the effect of a sharp-cutoff filter can be neglected, provided its cutoff frequency is at least twice the f_c of the rest of the system.* Thus, for example, a system with a 10-kHz Bessel filter in the patch clamp cascaded with a 5-kHz Bessel filter yields an effective bandwidth of 4.47 kHz; in this situation the presence of a DAT recorder with its sharp-cutoff 20-kHz filter would have essentially no effect on the final response.

For theoretical work, the Gaussian filter is convenient because its impulse response and step response are relatively simple functions of time; the results in Sections 3 and 4 of this chapter have been computed for a Gaussian response for this reason. Some properties of the Gaussian filter can be summarized as follows.

2.2.1. Properties of the Gaussian Filter

The frequency response function of the Gaussian filter is given by equation 1 or, numerically,

$$B(f) = \exp[-0.3466(f/f_c)^2] \tag{5}$$

The impulse response function (equation 2) can be written in terms of the cutoff frequency f_c as

$$h(t) = 3.011\, f_c \exp[-(5.336\, f_c t)^2] \tag{6}$$

The step response is

$$\begin{aligned} H(t) &= \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{t}{2^{1/2}\sigma_g}\right)\right] \\ &= \frac{1}{2}[1 + \operatorname{erf}(5.336\, f_c t)] \end{aligned} \tag{7}$$

In modeling the response to single-channel current pulses, it is useful to know the peak output of the filter in response to a rectangular pulse of length w and unit amplitude, which is

$$y_{\max} = \operatorname{erf}\left(\frac{w}{2^{3/2}\sigma_g}\right) = \operatorname{erf}(2.668\, f_c w) \tag{8}$$

*For Gaussian filters, each term in equation 4 is proportional to the second moment of the impulse response. Thus, the equation follows from the fact that when two functions are convolved, their second moments add. For sharp-cutoff filters, the second moment is approximately zero; indeed, for Butterworth filters, it is exactly zero.

The total noise variance of the output from the Gaussian filter when the input has a (one-sided) spectral density $S(f) = S_0 (1 + f/f_1 + f^2/f_2^2)$ is given by

$$\sigma_n^2 = \int_0^\infty |B(f)^2| S(f) df$$
$$= S_0[a_0 f_c + (a_1/f_1) f_c^2 + (a_2/f_2^2) f_c^3] \quad (9)$$

where $a_0 = 1.0645$, $a_1 = 0.7214$, and $a_2 = 0.7679$.

2.2.2. Risetime of the Filter

A particularly useful descriptive parameter for a filter is the risetime, T_r. Roughly speaking, T_r is the time for the output of a filter to make a transition when a square step is applied to the input. It therefore corresponds to the minimum length of a pulse to which the filter gives a nearly full-amplitude response. One commonly used definition for the risetime is the time between the 10% and 90% amplitude points of the transition in the output of the filter,

$$T_{10-90} = 2^{3/2} \sigma_g \text{erf}^{-1}(0.8)$$
$$= 0.3396/f_c \quad (10)$$

The definition we use here sets T_r equal to the reciprocal of the slope at the midpoint of the response $H(t)$ to a unit step input,

$$T_r = \left[\frac{dH(t)}{dt}\right]_{t=0}^{-1} \quad (11)$$

which is given by

$$T_r = (2\pi)^{1/2} \sigma_g$$
$$= 0.3321/f_c \quad (12)$$

For a Gaussian filter the two definitions of risetime give essentially identical values. T_r is inversely proportional to f_c and a 1-kHz Bessel or Gaussian filter has a risetime of about 330 μsec. It is often convenient to use T_r rather than f_c to specify the amount of filtering (e.g., one can say that "openings longer than 2 T_r were fitted").

2.3. Digitizing the Data

The data are always acquired, in the first place, in the form of a voltage (analog) signal; they are then converted to digital form for storage on digital tape (DAT or PCM/videotape), or for computer analysis, by an analog-to-digital converter (ADC). The ADC necessarily samples the voltage at discrete times; if the sample rate is too low, information about rapid voltage changes is lost. This loss of information can be described as frequency aliasing,

in which high-frequency components of the original signal become converted to lower-frequency ones.

A good criterion for the choice of the sampling frequency is to require that the digitized record, when interpolated by some convenient means, is indistinguishable from the original continuous record. Sampling at the Nyquist rate (i.e., at twice the filter cutoff frequency) is a special case of this criterion, but for our purposes, the Nyquist criterion requires two unreasonable assumptions. First, it requires that the original signal contain no frequency components above a given frequency f_0 to avoid aliasing. This is unreasonable because no practical filter can accomplish this entirely, and Bessel filters are particularly bad in this respect because of their gradual rolloff characteristic. Second, the samples (digitized at the Nyquist rate of $2 f_0$) must be interpolated using a very slowly decaying function of the form $\sin(xt)/xt$ in order to reconstruct the original signal properly. This sort of interpolation requires much computation and is not suitable for short records.

Interpolation is important when the original signal is sampled relatively sparsely; it allows one to reconstruct the record to any degree of smoothness for viewing while using a minimum of computer storage for the digitized data. Proper interpolation also reduces errors in certain transition-fitting procedures (see Section 4.1.2). When a cubic spline function is used to interpolate the points, a practical minimum sampling rate for Bessel-filtered data is about five times the -3 dB frequency of the filter, in which case the peak error in the reconstruction is about 2% (Fig. 1). In the cubic spline, cubic polynomials form the interpolation between every two points, with the second derivative being continuous throughout. The

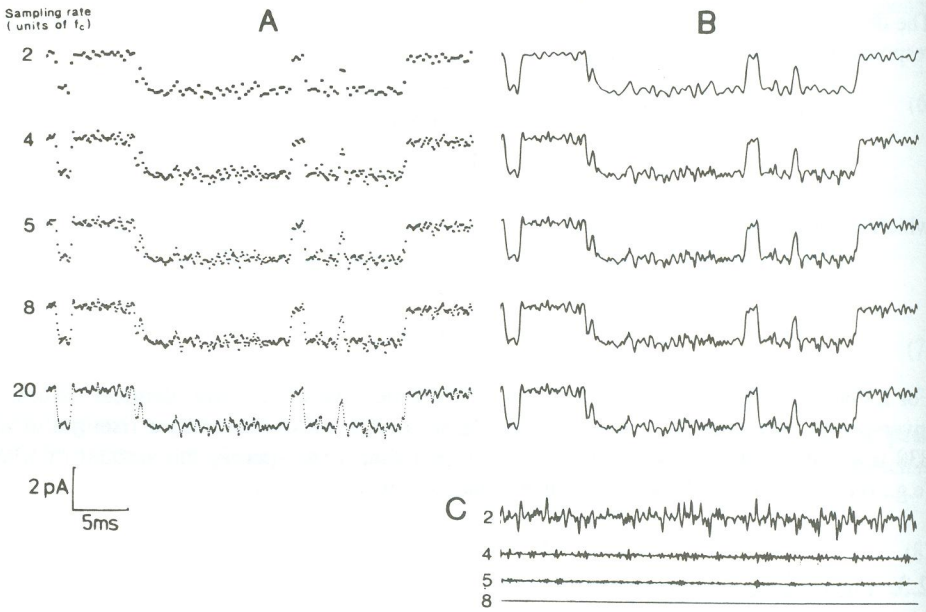


Figure 1. A single-channel current record sampled at various rates. Inward currents through ACh-receptor channels in a rat myoball were recorded cell-attached at 22°C with $V_m = -45$ mV and filtered at $f_c = 2$ kHz with a four-pole Bessel filter. A: Data points as sampled at 2, 4, 5, 8, and 20 times f_c. B: Result of cubic spline interpolation of the sampled data. C: Error traces, computed as the difference between the interpolated traces and the original data sampled at 20 f_c and scaled up by a factor of 4. The single-channel current was -1.5 pA in this recording, and the rms background noise level $\sigma_n = 0.15$ pA.

width of the interpolating function is quite narrow, so that "edge effects" (errors caused by the lack of surrounding data points) persist only about four points in from each edge of a record. A subroutine for spline interpolation is described in Appendix 3. Other interpolation techniques, including simple linear interpolation, can also be used but may require higher sample rates. If interpolation is not used, data sampled at the minimum rate appear sparse and are hard to evaluate by eye; higher rates, such as 10 to 20 times the filter's −3 dB frequency, are needed.

In general, it is best to digitize the entire experimental record. This is a good because it allows all of the data to be inspected directly and because it allows all dwell times, including the longest ones, to be measured directly. Sampling at a rate of 40 kHz (appropriate for a 2 to 4-kHz filter if interpolation is not used) generates 4.8 Mb of data per minute, assuming that data are stored as two-byte integers; thus, only a limited amount of data can be stored in computer memory. In order to digitize a long continuous record without gaps, the computer must have the ability to acquire samples into memory while simultaneously writing the data from memory to disk. This can be done by means of a separate memory buffer incorporated into the ADC system or by using direct memory access (DMA) transfer of data. For high sample rates (say 50 kHz or faster), attention must also be given to the speed at which data can be written to the storage device.

An example of a high-speed continuous acquisition program is the VCatch program for Macintosh computers. It acquires digital samples at a 94-kHz rate directly from the playback of a videotape recording using the VR-10 PCM adapter (Instrutech Corp, Mineola NY) or at sample rates up to 200 kHz using the ITC-16 ADC interface (Instrutech). In each case, the interface hardware includes an internal sample buffer (16k or 32k words of first-in/first-out buffer) that is emptied at regular intervals into a 1 Mb circular buffer in the computer's memory by an asynchronous "timer task" running on the host computer. The main program displays the incoming data and writes blocks of data from this buffer to a large-capacity hard disk. A similar facility is provided by the CED 1401-plus interface (Cambridge Electronic Design, Cambridge, U.K.) for IBM-compatible computers. It uses DMA to transfer ADC samples directly to a 64-kb circular buffer in the computer's memory, allowing analogue voltages to be digitized at rates up to 80 kHz while writing the data continuously to the hard disk. Some commercial interfaces allow continuous sampling and writing to disk only at lower rates than these, e.g., up to 30 kHz. For high-resolution data, this sampling rate may not be sufficient; however, if the original data recording is on FM tape, it is sometimes possible to slow down the tape speed while sampling the data to increase the effective sample rate.

An alternative to digitizing the entire record is to have some sort of automatic detection of the points at which opening transitions occur, and to digitize only the sections that contain openings. In this approach it is necessary that the detection method keep a record of the time intervals between openings, so that the distribution of shut periods can be constructed. This approach is satisfactory only to the extent that the detection system is reliable and the detection parameters have been properly set up before the recording starts. However, the availability of high-capacity disk drives that can store an entire recording makes this approach less attractive than it was in the past.

3. Finding Channel Events

The analysis of single-channel records first involves estimating the time and the amplitude of each transition in the current record. The list of these values is described as an

idealized record that approximates the true channel activity and serves as the data set for statistical analysis of the kinetics. In practice, some of the original transitions are missed in the analysis process. To a certain extent, corrections can be made for missing events (see Section 12 of Chapter 18, this volume; Section 6.11 below), but it is important that the idealized record be as complete and unbiased as possible, especially when multistate kinetics are involved.

Finding events and fitting the transitions are considered separately in this section and the next because the two operations are often carried out separately. For example, a simple transition finder can rapidly scan a digitized record for putative channel activity. Once each event is found, it can then be fitted to an idealized time course by a much more time-consuming fitting routine, which may even require the record to be filtered differently. On the other hand, event detection and characterization can be combined in the use of a simple threshold detector, which provides a simple but useful estimator of transition times for event characterization.

3.1. Description of the Problem

The basic problem in identifying channel activity in an experimental record is that short channel openings are indistinguishable from random noise fluctuations about the baseline; similarly, short gaps are indistinguishable from fluctuations away from the open-channel current level. This is because, as a result of filtering, narrow current pulses as well as random noise fluctuations take on roughly the same time course as the recording system's impulse response. Determining whether a particular blip is a channel opening can therefore be done only statistically. In order to estimate the reliability and the limits of detection, we consider a model situation and apply some classical results from communication theory to the problem.

We assume that the channel activity to be detected consists of widely spaced rectangular current pulses of random duration but fixed amplitude A_0. The baseline level is zero. The background noise has a spectral density $S_n(f)$ and is assumed to be Gaussian distributed and independent of the channel activity. (These last two conditions appear to hold in high-quality patch recordings.) The completely unfiltered current signal $x(t)$ (if it could be observed) is represented as the sum of noiseless channel activity $s(t)$ and a noise function $n(t)$, as illustrated in Fig. 2.

The detection strategy is the following: at each time point t we form a linear combination $y(t)$ of signal values according to

$$y(t) = \int_{-\infty}^{\infty} h(t - \tau)x(\tau)d\tau \tag{13}$$

where h is a normalized weighting function that determines, in effect, the amount of time

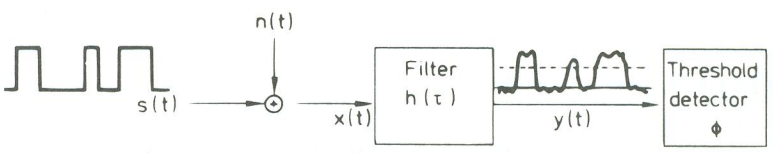


Figure 2. Model of single-channel event detection

averaging that is done in forming y. The value of y is then compared with a threshold ϕ; if $y > \phi$ at some time t, channel activity is said to be detected at t.

This detection scheme is general in the sense that it includes all possible linear signal-processing operations in the specification of the function h. It is also an optimum detection scheme in the sense that, for a signal consisting of pulses of defined shape and size, it can yield the lowest probability of error in detecting these pulses (VanTrees, 1968). We do not know, however, whether it is the optimum scheme for detecting pulses having random widths, as are actually encountered in single-channel records.

The operation described by equation 13 is a filtering operation; in fact, the function $y(t)$ is just what one obtains as the output from a filter with impulse response $h(t)$. Thus, we can represent a linear detection scheme of this kind simply as a filter followed by a threshold detector, as shown in Fig. 2. The filter in this diagram actually represents the transfer function of the entire recording system, including the characteristics of the pipette, patch-clamp amplifier, analog filter, and any computations that are performed on the digital samples. One step in event detection is often performed by a computer program in which y is computed as a weighted sum over discrete sample values rather than as an integral. This is equivalent to operating on the signal by a digital filter, which in turn is equivalent to continuous-time filtering, by the sampling theorem. Regardless of how the filtering is performed, the problem of determining the best way to detect events is reduced to finding a suitable value for the threshold ϕ and a suitable response characteristic for the filter.

3.2. Choosing the Filter Characteristics

3.2.1. Signal-to-Noise Ratio

The filter's cutoff frequency f_c and the form of the filter's frequency response characteristic can be varied to optimize the probability of detection of channel events. One strategy for doing this is to maximize the signal-to-noise ratio (SNR) for the response to a pulse of a given width, w, in the presence of noise. If SNR is defined to be the ratio of the peak amplitude y_{max} of the filtered pulse to the standard deviation of the filtered noise, it can be expressed in terms of the filter transfer function, $B(f)$, and the noise spectrum, $S_n(f)$, as

$$\text{SNR} = \frac{y_{max}}{\sigma_n} = \frac{\left|\int_{-\infty}^{\infty} B(f)X(f)df\right|}{\left[\int_{-\infty}^{\infty} |B(f)|^2 S_n(f)df\right]^{1/2}} \tag{14}$$

where $X(f)$ is the complex Fourier transform of the original pulse shape. We will see that the choice of the best filter setting depends strongly on the form of S_n. The background noise in the patch clamp should theoretically show flat spectral density at low frequencies (below about 1 kHz) and rise asymptotically as f^2 at high frequencies (see Chapter 4, this volume). In the frequency range between 1 kHz and 10 kHz, the spectral density typically is seen to rise roughly proportionally to f.

Two useful models for background noise are, therefore, the so-called "1 + f" spectrum, having the form

$$S_n = S_0\left(1 + \frac{|f|}{f_0}\right)$$

and "1 + f^2" noise,

$$S_n = S_0\left(1 + \frac{f^2}{f_0^2}\right)$$

In each case, f_0 is a characteristic "corner" frequency. In order to give numerical values for the results of calculations, we adopt a standard background noise spectrum of the $1 + f$ form with the (one-sided) spectral density $S_0 = 10^{-30}$ A^2/Hz and with $f_0 = 1$ kHz. This is a noise level that can be obtained with present-day amplifiers and pipette technology when some care is exercised.

If we assume a tunable filter with a variable cutoff frequency, f_c, of the form $B(f) = B_0(f/f_c)$, then we can calculate the dependence of σ_n on f_c by evaluating the denominator of equation 14. In the case that $S_n(f)$ is proportional to f_c^a for some exponent a, σ_n will be proportional to $f_c^{(a+1)/2}$.

In the case of a Gaussian filter response, σ_n can be computed directly from equation 9. The dependence of σ_n on f_c for various spectral types (flat, $1 + f$, and $1 + f^2$) is illustrated by the lower curves in Fig. 3.

The numerator of equation 14 is the peak value y_{max} of the filtered pulse. For a rectangular pulse of fixed width, y_{max} is small and proportional to f_c for low f_c values (heavy filtering). For a pulse of amplitude A_0 and width w, the size of the response is related to the filter risetime,

$$y_{max} \approx A_0 \frac{w}{T_r} \qquad w << T_r \tag{15}$$

As f_c is increased, T_r decreases, and y_{max} approaches the original pulse height when $w \geq T_r$. This last condition corresponds to filter bandwidths at which the rectangular shape of the original pulse can be resolved. The relation between y_{max} and f_c is shown by the upper curve in Fig. 3.

The choice of the optimum f_c for the three spectral types is indicated by the dashed lines in Fig. 3. In the case of a flat spectrum, the largest SNR value is obtained for a relatively high value of f_c because σ_n grows only as $f_c^{1/2}$, whereas y_{max} rises more quickly at low f_c values. For $S_n(f)$ rising proportionally to f, the choice of f_c is relatively uncritical, since σ_n and y_{max} rise in parallel. Finally, for $S_n(f)$ rising as f^2, f_c is best chosen to be small, since σ_n is rising relatively steeply, as $f_c^{3/2}$. Figure 3 presents an extreme case in which the pulse width w was chosen to be small (10 μs) compared with the time scale of the corner frequency f_0. As a result, the optimum f_c values differ widely. For longer pulses, the spread in optimal f_c values would be less.

3.2.2. Matched Filter

The exact form of the filter response that maximizes the SNR for a given noise spectrum and pulse shape is the so-called matched filter, which has the transfer function (see, for example, Van Trees, 1968)

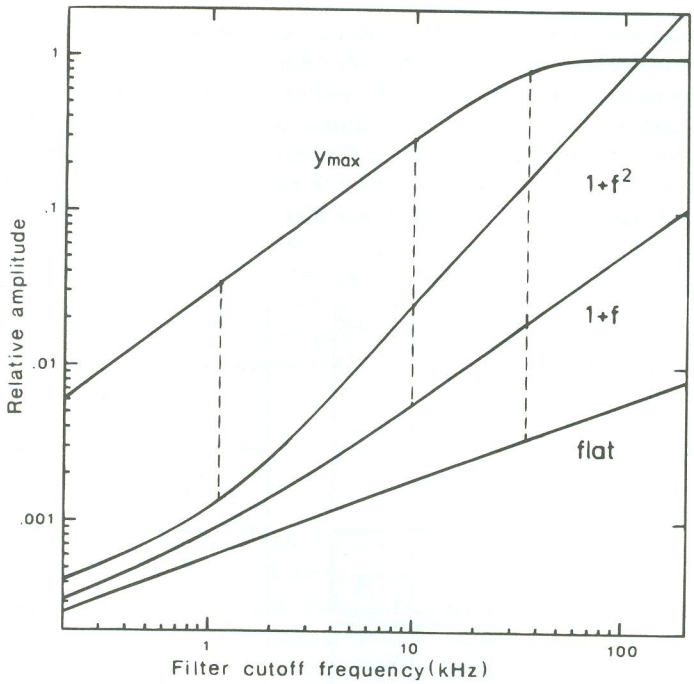


Figure 3. Effect of filter cutoff frequency f_c on signal and noise amplitudes. The upper curve shows the peak amplitude y_{max} of the response of a Gaussian filter to a 10-μs pulse of unit amplitude. Below about 40 kHz, the pulse is appreciably attenuated by the filter. The lower curves show the dependence of the rms noise amplitude σ_n on f_c assuming flat, $1 + f$, and $1 + f^2$ spectral characteristics. (The noise corner frequency was $f_o = 1$ kHz in each case, and S_o values were chosen arbitrarily.) The dashed lines indicate the points of widest separation between y_{max} and σ_n, i.e., the highest signal-to-noise ratios. The f_c values giving the best SNR were 36, 10, and 2 kHz for the three spectral types. In $1 + f^2$ noise, the optimally fitted pulse would be attenuated to only 6% of its original amplitude. The absolute value of σ_n for the "standard" noise spectrum ($S_o = 10^{-30}$ A²/Hz) can be read directly from the $1 + f$ noise curve if the relative amplitude values are multiplied by 50 pA.

$$B(f) = c\,\frac{X^*(f)}{S_n(f)} \tag{16}$$

where X^* is the complex conjugate of X, and c is an arbitrary gain factor. [The transfer function can be multiplied by an arbitrary delay factor of the form $\exp(-j\,2\pi\,ft_0)$, but we ignore this.] In the case of a flat noise spectrum, the matched filter's impulse response is a time-reversed copy of the matching signal—in our case, a pulse of width w; the filter is then just a running averager, averaging over a time w. If instead the noise spectrum is not flat, the matched filter has a different form.

It should be noted that the matched filter does not necessarily preserve the shape of the original pulse, since it is optimized only for the peak of the response. In the flat-spectrum case, for example, the response to the matched rectangular pulse is a triangular pulse.

3.2.3. Gaussian Filter

Although matched digital filters are not difficult to program, analog matched filters are difficult to make. Besides, one would prefer to have a general-purpose filter with only one

adjustable parameter, say, the cutoff frequency, as opposed to one with the complicated adjustments implied by equation 16. As was mentioned in Section 2.2, the Gaussian filter has various appropriate properties for single-channel analysis. Surprisingly, this filter also gives SNR values nearly as large as those from a matched filter. Figures 4A and D compare SNR values for the matched filter and the Gaussian filter as a function of the pulse width w, assuming noise spectral densities of the $1 + f$ and $1 + f^2$ types, respectively. The SNR values for the Gaussian filter were never less than 0.84 times the matched-filter values and

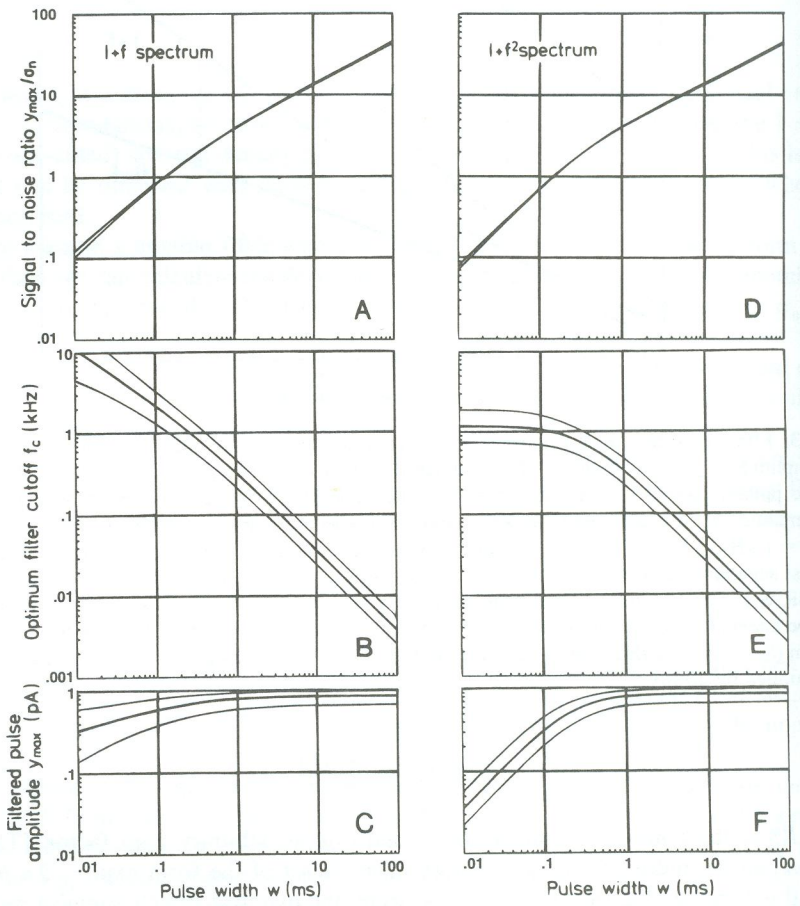


Figure 4. Filtering for optimum signal-to-noise ratios in the presence of background noise with $1 + f$ and $1 + f^2$ spectra. A and D: Ratio of peak signal, y_{max}, to rms noise, σ_n, for the matched filter (heavy curve) and the optimally tuned Gaussian filter (thin curve) as a function of the matched pulse width w. B and E: Gaussian filter cutoff frequency f_c yielding the SNR values plotted above. The choice of f_c is not extremely critical, as indicated by the thin curves, which denote the range of f_c curves giving at least 90% of the maximum SNR. C and F: The corresponding peak signal amplitudes after Gaussian filtering. The thin curves show the range of amplitudes resulting from the range of f_c values in B and E. The noise spectral densities were taken to be one-sided, $S_n = S_o(1 + f/f_o)$ and $S_n = S_o[1 + (f/f_o)^2]$, with $S_o = 10^{-30}$ A^2/Hz in each case, and the pulse amplitude $A_o = 1$ pA. The SNR, f_c, and y_{max} values from these curves can be scaled for other values of S_o, f_o, and A_o by forming the ratios $S = S_o/10^{-30}$ A^2 s, $f = f_o/1$ kHz, and $\hat{A} = A_o/1$ pA. The resulting values SNR′, f_o', and y_o', are given by SNR′ $= [\hat{A}/(Sf)^{1/2}]$SNR(wf); $f_c' = f f_c(wf)$; and $y_{max}' = \hat{A}\, y_{max}(wf)$.

$\mathbf{r}_1$, $\mathbf{r}_0$, and $\mathbf{r}_2$ are not exactly parallel. Figure 7.7 shows the geometry. We use the law of cosines to write [remember that $\cos(\pi-\theta)=-\cos\theta$]

$$r_1=r_0[1+(2x_1/r_0)\cos\theta+x_1^2/r_0^2]^{1/2},$$
$$r_2=r_0[1-(2x_2/r_0)\cos\theta+x_2^2/r_0^2]^{1/2}.$$

When these are inserted in Eq. (7.4) and a Taylor's-series expansion is done to second order in both x_1/r_0 and x_2/r_0, the result is

$$v=\frac{2\pi a^2}{4\pi r^3}\frac{\sigma_i}{\sigma_o}\frac{\Delta v_i(x_1+x_2)}{2}\frac{3\cos^2\theta-1}{2}. \quad (7.16)$$

The constants have been arranged to show that the term $\Delta v_i(x_1+x_2)/2$ is the area under the impulse when v is plotted as a function of distance along the cell. The angular factor as written with its factor of 2 in the denominator is tabulated in many places as the "Legendre polynomial $P_2(\cos\theta)$." The exterior potential now falls off more rapidly with distance, as $1/r^3$. The angular dependence, shown in Fig. 7.8, is symmetric about $\pi/2$. This shows the angular dependence as one moves around the impulse at a constant distance from it. This is a very different situation and a very different curve from the potential measured at a fixed point near the cell as an impulse travels past. In the latter case r as well as θ is changing. This behavior is discussed in Problems 7.6 and 7.7. The results are shown in Fig. 7.9. The potential from the depolarization is biphasic; that from the complete pulse is triphasic, being positive, then negative, then positive again.

For a single axon in an ionic solution the exterior conductivity is usually higher than in the cell, so $\sigma_i/\sigma_o=0.2$. The conductivity of tissue is considerably less than the conductivity of an ionic solution, and the ratio becomes greater

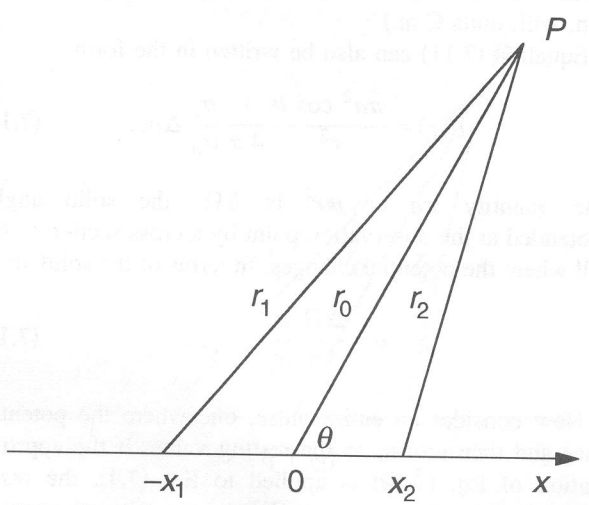


FIGURE 7.7. When the observation point is not so far away, or when a complete nerve impulse is being considered, the law of cosines must be used to relate r_1 and r_2 to r_0.

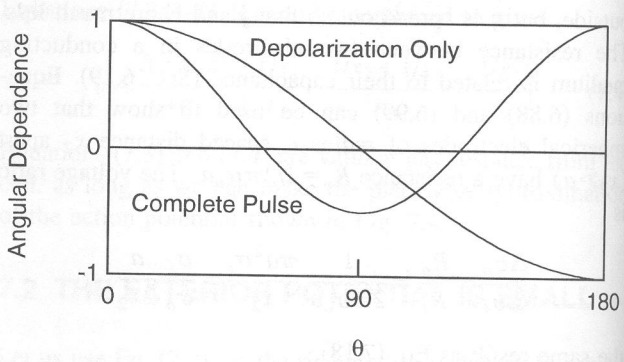


FIGURE 7.8. Plot of the angular dependence of the potential from the entire impulse, Eq. (7.16).

than one. For the electrocardiogram it will be more appropriate to use $\sigma_o=0.33$ S m^{-1} (muscle) or 0.08 S m^{-1} (lung), in which case σ_i/σ_o is 6 or 25. We will use an approximate value of 10.

7.4 THE EXTERIOR POTENTIAL FOR AN ARBITRARY PULSE

We have derived the results of the previous sections for an action potential that varies linearly during depolarization and repolarization, a piecewise-linear approximation. In general the action potential does not have sharp changes in slope. We will now consider the general case and find that the results are very similar. For depolarization alone, we will again have a potential depending on the dipole moment. For a complete pulse the potential will depend on the area under the pulse curve.

Again, the axon is stretched along the x axis in an infinite, homogeneous conducting medium. Consider a small segment of axon between x and $x+dx$. If the current entering this segment at x is greater than the current leaving at $x+dx$, the difference must flow into the external medium. From Eq. (6.45b),

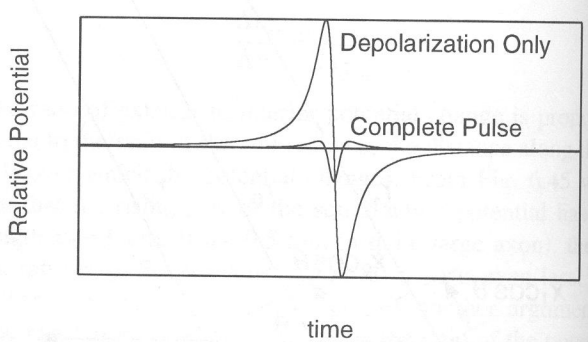


FIGURE 7.9. The potential far from the axon as a function of time as an impulse travels from left to right along the axis. The potential from the complete pulse has been multiplied by a factor of 10 in order to show it.

spectrum is flat, whereas $k = 1.25$ for S_n proportional to f^2; practical recording situations correspond to intermediate values.

The function in equation 17 is plotted in Fig. 5 assuming $f_c = 1$ kHz. The false event rate is seen to be a very steep function of the ratio ϕ/σ_n decreasing from about 10 events/s at $\phi/\sigma_n = 3$ to 0.004 events/s at $\phi/\sigma_n = 5$. What constitutes an acceptable value of λ_f depends on the frequency of true events. For detecting relatively rare channel openings, λ_f should be at least one or two orders of magnitude smaller than the opening rate, which implies a ϕ/σ_n ratio of perhaps 5 or more. On the other hand, in the case that a burst of channel openings has been found, the problem might then be to find all channel-closing events. Since the true events in this case would be much more frequent, λ_f could be larger, and ϕ/σ_n might be chosen to be 3, for example. It is a good idea to be conservative and choose a somewhat larger value for ϕ/σ_n than that given by equation 17 or Fig. 5 to allow for possible errors in the estimation of the baseline level or small changes in the noise level, which could have a large effect on the false-event rate.

The threshold must also be chosen low enough that the desired events will be detected. One strategy for choosing ϕ would be to optimize the detection of the shortest possible events. Let w_{min} be the minimum detectable event width, and y_{max} the peak amplitude of a filtered pulse of this width. If we set $\phi = y_{max}$, approximately half of all such events will be detected, since noise fluctuations will cause some events to cross the threshold and others

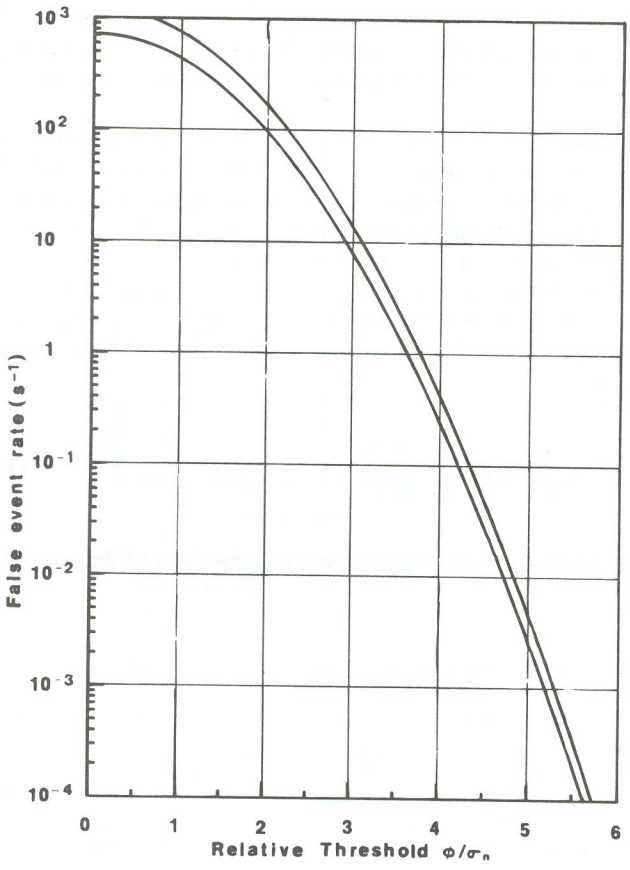


Figure 5. False-event rate, λ_f, as a function of the threshold-to-rms-noise ratio. The curves were calculated according to equation 17 with $f_c = 1$ kHz for the case of $1 + f^2$ (upper curve) and flat spectral densities of background noise. False-event rates corresponding to practical background noise spectra are expected to fall between the curves. Note that λ_f should be scaled proportionally to f_c for other f_c values.

to remain below it. To determine the value of w_{min}, we can use the signal-to-noise ratio curves of Fig. 4. The SNR in this case is just equal to the desired ϕ/σ_n ratio. Given this, the values for w_{min} and f_c can be read from the curves. Unfortunately, this procedure requires that parameters S_0 and f_0 of the noise spectrum be known in order to scale properly the results from Fig. 4.

A simpler approach is suggested by the fact that for the $1 + f$ spectrum, y_{max} varies only weakly with w (Fig. 4C), and for each w, a considerable range of y_{max} values can result in nearly maximum SNR values. Thus, one could pick ϕ equal to a reasonable y_{max} value and then tune the filter while measuring σ_n to give the desired ϕ/σ_n ratio. What is a reasonable y_{max} value? This issue is discussed in Appendix 1; in summary, a good choice of ϕ is 0.7 A_0 in the case of small-amplitude events, which will require heavy filtering ($f_c \leq f_0$), or 0.5 A_0 for larger-amplitude events for which a wider filter bandwidth will be used. This choice of $\phi = 0.5\ A_0$ is of practical interest because it allows simple event characterization as well, as described in Section 4.1.

3.4. Practical Event Detection

3.4.1. Optimal Threshold Detection

A general procedure for setting up the filter and threshold detector can now be summarized as follows: (1) given the channel amplitude A_0, pick a threshold level ϕ, e.g., in the range 0.4 to 0.7 times A_0; (2) adjust the filter's corner frequency to bring the rms noise, σ_n, to the desired fraction, e.g., one-fifth, of ϕ (3) optionally, ϕ can be readjusted slightly in view of the relationship between f_c and the frequency of the corner of the noise spectrum.

In typical patch recordings the background noise spectrum has, up to now, commonly been of the $1 + f$ form, for which the above strategies are appropriate. The final asymptote of the noise spectral density is, however, proportional to f^2, and it is likely that as techniques improve and extraneous noise sources are eliminated, the background noise in practical recordings will more nearly approach this asymptote. Once the noise density is seen to rise more steeply than linearly with frequency, a different strategy for choosing the threshold and filter frequency should be used. Recall that in this case the SNR is not improved when f_c increases beyond a critical value (Fig. 4E); therefore, it would be best in the case of large events to set the filter first to the critical frequency, about 1.2 times f_0. Then, the threshold level can be chosen to be the proper multiple of σ_n to achieve an acceptably low false-event rate.

Some convenient means for measuring σ_n is clearly required in order to set up the filter and threshold in the ways just described. A "true rms" voltmeter can be used to read σ_n directly, provided that sufficiently long event-free stretches are available for the measurement to be made. If the record is digitized, a segment can first be checked visually for the absence of obvious events. A calculation of the standard deviation of all the points in the segment then yields σ_n. A fairly long segment (or collection of segments) is needed for a precise estimate; 1000 points yields a standard deviation for σ of roughly 5%, depending on the spectral type and the relative sampling rate. For example, if the sampling rate is higher than $5f_c$, more points will be required because of the increased correlation between adjacent samples.

Throughout this section, we have assumed that the baseline level is zero. Since in experimental records the baseline current level is nonzero and typically shows a slow drift with time, any event-finding procedure needs to compensate for this. One strategy for

automatic compensation is to identify event-free segments of the record and to correct the baseline estimate continuously by a small amount proportional to the difference between the latest segment and the baseline estimate. The estimate is then subtracted to give a zero-baseline record for event detection. This procedure is similar in effect to a first-order high-pass filter and is suitable for records with small drifts and moderate levels of channel activity. Automatic routines can, however, be "confused" by records with high activity (i.e., with little time spent at the baseline level) and by sudden changes in the baseline level. The most reliable technique is probably to fit the baseline, for example, by using a computer display of the data with a superimposed movable baseline cursor. In the method described in Section 4.2, the baseline position is constantly updated by means of a least-squares fit to any section of baseline that is on the screen.

Finally, it should be emphasized that the conditions described in this section for optimum *detection* of channel events are not necessarily the best conditions for *characterizing* channel events. Specifically, the best signal-to-noise ratios for event detection are sometimes obtained with relatively heavy filtering that distorts the shape of brief events. This presents no problem when the goal is to detect short, widely spaced events. However, as is shown in the next section, less filtering is desirable when one wants to discriminate the occurrence of two closely spaced short events from a single longer event or if one wants to determine the amplitude and duration of an event simultaneously.

3.4.2. Alternative Approaches to Event Detection

Sometimes it is not essential to minimize the probability of false events. In the time-course-fitting technique one intentionally places the threshold close to the baseline. Whenever this threshold is crossed, the computer displays the event that has been detected. It is then left to the operator to decide whether to fit the event or not. If an event is obviously false, there is no point in fitting it, but the decision about whether to fit or not is not critical as long as the resolution that is eventually imposed on the data (see Section 5.2) is such as to produce an acceptable false-event rate. The advantage of this approach is that it ensures that all events that are longer than the subsequently-imposed resolution are fitted.

A practical way to check the false-event rate, one that does not require careful measurement of σ_n, the baseline drift, etc., is simply to observe the frequency of detected events having the "wrong" polarity. If, for example, the true channel currents are positive going, any negative-going current pulses are most likely false events.

4. Characterizing Single-Channel Events

Since most single-channel current events appear to be rectangular steps of one or more amplitudes, the crucial step in analyzing a current record containing a single class of channel events is to determine the time of each current transition. These times can then be used for a kinetic analysis of the channel activity. The technical challenge is to characterize as many of the actual channel transitions as possible, including the briefest openings or gaps. In many cases the record can be modeled as a series of brief, widely spaced pulses having a width w that we wish to measure. Depending on the nature of the channel, these pulses could represent either openings or gaps. Special difficulties arise in the fitting process when the pulses are not widely spaced; the interpretation of histograms (see Section 6) is also compli-

cated in this case when the channel openings and gaps are both brief and roughly equal in duration.

It is the rule, rather than the exception, for records to contain more than one open-channel amplitude. This may result from the presence of more than one channel type in the patch and/or from the presence of one sort of channel that can open to more than one level. The question of how constant these levels are is discussed in Section 5.3.1, but regardless of this, the existence of multiple levels causes considerable problems, especially for the more automated methods of analysis. Sometimes amplitude estimates are just averaged together to give the "mean single-channel current" and although this is sometimes a reasonable procedure, it more usually is not. In practice, estimating the amplitude of long events is straightforward, but for short events, the estimation of the amplitude not only is unreliable but also increases the uncertainty in the transition time estimates. The usual practice, therefore, is to fit only the duration of brief events, with the amplitude constrained to some average value.

The methods of channel characterization we consider here are simple ones in which an attempt is made to detect channel-opening and closing events with a minimum of ambiguity. We recommend these methods because the bias and statistical errors in the characterization are relatively well known and because the detection of each event can be readily verified by the user. More sophisticated transition-detection schemes have been applied to single-channel data, including the Hinckley detector and T-test methods (see Chapter 3, this volume); these methods are not much better at characterizing simple isolated channel events but show promise in allowing better characterization of rapid bursts of events and subconductance levels. Still other methods exist that do not rely on the detection of individual events at all but obtain indirect information about dwell times and amplitudes from the statistics of the entire record. Examples of these are power spectra and all-points histograms computed from single-channel records. These provide less information than a full evaluation of closed and open times but can be used to fit simple models and thus estimate dwell times and amplitudes of rapidly switching channels in cases where these parameters cannot otherwise be obtained (see Section 5.3.2; Chapter 3, this volume). A more general technique is the application of "hidden Markov model" signal-processing algorithms (Chung *et al.*, 1990), which allow a complete model of the channel activity to be evaluated from all of the information in the recording. This technique allows the extraction of useful kinetic information from records having a signal-to-noise ratio several times lower than that required for the simple methods discussed here. However, it has not yet been applied widely to real data, or tested directly against alternative methods of analysis; its usefulness as a routine method is, therefore, not yet known.

4.1. Half-Amplitude Threshold Analysis

4.1.1. The Technique

The use of a simple threshold detector is the most widely used method of single-channel analysis and is readily applied to channels having only one nonzero conductance level. An estimate of the channel amplitude A_0 is used to set a threshold level, assumed here to be $A_0/2$. Every crossing of the threshold is interpreted as an opening or closing of the channel, so that the time spent above the threshold, w_t, is taken as an estimate of the channel-open time. As was pointed out by Sachs *et al.* (1982), choosing the threshold to be $A_0/2$ is convenient because w_t is then an unbiased estimate of the true pulse width w_0 for long pulses

of either polarity and can therefore be used to estimate both open and closed times. However, for short events with w_0 of the order of the filter risetime T_r, w_t underestimates w_0 (see Fig. 6). Events shorter than a dead time of about $T_r/2$ are missed altogether, because, after filtering they never reach the threshold.

The exact value of the dead time T_d of this detection technique can be either measured experimentally or calculated by finding the pulse width that gives a half-amplitude response from the recording system. If, for example, an analogue filter is used and has its bandwidth set far below that of the other parts of the recording system, it suffices to observe its output while variable-width pulses are applied to the input by a stimulator. In the case of a Gaussian filter, T_d is found (see equation 8) according to

$$\mathrm{erf}(T_d/2^{3/2}\sigma_g) = \frac{1}{2} \tag{19}$$

which yields $T_d = 0.538\ T_r$ or, equivalently, $T_d = 0.179/f_c$. If, for example, a sample rate of $10f_c$ is used (see Section 2.3), T_d is 1.79 sample intervals. Alternatively, a dead time can be imposed retrospectively, as described in Section 5.2 (as long as all events longer than the chosen value have been measured). This method ensures a consistent dead time throughout.

If not only the dead time but also the complete relationship between w_t and w_0 is known, then the distorting effect of the threshold-crossing analysis can be estimated. In terms of the filter step response $H(t)$, which is assumed for simplicity to be symmetrical about $t = 0$, the

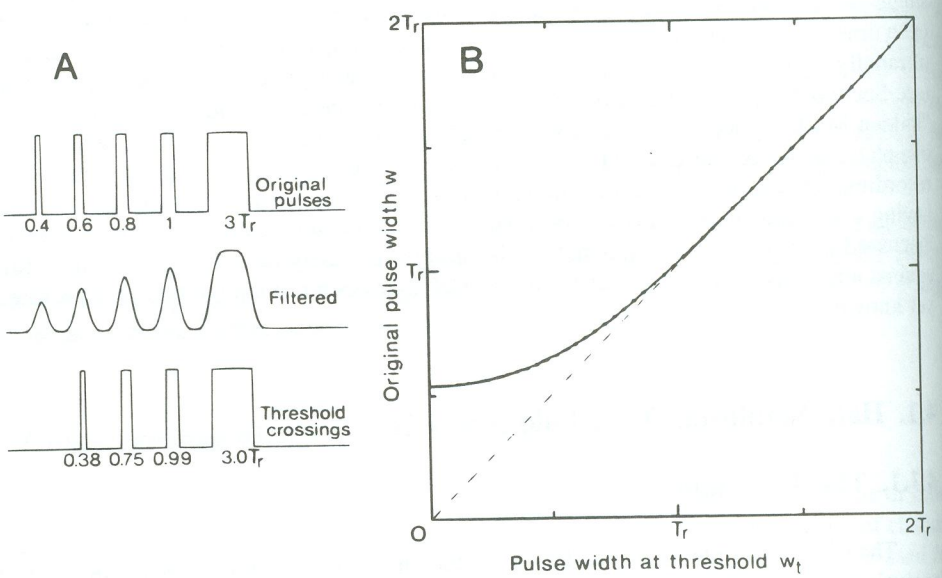


Figure 6. Relationship between true pulses with width w_o and the width w_t at the 50% threshold for Gaussian-filtered pulses. A: Simulated pulses with lengths given in units of T_r. The shortest pulse fails to reach threshold, and the pulses of intermediate width result in low values of the threshold-crossing width, w_t. B: The relationship between w_t and true pulse width in the absence of noise. For w_t equal to T_r or longer, w and w_t are essentially equal (dashed line). The points (barely visible under the curve) are values of the approximation function (equation 21).

relationship is given implicitly by

$$H\left(\frac{w_t + w_0}{2}\right) + H\left(\frac{w_t - w_0}{2}\right) = \frac{1}{2} \tag{20}$$

which must be evaluated numerically. Figure 6B shows this relationship for Gaussian-filtered pulses. A convenient approximation to the relationship, having relative errors less than 10^{-3}, is given by the function

$$\begin{aligned} w_0 &= g(w_t) \\ &= w_t + a_1 \exp(-w_t/a_1 - a_2 w_t^2 - a_3 w_t^3), \qquad w_t > 0 \end{aligned} \tag{21}$$

with $a_1 = 0.5382\ T_r$, $a_2 = 0.837\ T_r^{-2}$, and $a_3 = 1.120\ T_r^{-3}$. These coefficients are alternatively given in terms of the filter cutoff frequency as $a_1 = 0.1787\ /f_c$, $a_2 = 7.58\, f_c^2$, and $a_3 = 30.58\ f_c^3$. The function g can be used directly to convert the observed w_t values to effective w_0 values. Alternatively, the function can be used to predict the probability density function (pdf) of threshold-crossing intervals, $f_t(w_t)$, from the pdf of true durations $f(w_0)$ according to

$$f_t(w_t) = f[g(w_t)]g'(w_t) \tag{22}$$

Thus, in the absence of effects from noise, the distortions of this simple analysis scheme can be compensated by the fitting of a modified distribution to the resulting duration estimates.

4.1.2. Effect of Noise

Noise can be thought of as an instantaneous variation of the threshold level. For relatively long events, the Gaussian-distributed threshold fluctuations cause an approximately Gaussian-distributed random error in the determination of each threshold-crossing time. The standard deviation in the apparent corrected width, w, is approximately

$$\sigma_w = 2^{1/2} \frac{\sigma_n}{A_0} T_r \tag{23}$$

If the duration, w, of short events is corrected, for example, according to equation 21, the error in these estimates for w near T_d is

$$\sigma_w = \frac{\sigma_n}{A_0} T_r \tag{24}$$

and is also approximately Gaussian distributed.

The threshold-crossing technique automatically excludes events with (apparent) w_0 values less than T_d, but because of the effect of noise, some events with true w_0 values less than T_d will be counted, and some larger events will be missed. The general effect of noise is, therefore, a broadening and distortion of the distribution of apparent event durations as a result of randomness in the estimates of w_0. This broadening is most serious when the

underlying distribution $f(w_0)$ is rapidly varying. Figure 7 compares theoretical distributions of w_0 with distributions calculated on the basis of threshold-crossing analysis in the presence of a fairly high noise level (with $\phi/\sigma_n = 4$). When the time constant of the underlying w_0 distribution is less than T_r, an exponential fit to the observed, corrected distribution would yield a time constant that is too large (compare curves 1 and 3 in Fig. 7A). A similar effect of noise is to be expected on duration estimates obtained by the time-course-fitting technique, as errors in the estimates cause a "smearing out" of rapidly varying portions of the distribution.

For the best performance of the threshold-crossing analysis, it is generally best to decrease T_r as much as possible (that is, increase the filter cutoff f_c) to reduce the number of missed events. The same consideration applies here as in Section 3.4 above, however, about choosing a sufficiently large ϕ/σ_n ratio to give an acceptable false-event rate. When the threshold-analysis technique is implemented on a computer, an additional problem arises from the nature of digitized records. Because of the finite sample interval, it is possible for a set of digitized current values to lie below a certain threshold even when the original current trace crosses the threshold. This introduces an additional, biased error in the estimates of event durations. It is a good idea to use interpolation in order to minimize this effect, especially when the sample interval is relatively long.

The performance of the threshold-analysis technique has been considered so far only in the case of widely spaced events. A problem with the technique is that it responds poorly to short events that come very closely spaced in time. For example, a brief pulse can be counted as a longer one when it occurs in the vicinity of a second pulse (Fig. 8). This effect becomes significant when both the pulse length and the gap between pulses are roughly T_d or smaller. The systematic errors that are introduced by this failure have not been characterized, but they are probably not serious when either the mean open time or the mean gap time is at least several times T_r.

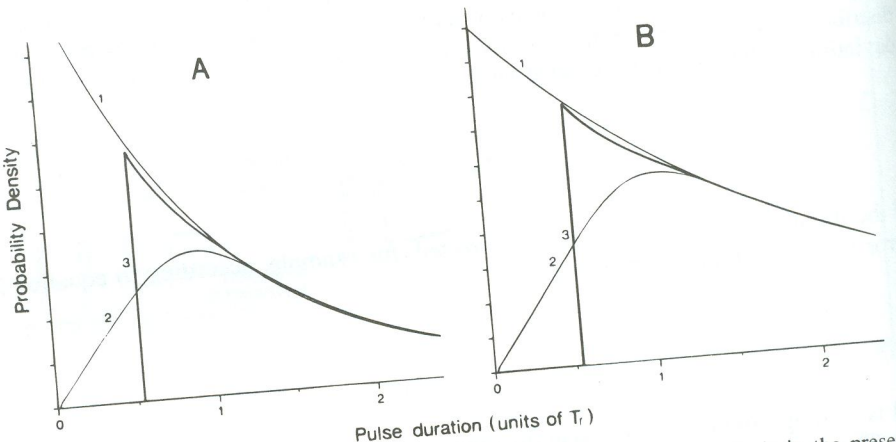


Figure 7. Distortion of pulse duration distributions by the threshold-crossing analysis in the presence of noise. The original distributions of durations w are shown by curve 1 in each part of the figure. The time constants of the distributions were T_r in A and $2T_r$ in B. The distribution of threshold-crossing times, w_t, is shown as curve 2 in each part, and the corrected distribution $g(w_t)$ is shown as curve 3. In the absence of noise, curve 3 would superimpose on the original distribution. A flat background noise spectrum with $\sigma_n = \phi/4$ was assumed.

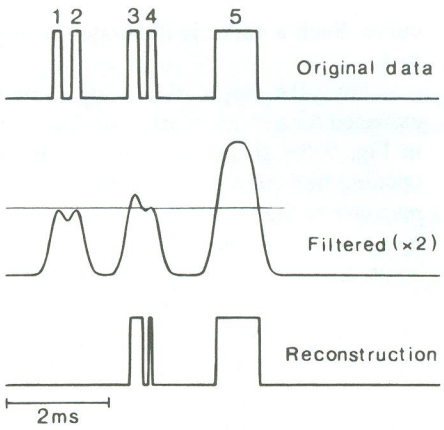


Figure 8. Threshold analysis of closely spaced events. Simulated rectangular events (top trace) were filtered with a 1-kHz Gaussian filter (T_r = 330 μs) and displayed (middle trace) with twice the vertical scaling. The reconstruction obtained from the threshold-crossing analysis is shown in the lower trace. Events 1 and 2 had lengths of $0.5T_r$ and were just below threshold. Event 4 was even shorter ($0.47T_r$) but was detected because it followed event 3 ($0.67T_r$) by a shut interval of only $0.5T_r$. Event 5 had a length of $2.5T_r$.

4.1.3. Estimating the Amplitude

The threshold-crossing technique assumes that the event amplitude is known *a priori,* so that the threshold can be set correctly. In practice, this presents little problem in interactive (as opposed to entirely automatic) fitting programs, since the operator can usually find sufficiently "square" events to provide an initial estimate for the amplitude. An estimate of the amplitude of an individual event, provided it is long enough, can be made by averaging the amplitude of the trace between threshold crossings, excluding the points within a given distance (e.g., 0.7 T_r) of the threshold-crossing points. Because of this exclusion, only events longer than about $2T_r$ can be used for determining the amplitude. As will be shown below, the time-course-fitting technique can give amplitude estimates for events shorter than this, but only at the expense of increased error in the duration estimates. This method suffers from the problem that the amplitude estimates so found will be too low if the region of the trace that is averaged contains brief shuttings that have not been detected because they did not cross the threshold level. If such brief shuttings are at all common (which is often the case), then it is necessary to inspect each amplitude fit to make sure that such bias has not occurred.

4.2. Direct Fitting of the Current Time Course

4.2.1. The Technique

A theoretical time course of the current can be computed on the basis of the step response of the recording system and fitted to the actual record. The step response can be measured by injecting a square-wave signal into the input of the patch-clamp amplifier, for example using a built-in integrator (see Chapter 4, this volume), or by coupling the triangle-wave output of a function generator into the headstage input through a small capacitance (e.g., by simply holding a wire near the headstage). A high-quality triangular wave is needed for this job. The resulting output signal, filtered and digitized in the same way as the data to be analyzed, is stored in a computer file for subsequent use. Usually, a suitable trigger pulse is also recorded, so that several sweeps can be averaged to obtain a smooth output

curve. Such a curve is illustrated in Fig. 9A; it is scaled so that it covers the range from 0 to 1.

Once the output of the apparatus to a step is known, it is easy to calculate the output expected for a series of steps such as a channel opening and shutting. The process is illustrated in Fig. 9 for single-channel openings of two different durations, t_o. The response to the opening transition is simply the step response function, which has already been stored. The response to the shutting transition is exactly the same but inverted and displaced to the right by t_o seconds. If these two curves are added, we obtain the expected output to a rectangular input, as illustrated for two examples in Figs. 9C and F.

This calculated output can be used to fit actual data as follows. The data are displayed on the screen, on which is superimposed the calculated response (output) to a rectangular input, which has been scaled by multiplying it by the amplitude of the opening. The amplitude cannot be measured from the event itself if it is very brief, so the amplitude must then be taken as the mean amplitude of all previous openings that have been fitted or as the amplitude of the last opening fitted. The times of the two transitions are then adjusted until the calculated output superimposes, as well as possible, on the data, as illustrated in Fig. 12. The adjustment of the amplitudes and transition times can be done manually or by means of a least-squares fit, as described below.

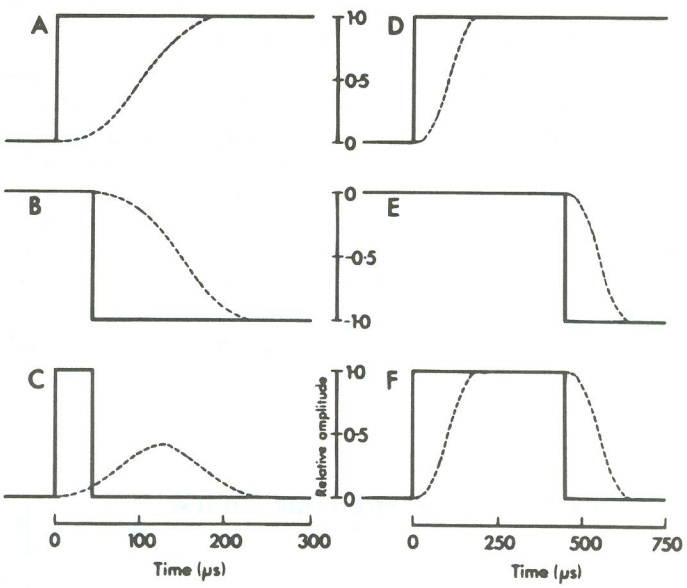


Figure 9. Illustration of the method of calculation of the expected response of the system from the measured response to a step input. The left-hand column illustrates a short (45-μs) pulse, and the right-hand column a longer (450-μs) pulse. The dashed lines in A and D show (on different time scales) the experimentally measured response to a step input, shown schematically as a continuous line, for a system (patch clamp, tape recorder, and filter) for which the final filter (eight-pole Bessel) was set at 3 kHz (−3 dB). A: The response to a unit step at time zero is shown. B shows the same signal but shifted 45 μs to the right and inverted. The sum of the continuous lines in A and B gives the 45 μs unit pulse shown as a continuous line in C. The sum of the dashed lines in A and B is shown as a dashed line in C and is the predicted response of the apparatus to the 45-μs pulse. It reaches about 41% of the maximum amplitude, which is very close to the value of 39% expected for a Gaussian filter (see equation 8). D, E, and F show, except for the time scale, the same as A, B, and C but for a 450-μs pulse, which achieves full amplitude.

4.2.2. Theory

The formal justification of the procedure illustrated in Fig. 9 is as follows. The step input at $t = 0$ is denoted $u(t)$, which is zero for $t < 0$ and unity for $t > 0$. A rectangular pulse input extending from time 0 to time w is therefore

$$s(t) = u(t) - u(t - w) \tag{25}$$

The output expected for this input can then be found (as long as the system behaves linearly) by convolving this input with the impulse response function of the system $h(t)$; i.e.,

$$y(t) = \int_0^t [u(t) - u(t - w)]h(t - \tau)d\tau \tag{26}$$

The system's response to a unit step input $u(t)$ is defined to be the system step response $H(t)$, which is the integral of h. Expressed in terms of $H(t)$, equation 26 simplifies to

$$y(t) = H(t) - H(t - w) \tag{27}$$

This is the calculation illustrated in Fig. 9. When the form of the input is inferred by superimposing this calculated response on the experimental data, we are performing a sort of graphic deconvolution.

This process can be extended to any number of transitions. In Fig. 10, some of the outputs that can result from four transitions (two rectangular pulses) are illustrated. If the transitions are well separated, the output, of course, simply looks like two somewhat rounded rectangular pulses (Fig. 10A). If the middle two transitions are close together, we have an opening with an incompletely resolved short gap (Fig. 10B). If the first three transitions are close together, the response looks like a single opening with an erratic rising phase (Fig. 10C). And if all four transitions are close together, the response looks like a (rather noisy) opening of less than full amplitude (Fig. 10D). If the channel were initially open in Fig. 10D, the response might be mistaken for an incomplete shutting to a conductance sublevel.

Before we go on to discuss the practical aspects of time course fitting, it is appropriate first to discuss the problems that may arise in attempting to fit both duration and amplitude simultaneously.

4.2.3. Simultaneous Determination of Amplitude and Duration

In theory, both the times and amplitudes of transitions in the theoretical trace could be varied to provide a best fit to the time course of the experimental record. The practical difficulty is that for pulse widths, w, shorter than the recording system risetime, T_r, the shape of the observed current pulse is relatively insensitive to w. In Fig. 11A, we compare the time courses of Gaussian-filtered pulses that have widths that differ by a factor of two but equal areas. Even in the absence of noise, the time courses are nearly indistinguishable for w less than about $T_r/2$.

To obtain a quantitative estimate of the errors to be expected in fitting the amplitude and duration simultaneously, the performance of a least-squares fitting routine for fitting the time course was evaluated. Figure 11B shows the behavior of the expected standard deviations,

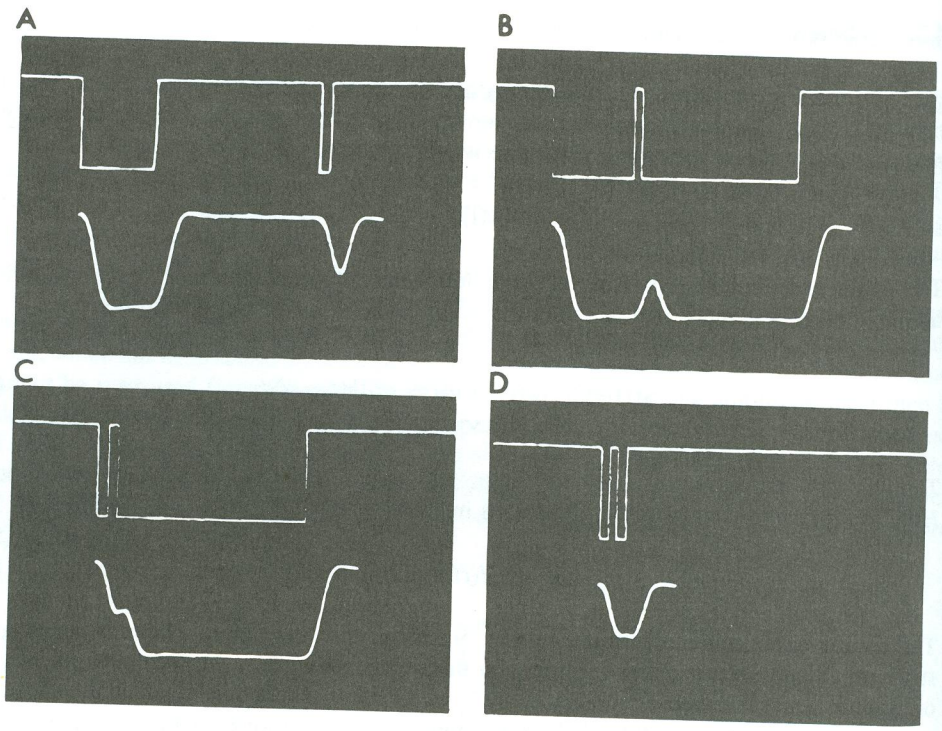


Figure 10. Examples of the calculated output of the apparatus (lower traces) in response to two openings of an ion channel (upper traces). The step response function used to generate the response is that specified in Fig. 9. The curves are generated by a computer subroutine and were photographed on a monitor oscilloscope driven by the digital-to-analogue output of the computer. Openings are shown as downward deflections. A: A fully resolved opening (435 μs) and gap (972 μs) followed by a partially resolved opening (67 μs). B: Two long openings (485 and 937 μs) separated by a partially resolved gap (45.5 μs). C: A brief opening (60.7 μs) and gap (53.1 μs) followed by a long opening (1113 μs); this gives the appearance of a single opening with an erratic opening transition. D: Two short openings (both 58.2 μs) separated by a short gap (48.1 μs); this generates the appearance of a single opening that is only 55% of the real amplitude but appears to have a more-or-less flat top, so it could easily be mistaken for a fully resolved subconductance level.

σ_A and σ_w, for the estimates of the amplitude and width, respectively, that are found using a linearized fitting process. Because the errors are proportional to the background noise standard deviation, σ_n, the values plotted in the figure are normalized with respect to σ_n; i.e., they are σ_A/σ_n and $\sigma_w A_0/\sigma_n T_r$. The behavior of the errors as a function of the original pulse width, w, depends on the form of the background noise spectrum; the two extreme cases of a flat spectrum and an f^2 spectrum are shown.

For long pulses, the error in the estimation of w is constant and is approximately 1.8 and 1.3 times $T_r\sigma_n/A_0$ for the flat and f^2 spectra, respectively. In a typical situation, $A_0/\sigma_n = 10$, which yields σ_w values in the range of 10–20% of T_r. The fact that σ_w is constant at large w can be understood from the way the duration of a long pulse is measured, as the interval between two transitions. If the transitions are far enough apart, the errors caused by noise in the determination of the transition times will be uncorrelated and independent of the time between them. On the other hand, amplitude estimates become more precise for

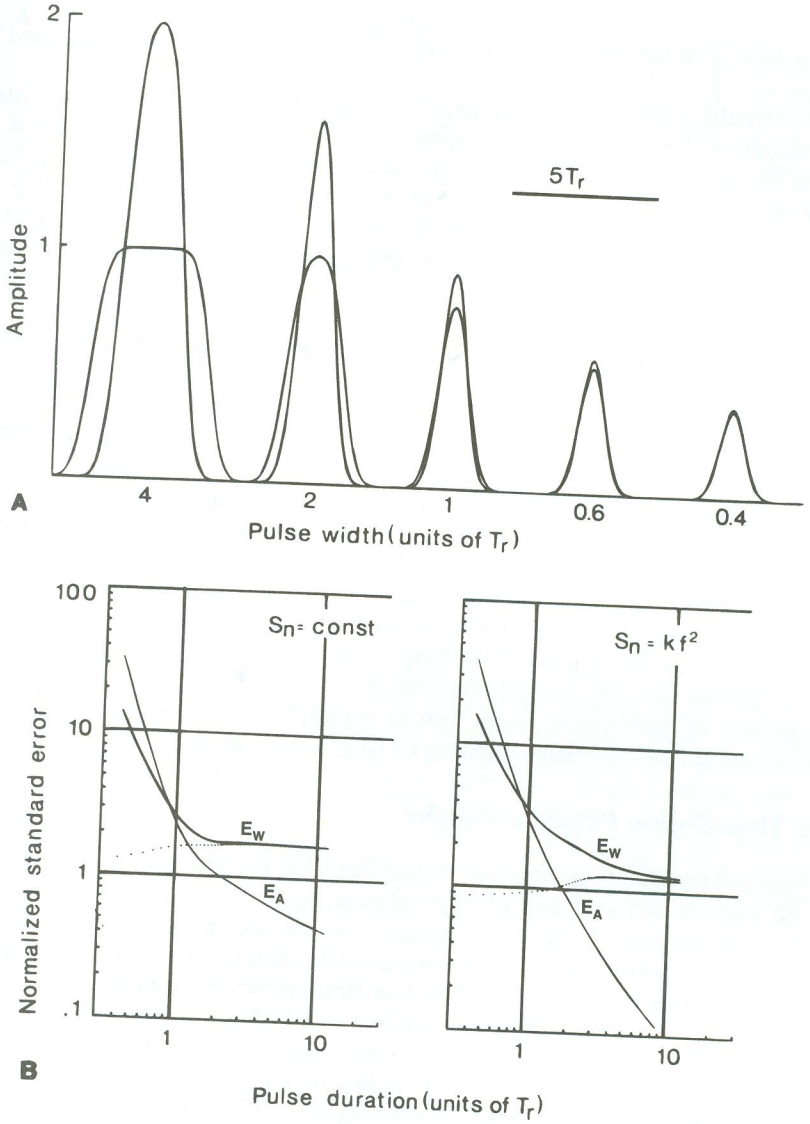


Figure 11. Errors in simultaneous fitting of amplitude and duration. A: Gaussian-filtered pulses of the widths indicated are superimposed with pulses having half the width but twice the amplitude. As the widths become shorter, the time courses become indistinguishable. B: Standard deviations of the estimates of amplitude and duration of Gaussian-filtered pulses in the presence of noise having either a flat or $1 + f^2$ spectrum. Pulse durations are given in units of the filter risetime T_r. The expected errors ϵ_A and ϵ_w are normalized to the background noise σ_n and other parameters according to $\epsilon_A = \sigma_A/\sigma_n$ and $\epsilon_w = \sigma_w A_o/\sigma_n T_r$. The dotted curve gives ϵ_w when the amplitude estimate is constrained to the correct value A_o.

longer pulses, with the error decreasing as $w^{-1/2}$ in the case of a flat background spectrum and large w.

As the pulse width becomes comparable to T_r or shorter, the errors of estimates of both amplitude and width increase sharply, becoming double their asymptotic values at about 0.8 T_r in the flat-spectrum case. This sharp rise does not occur if the amplitude is constrained and the duration alone is fitted, as illustrated by the dotted curves in Fig. 11B. This rise reflects the difficulty of simultaneous fitting. Because it occurs in the vicinity of T_r, it can be seen that a small T_r, i.e., the largest possible filter bandwidth, is best for simultaneous fitting. Of course, in practice the filter bandwidth must be chosen low enough to avoid false events.

If the duration alone is fitted, with the amplitude held fixed, the error in the duration estimate depends only weakly on w and, in fact, decreases slightly as w becomes small, as shown by the dotted curves in Fig. 11B. The absolute size of the error is much smaller, and the criterion for choosing f_c to minimize the error (which is essentially proportional to $T_r\sigma_n$) is similar to that for event detection.

In conclusion, it is possible to obtain some amplitude information from events shorter than the recording system risetime T_r. In practice, this information is difficult to obtain because it is based on fine details of the pulse shape, but it could conceivably be useful for statistically testing hypotheses such as the existence of multiple channel populations. Much more precise estimates for the duration of short channel events can be obtained by fixing the amplitude in the fitting process. For longer events with $w \geq 2\ T_r$ the concurrent estimation of the amplitude has only a small effect on the error of the duration estimates. Estimating the amplitude of these events would then be worthwhile provided that the size of the error in the amplitude estimates (approximately equal in magnitude to σ_n at $w = 2\ T_r$) is acceptable.

4.2.4. Time-Course Fitting in Practice

Personal computers are now fast enough that it has become feasible to fit simultaneously both the duration and amplitude of single-channel openings. With the program SCAN, which is under development at University College London, the fit does not take any noticeable length of time for fitting up to four transitions and is still quite acceptable for fitting say ten or more transitions when run on an 80486 or Pentium processor machine. The data trace is scrolled across the screen until an event is detected, as described in Section 3.4.2. The trace is then expanded, contracted, or shifted as necessary to get a suitable section of data for fitting on the screen. The program then makes initial guesses for the positions of all the transitions and amplitudes, performs a least-squares fit on the basis of these guesses, and displays the fitted curve for acceptance, rejection, or modification by the operator. Three examples of fits done in this way are shown in Fig. 12.

When the channel is shut at each end of the fitted region, as in Fig. 12A and B, fitting n transitions involves estimation of $2n + 1$ parameters (the time at which each transition occurs, the amplitude following each transition, and the amplitude before the first transition). The fit of the amplitude after the last transition is taken as a new estimate of the current baseline position. The program "knows" that the channel is shut at this point, so the next transition must be an opening; the average of a section of trace before the next opening can therefore be taken as a temporary baseline estimate, even if drift has occurred, thereby allowing reasonable initial guesses to be made for the next fitting. In this way it is possible to keep track of the baseline level. Other options in the program allow fits to be done with only one open level for all openings or to be done by specifying the amplitudes in advance and fitting only the transition times (as was always done with earlier programs).

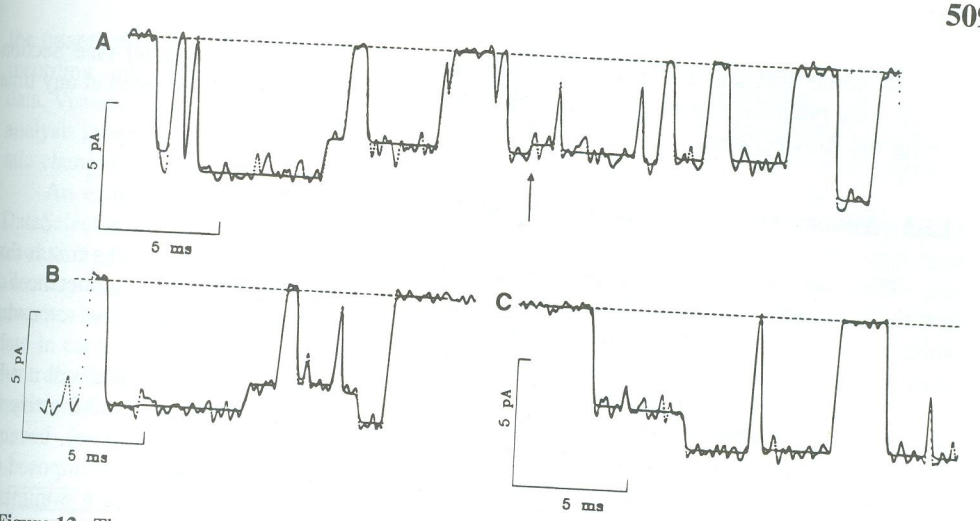


Figure 12. Three examples of fitting with the SCAN program. The record is from an NMDA-type glutamate receptor in a dentate gyrus granule cell (outside-out patch at −100 mV, glutamate 200 nM + glycine 1 μM with 1 mM Ca and no added Mg, eight-pole Bessel filter at 2 kHz, −3 dB, risetime 166 μs; methods as in Gibb and Colquhoun, 1991; data of A. J. Gibb). The dashed line shows the baseline (shut) level. The record was sampled at 50 kHz (though 20 kHz would have been sufficient and more usual). The transition times, and amplitudes (for events that were longer than two risetimes), were fitted simultaneously by least squares. Shut periods shorter than two risetimes had their amplitudes fixed to zero. Open periods shorter than two risetimes had their amplitude constrained to be the same as that of the closest opening that was longer than three risetimes. The fitted curve is the continuous line. A: Two contiguous fittings. The durations and amplitudes in this fit, starting from the first opening, are as follows: 0.707 ms, −4.48 pA; 0.491 ms, 0 pA; 0.248 ms, −5.22 pA; 0.321 ms, 0 pA; 5.33 ms, −5.24 pA; 0.894 ms, −3.69 pA; 0.802 ms, 0 pA; 3.08 ms, −3.75 pA; 0.216 ms, 0 pA; 0.074 ms, −3.73 pA; 1.83 ms, 0 pA; 0.092 ms, −3.91 pA; 0.448 ms, 0 pA; 1.02 ms, −3.91 pA; 1.07 ms, −3.52 pA; 0.131 ms, 0 pA; 3.17 ms, −3.74 pA; 0.156 ms, 0 pA; 0.756 ms, −4.04 pA; 0.511 ms, 0 pA; 1.39 ms, −3.80 pA; 0.865 ms, 0 pA; 2.47 ms, −3.75 pA; 1.92 ms, 0 pA; 1.59 ms, −5.06 pA. Note that the transition from −3.91 pA to −3.52 pA (marked with arrow) is dubious, and this would probably be removed later, at the stage when the resolution is imposed on the data (see text, Section 5.2), when adjacent openings that differ in amplitude by less than some specified amount are concatenated into a single opening (with the average amplitude). B and C: Two more examples. In B there is a very small transition (from −4.98 to −4.90 pA) shortly after the first opening transition; this was triggered by the wobble in the data at this point but would certainly be removed before analysis (see A).

The fitting of amplitudes in this way will be biased if the regions of trace that are fitted contain brief shuttings, as discussed *a propos* threshold-crossing analysis in Section 4.1.3. This problem can be minimized by allowing the program to fit very brief events, even though most of them will be rejected later, when a realistic resolution is imposed (see Section 5.2).

It is, as discussed in Section 4.2.3, not feasible to fit both amplitude and duration to very short openings or shuttings. Shut periods shorter than a specified length (usually two risetimes) have their amplitudes fixed to zero. Open periods shorter than a specified length (also usually two risetimes) have their amplitude constrained to be the same as that of the closest opening that is longer than, say, three risetimes, if such an opening is present in the region of trace being fitted. Otherwise, the amplitude of short openings is fixed at the current mean full amplitude (or some other specified value).

Once a satisfactory fit has been obtained, the data points in the fitted region can be entered into an all-points histogram. Also, those data points that are in regions where the fitted curve is flat can be entered separately into shut-point and open-point histograms, which

exclude points that are in the region of transition from one level to another (see Section 5.3.2). This procedure means that these three sorts of histogram can be viewed at any time during the fitting process.

4.2.5. Advantages and Disadvantages of Time-Course Fitting

There are two major advantages in using the time-course-fitting method. The first is that it is the only well-tested method for dealing with records that contain multiple conductances or subconductance states. The second is that the resolution of measurements can be somewhat greater than can be obtained with the threshold-crossing method.

It is quite likely that, during time-course fitting, some of the events fitted will not be real openings or shuttings of the ion channel but merely random noise or small artifacts. This is not really a disadvantage of the method (except insofar as it takes time), because such events should be eliminated at a later stage, when a realistic resolution is imposed on the idealized record (Section 5.2). In fact, it is actually an advantage, because it minimizes the bias in amplitude estimates that result from the presence of brief events that may be detectable but would not normally be fitted.

There will, from time to time, be events on the screen that are ambiguous. It may be impossible to tell whether an event is a genuine channel opening at all, or whether it is some form of interference. And even if the event is "obviously" an opening, it may be impossible to be sure whether it is an opening to a subconductance level or whether it is two or more brief full openings separated by short gap(s) (as illustrated in Figs. 8 and 10). Such events will necessitate a subjective decision by the operator about the most likely interpretation of the data. Magleby (1992) has criticized the method because of the "operator bias" that is introduced into the analysis in this way. However, exactly the same sort of operator bias will occur in any form of threshold-crossing analysis in which the operator inspects and approves or disapproves what the program has done. As mentioned above, it is highly desirable that the operator should know what the program has done. It is equally very desirable that the operator should be aware that the data contain ambiguous events, even if he/she is not sure what to do with them. The only case in which the argument about operator bias seems to be valid is when data are analyzed automatically by the "total simulation" method proposed by Magleby and Weiss (1990). In this case, it is necessary that a completely automatic method of analysis be used because of the immense amount of computation that is involved, and it is necessary that the simulated and experimental records be analyzed by identical methods (including the ambiguous bits). In all other cases, there is little to be gained by sweeping the ambiguities under the carpet.

The question of ambiguous events has been discussed at some length. However, it is probably true, at least for channels that have a reasonably good signal-to-noise ratio, that such events are sufficiently rare that the conclusions from the analysis are unlikely to be much altered by the subjective decisions that must occasionally be made.

4.3. Event Characterization Using a Computer

4.3.1. Data Display

The single most important feature of a computer system for analyzing single-channel data is a responsive and flexible means of displaying the digitized data. Before and during

the quantitative event characterization, it is essential that the user be able to examine the recording, millisecond by millisecond if necessary, to be able to judge the quality of the data. Visual inspection can show features that could be missed or misinterpreted by automatic analysis programs, such as the presence of artifacts or superimposed channel events, systematic changes or "rundown" of the channel activity, and subconductance levels.

An example of a suitable display for long, continuous data recordings is that of the DataSelector program shown in Fig. 13. Here the data are shown at three different time resolutions, providing an overview of the entire multimegabyte file (top trace) while also allowing inspection of a selected region at high resolution. One important feature of the program is the ability of the user to select the position and degree of magnification of the data in each trace. As the box in a trace is dragged or resized using the computer's mouse, the trace below it is redrawn to correspond to the region enclosed by the box. Another important feature of the program is the rapid, flicker-free redrawing of the traces as they are

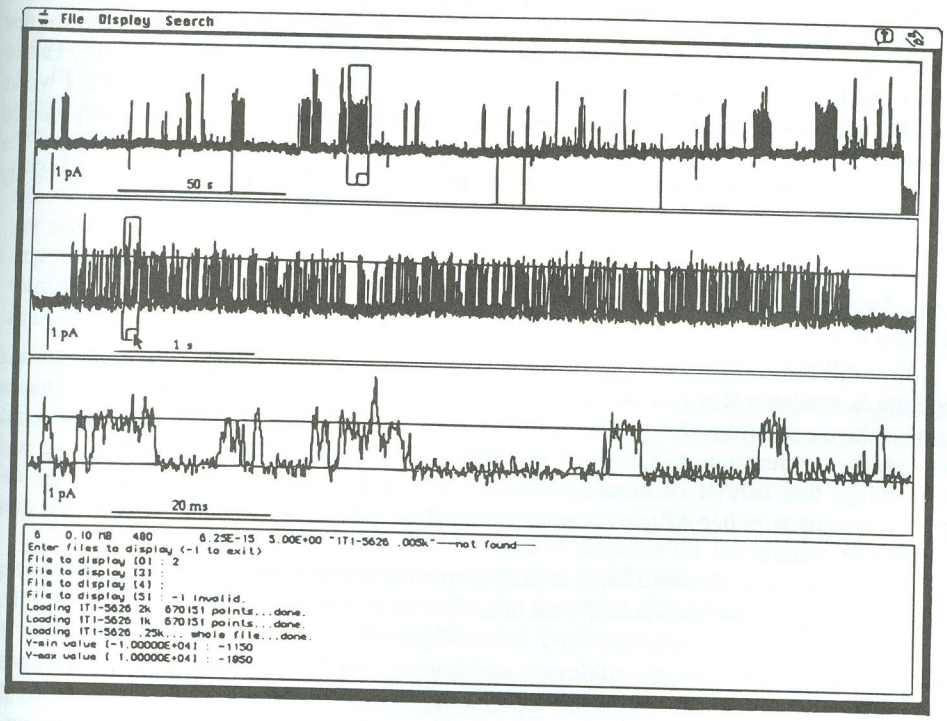


Figure 13. Perusal of a recording using the DATA SELECTOR program. Data are shown from a 4-min recording of potassium-channel currents that includes a slow baseline drift and several spikes from electrical interference. The top trace shows an overview of the entire recording; the region indicated by the box, about 10 s in duration, is expanded as the middle trace. The mouse cursor is positioned to change the size of the box in the middle trace, which selects the 150-ms segment shown in the bottom trace. A brief upward spike that is visible in the upper traces is seen in the bottom trace to be too broad to be a simple noise spike; it also has approximately twice the amplitude of the main channel events, suggesting that it represents an overlapping channel opening. The original recording was obtained with a VR-10 PCM/VCR recording system; the data were transferred directly to the Macintosh computer, creating a 49-MB data file at the 94-kHz sample rate. An off-line Gaussian filtering program, in turn, created synchronized, filtered files with bandwidths of 250 Hz, 1 kHz, and 2kHz. The DATA SELECTOR program reads data from these files as needed to draw and update the display.

rescaled. This is accomplished by first drawing each trace on an off-screen pixel map and then copying it to the screen buffer. The copying operation is very fast, providing an essentially instantaneous update. The drawing operation itself is also fast enough (usually taking less than 100 ms) so that the scrolling and changes of magnification appear smooth and continuous to the user.

For this sort of display, it is important to have fast graphics. The trace-drawing routine used in DataSelector was written in assembly language and is optimized for rapidly graphing arrays of thousands of data points.* It draws directly to the offscreen pixel-map memory rather than making calls to the operating system's graphics routines. Similarly, high-speed displays on IBM-compatible personal computers typically use graphics subroutines that write directly to the video memory rather than using the BIOS interrupts.

For the characterization of events the computer display must also be able to superimpose cursors or reconstructed transitions over the raw data and allow the user to make manual adjustments and corrections. For the 50% threshold analysis, it is sufficient to use the computer's mouse to adjust two variable parameters, the estimated current amplitudes before and after a transition. Time-course fitting requires more adjustable parameters, and for that purpose a set of knobs (i.e., potentiometers that are read by the computer's ADC) can be more flexible than the mouse, though when the method described in Section 4.2.4 works well, the number of manual adjustments that are needed is small, and mouse/keyboard operation is feasible. Use of the numerical keypad, rather than letter keys or mouse, for making menu choices is much more ergonomically satisfactory for operations that are highly repetitive (and single channel analysis is certainly in this category).

4.3.2. Programs

It is still the case, 18 years after the invention of the patch clamp, that no commercial program is available that can perform all of the methods that are described in this chapter. Perhaps the most serious thing that is lacking is a satisfactory program for analyzing records that contain conductance sublevels or multiple conductance levels. At present, if you wish to do things that cannot be done by the commercially available programs, there are two options. You must either write a program yourself or get one from somebody who has done the job you require.

Many programs offer the choice of fully automatic analysis, without any visual inspection of how the program has interpreted your data. Use of such methods is very dangerous unless your data are of high quality and have been subjected to some preliminary check that the baseline stability, conductance sublevels, ambiguous events, and artifacts are all within the range that the program can cope with safely (e.g., see Magleby, 1992). If done thoroughly, such a check may take almost as long as checking individual fits unless your recording is of exceptionally high quality. The speed of automatic methods obviously makes them very

*The drawing algorithm is based on the observation that the display of a trace can be generated by a set of vertical lines, one for each horizontal pixel position in the display. Often there are many more data points to be graphed, say 10^4 or 10^5, than the number of horizontal pixel positions, which might be only 640 or 1024. In simplified form the algorithm can be understood as follows: let n be the number of data points corresponding to a given horizontal pixel position. The endpoints of the vertical line to be drawn at that position are chosen simply to be the minimum and maximum values of $n + 1$ data points (including one from the set of points corresponding to the next horizontal position). Because only vertical lines are to be drawn, the actual drawing routine can be very simple and efficient.

attractive, but the computer maxim "garbage in, garbage out" certainly applies to single-channel analysis, and it may require some investment of time to ensure that you do not get "garbage out."

The earlier forms of time-course fitting were substantially more time consuming than threshold-crossing analysis, even when the fits produced by the latter were inspected. However, the methods described above are faster, and there is now probably not much difference, at least for data that are good enough that initial guesses for transition times and amplitudes are usually satisfactory, so few manual adjustments are needed. As personal computers get faster, so the time taken for least-squares fitting of many parameters will be reduced still further, and the difference between the various methods will become negligible. The speed of the analysis will depend only on the amount of visual checking that is done.

4.3.3. Storing the Idealized Record

The output from these programs is a list of numbers representing the time of each transition in the current record and the amplitude of the transition. This list contains all of the information present in the idealized record that is constructed in the fitting process. Generally, this information is stored in a file by the computer for further processing, such as sorting into histograms or fitting of distributions. Although in principle only two numbers need be stored for each transition in the original record, it is a good idea to include some more information in the file to allow for mistakes that inevitably occur in the analysis process. For example, if the only clue to the number of channels open is the number and polarity of step amplitude values, the corruption of a single entry could cause much confusion. One format for the storage of data, used by the TAC program, which performs threshold-crossing analysis, stores a record containing the following information as an entry for each transition:

1. AbsTime, the time of the transition (LONGREAL in seconds)
2. EventType, the kind of event. This is an enumerated type, having values corresponding to (1) normal transition, (2) interval of data to be ignored, (3) transition between conductance levels, etc.
3. Level, the number of channels open after the transition (INTEGER)
4. PreAmp, the current amplitude before the transition (REAL, in amperes)
5. NumPre, the number of data samples used to estimate the preamplitude (zero if the amplitude was not determined automatically; INTEGER)
6. PostAmp, the current amplitude after the transition (REAL, in amperes)
7. NumPost, the number of data samples used to estimate the postamplitude (INTEGER)

The use of a LONGREAL (64-bit floating-point) value provides sufficient numerical resolution (better than 1 nanosecond in 24 hr) to allow the absolute time of each event to be stored, even when the transition time has been interpolated to a fraction of a sample interval. This greatly simplifies operations in which individual transition records are edited and also allows each event to be synchronized with its position in the raw data file. The current amplitudes are documented by their values as well as the number of points used to estimate them (when automatic amplitude estimation is in effect) so that the reliability of the values can be estimated. This record structure occupies 24 bytes of storage for each event. The resulting event list files are nevertheless much shorter than the raw data files they describe.

The NumPre, NumPost, and EventType indicators allow the subsequent analyses to be carried out with certain values (e.g., ambiguous amplitudes) either included or excluded. This will allow a judgment as to the influence of the ambiguities on the conclusions from the analysis.

5. The Display of Distributions

Analysis of the experimental results by one of the methods described in Section 4 produces an idealized record. This takes the form of an event list that contains the duration of each event and the amplitude of the single-channel current following each transition (or, for some sorts of analysis, only a record of whether the channel was open or shut). We now wish to move on to discuss the ways in which the information in this event list can be viewed and fitted with appropriate curves.

5.1. Histograms and Probability Density Functions

5.1.1. Stability Plots

This section deals mainly with the display of measurements that have been made at equilibrium, so the average properties of the record should not be changing with time. In practice, it is quite common for changes to occur with time, and any such change can easily make the corresponding distribution meaningless. It is, therefore, important to check the data for stability before distributions are constructed or fitted. This can be done by constructing *stability plots* as suggested by Weiss and Magleby (1989). In the case, for example, of measured open times, the approach is to construct a moving average of open times and to plot this average against time or, more commonly, against the interval number (e.g., the number of the interval at the center of the averaged values). A common procedure is to average 50 consecutive open times and then increment the starting point by 25 (i.e., average open times 1 to 50, 26 to 75, 51 to 100, etc). The overlap between samples smoothes the graph (and so also blurs detail). An exactly similar procedure can be followed for shut times and for open probabilities. In the case of open probabilities, a value for P_{open} is calculated for each set of 50 (or whatever number is chosen) open and shut times as total open time over total length. If a shut time is encountered that has been marked as "unusable" during analysis (see Section 4.3.3), then the set must be abandoned and a new set started at the next valid opening.

Figure 14 shows examples of stability plots for amplitudes (in A, C, and E) and for open times, shut times, and P_{open} (in B, D, and F). Graphs for A–D are from experiments with recombinant NMDA receptors. The two amplitude levels are stable throughout the recording for the experiment shown in Fig. 14A and B, though there is a modest tendency in B for shut times to decrease and for P_{open} to increase correspondingly during the experiment. In contrast, Fig. 14C shows a different experiment in which the two amplitude levels both show a sudden decrease after about the 900th interval. Amplitude histograms from such an experiment would show three or four levels but would of course give no hint that there had been a sudden change in the middle of the experiment. The corresponding stability plots for open times, shut times, and P_{open}, shown in Fig. 14D, also show instability; shut times decrease, and P_{open} correspondingly increases, at about the same point in the experiment where the amplitude changes. The open times, however, remain much the same throughout in D, as is also the case for B and F. Figure 14E and F show similar plots from an experiment on adult frog endplate nicotinic receptors, in which all the measured quantities remain stable throughout the recording; data from this experiment were used to construct the shut-time histogram shown in Fig. 15.

Plots of this sort can be used to mark (e.g., by superimposing cursors on the plot)

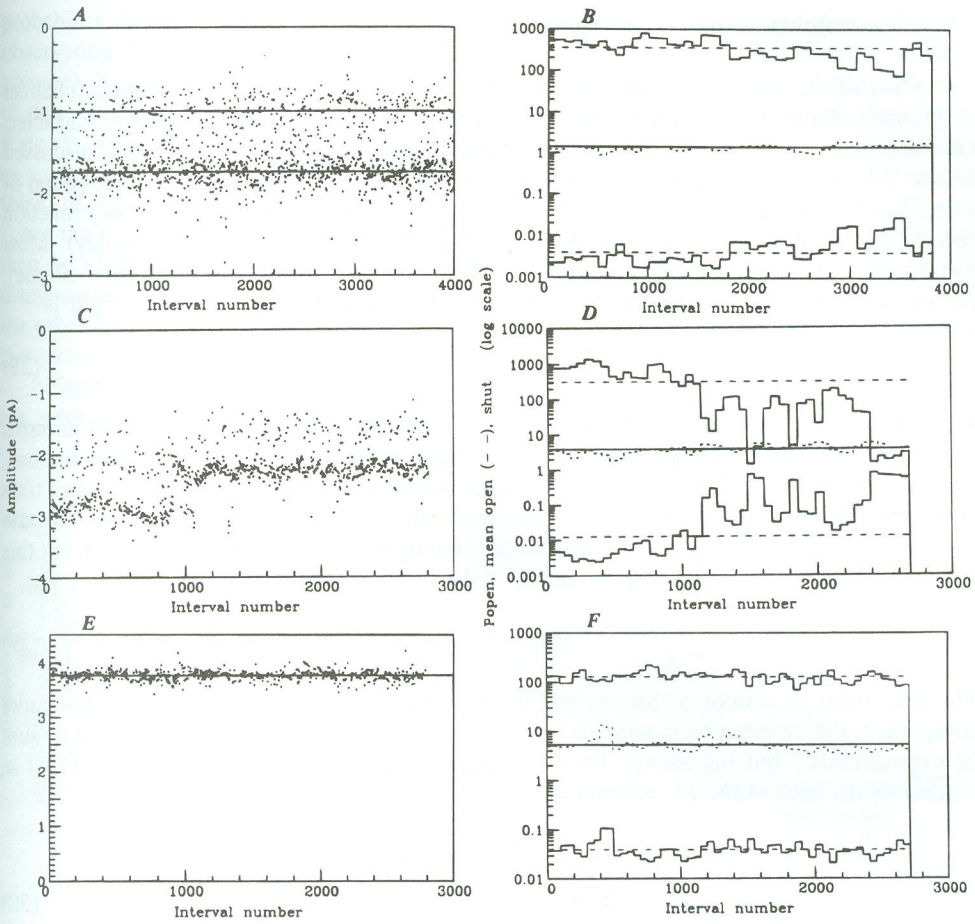


Figure 14. Examples of stability plots. Data for A, B, C, and D are from NMDA-type glutamate receptors expressed in oocytes (unppublished data of P. Stern, P. Béhé, R. Schoepfer, and D. Colquhoun; methods as in Stern *et al.,* 1992). Oocytes were transfected with NR1 + NR2C subunits in A and B (4002 resolved intervals) and with NR1 + NR2A + NR2C subunits in C and D (2810 resolved intervals). A and C show amplitude stability plots; the horizontal lines in A mark the amplitudes that were fitted to the amplitude histogram, −1.01 pA and −1.75 pA. B and D show stability plots for shut time (top), open time (middle), and P_{open} (bottom). Average of 50 values plotted, with increment of 25 intervals. Horizontal lines show the average values for the whole run. E and F show the same two types of stability plot for the same frog endplate nicotinic receptor data that was used to construct the histograms in Fig. 15 (amplitudes are plotted as positive numbers in E).

sections of the data that are to be omitted from the analysis. For example, this approach has been used to inspect, separately, the channel properties when the channel is in a high-P_{open} period and when it is behaving normally.

It should be noted that when the average P_{open} value (the value for the whole of the data) is plotted on the stability plot, it can sometimes appear to be in the wrong position. This may happen when the record contains a very long shut period that reduces the overall P_{open} but affects only one point on the stability plot (which is normally constructed with interval number on the abscissa rather than time).

5.1.2. Probability Density Functions

Most of the data with which we have to deal consist of continuous variables (channel amplitudes, durations of open periods, etc.) rather than discontinuous or integer variables. One exception is the distribution of the number of openings per burst, which is discussed below; this number can, of course, take only integer values. The probability distribution of a continuous variable may be specified as a probability density function, which is a function specified such that the area under the curve represents probability (or frequency). Most commonly, the pdf is an exponential or sum of exponentials (see Chapter 18, this volume). For example, if a time interval has a simple exponential with mean $\tau = 1/\lambda$, its pdf is

$$f(t) = \lambda e^{-\lambda t} \qquad t > 0 \tag{28}$$

which has dimensions of s^{-1}. Alternatively, the exponential density can be written in terms of the time constant, τ, rather than the rate constant, λ. This is preferable for two reasons. First, it is easier to think in terms of time rather that rate or frequency. Second, use of time constants prevents confusion between *observed* rate constants (denoted λ) and the rate constants for transitions between states in the underlying mechanism (see Chapter 18, this volume). Thus, equation 28 will be written in the form

$$f(t) = \tau^{-1} e^{-t/\tau} \tag{29}$$

The area under this curve, as for any pdf, is unity. When there is more than one exponential component, the distribution is referred to as a *mixture of exponential distributions* (or a "sum of exponentials," but the former term is preferred since the total area must be 1). If a_i represents the area of the ith component, and τ_i is its mean, then

$$\begin{aligned} f(t) &= a_1 \tau_1^{-1} e^{-t/\tau_1} + a_2 \tau_2^{-1} e^{-t/\tau_2} + \cdots \\ &= \Sigma a_i \tau_i^{-1} e^{-t/\tau_i} \end{aligned} \tag{30}$$

The areas add up to unity; i.e.,

$$a_1 + a_2 + \cdots = 1$$

or

$$\Sigma a_i = 1 \tag{31}$$

and they are proportional, roughly speaking, to number of events in each component. The overall mean duration is given by:

$$\text{mean duration} = \Sigma a_i \tau_i \tag{32}$$

In practice, the data consist of an idealized record of time intervals constructed by one of the methods described above (see Section 4). This record may be revised to ensure consistent time resolution (see Section 5.2). The open times, shut times, and other quantities of interest can be obtained from it. For example, the data might consist of a series of n open times $t_1, t_2, \ldots, t_n$. They might be, for example, 1.41, 5.82, 3.91, 10.9 . . . , 6.43 ms. The

probability density function is, roughly speaking, proportional to the probability that the observation falls within an infinitesimal interval (from t to $t + dt$; see Chapter 18, this volume). But we have not got an infinite data set, so the pdf of the data looks like a series of delta functions (one at each measured value). This sort of display is not very helpful as it stands, so we smooth it by using a finite binwidth. In other words, we display a histogram as an approximation to the pdf by counting the number of observations that fall in intervals (bins) of specified width. In the example above, we might use 1 ms as the bin width and count the number of observations between 0 and 1 ms, 1 and 2 ms, and so on. These can then be plotted on a histogram as illustrated, for example, in Fig. 15. The histogram is discontinuous, and its ordinate is a dimensionless number. The pdf it approximates is, on the other hand, a continuous variable with dimensions of s^{-1}, so care is needed when both histogram and pdf are plotted on the same graph (see Section 5.1.5).

Figure 15A shows a histogram of shut times, with a time scale running from 0 to 1500 ms, with a bin width of 80 ms. This range includes virtually all the shut times that were

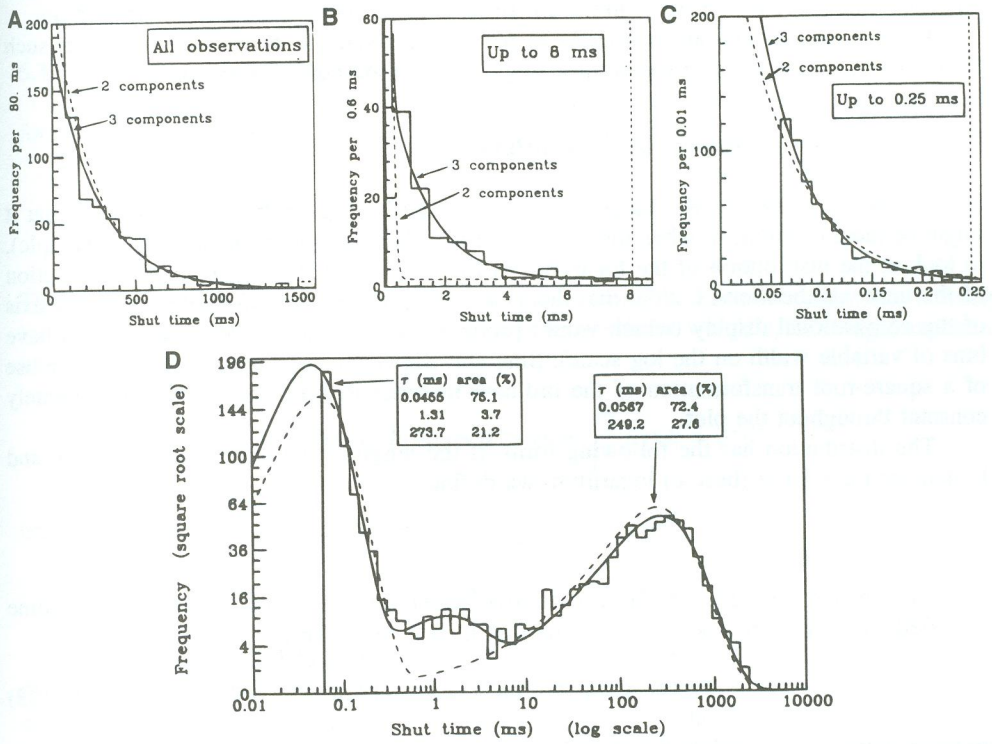


Figure 15. Example of a distribution of shut times. In A, B, and C, the histogram of shut times is shown (on three different time scales), and in D the distribution of log(shut times) for the same data is shown. The data are from nicotinic channels of frog endplate (suberyldicholine 100 nM, −130 mV). Resolutions of 80 μs for open times and 60 μs for shut times were imposed as described in the text; this resulted in 1348 shut times, which were used to construct each of the histograms. The dashed bins (which are off scale in B and C) represent the number of observations above the upper limit. The data were fitted by the method of maximum likelihood with either two exponentials (dashed curve) or three exponentials (continuous curve). The same fits were superimposed on all of the histograms. The estimated parameters are shown in D. (D. Colquhoun and B. Sakmann, unpublished data.)

observed. The first bin actually starts at $t = 60$ μs rather than at $t = 0$ because a resolution of 60 μs was imposed on the data (see Section 5.2 below), so there are no observations shorter than this. All that is visible on this plot is a single slowly decaying component with a mean of about 250 ms, though the first bin, the top of which is cut off on the display, shows that there are many short shut times too. The same data are shown again in Fig. 15C, but only shut times up to 250 μs are shown here, with a bin width of 10 μs; the 60-μs resolution is obvious on this plot. There are many shut times longer than 250 μs of course, and these are pooled in the dashed bin at the right-hand end of the histogram (the top of which is cut off). Again, the histogram looks close to a single exponential, but this time with a mean of about 50 μs. Although it is not obvious from either of these displays, there is in fact a (small) third component in this shut-time distribution. It is visible only in the display of the same data in Fig. 15B, in which all shut times up to 8 ms are shown (with a bin width of 0.6 ms), where an exponential with a mean of about 1 ms is visible. The data were not fitted separately for Figs 15A, B, and C, but one fit was done to all the data (by maximum likelihood—see Section 6) with either two exponential components (dashed line) or 3 exponential components (solid line). This same fit is shown in all four sections of Fig 15. The inadequacy of the two-component fit is obvious only in the display up to 8 ms.

Clearly, the conventional histogram display is inconvenient for intervals that cover such a wide range of values. The logarithmic display described next is preferable.

5.1.3. Logarithmic Display of Time Intervals

It was suggested by McManus *et al.* (1987) and by Sigworth and Sine (1987) that it might be more convenient, when intervals cover a wide range (as in the preceding example), to look at the distribution of the logarithm of the time interval rather than the distribution of the intervals themselves. Note that this is not simply a log transformation of the x axis of the conventional display (which would produce a curve with no peak, and would have bins of variable width on the log scale). Sine and Sigworth suggested, in addition, the use of a square-root transformation of the ordinate in order to keep the errors approximately constant throughout the plot.

The distribution has the following form. If the length of an interval is denoted t, and ln denotes the natural (base e) logarithm, we define

$$x = \ln(t)$$

then we can find the pdf of x, $f_x(x)$, as follows. First we note that if a t is less than some specified value t_1, then it will also be true that $\ln(t)$ is less than $\ln(t_1)$. Thus,

$$\text{Prob}[t < t_1] = \text{Prob}[\ln(t) < \ln(t_1)] = P \tag{33}$$

In other words, the cumulative distributions for t and $\ln(t)$ are the same. Now it is pointed out in Chapter 18 (this volume, Section 3.1) that the pdf can be found by differentiating the cumulative distribution. Thus, denoting the probability defined in equation 33 as P,

$$\begin{aligned} f_x(x) = \frac{dP}{dx} = \frac{dP}{d\ln(t)} &= \frac{dt}{d\ln(t)} \cdot \frac{dP}{dt} \\ &= tf(t) \\ &= \Sigma\, a_i \tau_i^{-1} \exp(x - \tau_i^{-1} e^x) \end{aligned} \tag{34}$$

er than at $t = 0$ because a resolution
elow), so there are no observations
gle slowly decaying component with
of which is cut off on the display,
ne data are shown again in Fig. 15C,
h a bin width of 10 μs; the 60-μs
times longer than 250 μs of course,
d end of the histogram (the top of
a single exponential, but this time
from either of these displays, there
stribution. It is visible only in the
mes up to 8 ms are shown (with a
of about 1 ms is visible. The data
ne fit was done to all the data (by
onential components (dashed line)
shown in all four sections of Fig
nly in the display up to 8 ms.
nient for intervals that cover such
next is preferable.

igworth and Sine (1987) that it
e (as in the preceding example),
rval rather than the distribution
og transformation of the x axis
with no peak, and would have
suggested, in addition, the use
keep the errors approximately

f an interval is denoted t, and

e that if a t is less than some
n $\ln(t_1)$. Thus,

$$P \tag{33}$$

the same. Now it is pointed
found by differentiating the
in equation 33 as P,

$$^{-1}e^{x}) \tag{34}$$

The second line here follows because dP/dt is simply the original distribution of time intervals, $f(t)$; it shows, oddly, that the distribution of $x = \ln(t)$ can be expressed most simply not in terms of x but in terms of t. When $f(t)$ is multiexponential, as defined in equation 30, and we express $f_x(x)$ in terms of x by substituting $t = e^x$, we obtain the result in equation 34. This function is not exponential in shape but is (for a single exponential component) a negatively skewed bell-shaped curve, the peak of which, very conveniently, occurs at $t = \tau$.

The same data that were displayed in Fig 15A, B, and C are shown in Fig 15D as the distribution of log(shut times). The same fitted curves are also shown (the fitting uses the original intervals, not their logarithms), and the three-component fitted curve shows peaks that occur at the values of the three time constants. It is now clearly visible, from a single graph, that the two-exponential fit is inadequate. (The slow component of the two-exponential fit also illustrates the shape of the distribution for a single exponential because it is so much slower than the fast component that the two components hardly overlap.) This sort of display is now universally used for multicomponent distributions. Its only disadvantage is that it is hard, in the absence of a fitted line, to judge the extent to which the distribution is exponential in shape.

5.1.4. The Cumulative Distribution

The area under the pdf up to any particular value, t, of the time interval is the cumulative form of the distribution, or *distribution function,* namely

$$F(t) = P(\text{time interval} \leq t) = \int_0^t f(t)dt = 1 - e^{-t/\tau} \tag{35}$$

This is a probability and is dimensionless; it increases from 0 to 1 as t increases. Alternatively we may consider the probability that an interval is *longer than t,* which is, for a single exponential,

$$1 - F(t) = P(\text{interval} > t) = \int_t^\infty f(t)dt = e^{-t/\tau}$$

or, for more than one component, the sum of such integrals:

$$1 - F(t) = P(\text{interval} > t) = \Sigma a_i e^{-t/\tau_i} \tag{36}$$

Occasionally, the data histogram is plotted in this cumulative form with the fitted function (36) superimposed on it. This presentation will always look smoother than the usual sort of histogram (the number of values in the early bins is large), but it should *never* be used, because the impression of precision that this display gives is *entirely spurious.* It results from the fact that each bin contains all the observations in all earlier bins, so adjacent bins contain nearly the same data. In other words, successive points on the graph are not independent but are strongly correlated, and this makes the results highly unsuitable for curve fitting.

To make matters worse, it may well not be obvious at first sight that cumulative distributions have been used, because the curve, equation 36, has exactly the same shape as the pdf, equation 30. There are no good reasons to use cumulative distributions to display data; they are highly misleading. In any case, it is much easier to compare results if everyone uses the same form of presentation.

5.1.5. Superimposition of a Probability Density Function on the Histogram

It is helpful to regard the ordinate of the histogram not as a dimensionless number but as a "frequency" or "number per unit time" with dimensions of reciprocal time; the ordinate then becomes directly analogous to probability density. Rather than regarding the height of the histogram block as representing the number of observations between, say, 4 and 6 ms, we regard the *area* of the block as representing this number. The ordinate, the height of the block, will then be the number per 2-ms bin. This is illustrated in Fig. 16 for a hypothetical example of a simple exponential distribution of open time durations with mean $\tau = 10$ ms and rate constant $\lambda = 1/\tau = 100\ s^{-1}$. The pdf is thus $f(t) = 100e^{-100t}\ s^{-1}$. It is supposed that there are $N = 494$ observations altogether (including those that might be too short to be seen in practice—see Section 6.1).

The histogram is plotted with a bin width of 2 ms, so the ordinate is number per 2-ms bin. The pdf has, of course, unit area. In order to obtain a curve that can be superimposed on the histogram, we must multiply the pdf by the total number of events and convert its units from s^{-1} to $(2\ ms)^{-1}$ by dividing by 500. The continuous curve is therefore $g(t) = (494/500)f(t) = 98.8e^{-100t}\ (2\ ms)^{-1}$. The number of observations that are expected between 4 and 6 ms is the area under the continuous curve; i.e., from equation 35 or 36, it is $494(e^{-4/\tau} - e^{-6/\tau}) = 60.6$. This is almost the same as the ordinate of the continuous curve at the midpoint ($t = 5$ ms) of the bin: $g(t) = 98.9e^{-5/\tau} = 59.9$ (per 2 ms). This approximation will always be good as long as the bin width is much less than τ. Thus, if we actually observed the expected number of observations (60.6) between 4 and 6 ms, the histogram bin would fit the continuous curve closely, as shown in Fig. 16.

Generalizing this argument, the function, $g(t)$, to be plotted on the histogram is

$$g(t) = Nd\,f(t) \tag{37}$$

where $f(t)$ is the probability density function, with units s^{-1} (estimated by fitting the data as described in Section 6), d is the bin width (with units of seconds), and N is the estimated total number of events as calculated by equations 87, 91, or 101, as appropriate. Note that

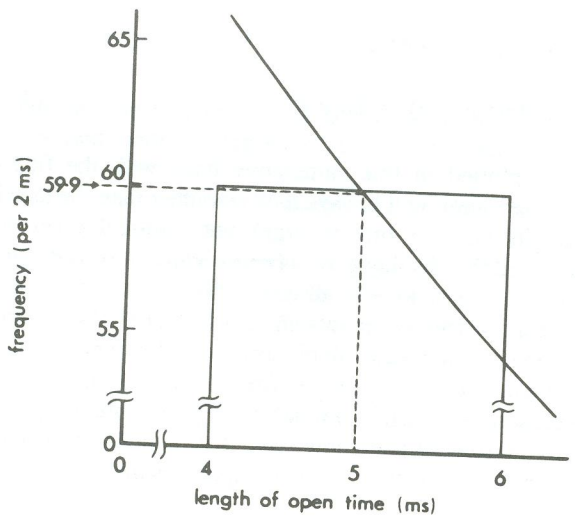


Figure 16. Schematic illustration of the superimposition of a continuous curve (proportional to the fitted theoretical pdf) to a histogram of observed frequencies. The block corresponds to 60 observations between 4 and 6 ms and has an area equal to that under the continuous curve between 4 and 6 ms. The ordinate of the continuous curve at the midpoint of the bin ($t = 5$ ms) is 59.9 $(2\ ms)^{-1}$. See text for further details.

equation 37 is dimensionless, so it is really the pdf that is scaled to the data rather than the other way around.

In the case where the log(interval length) is displayed, as described in section 5.1.3, the probability density function, $f(t)$, would usually be fitted, as described in Section 6, by the method of maximum likelihood applied to the original observations (not to their logarithms). The distribution of $\log_{10}(t)$ is, from equation 34, $2.30259tf(t)$, where the factor 2.30259 [= ln(10)] converts from natural logarithm units to common logarithm units. The curve, $g(t)$, to be plotted on the logarithmic histogram is thus

$$g(t) = Nd' \, 2.30259 \, tf(t) \tag{38}$$

where d' denotes the bin width in $\log_{10}$ units.

5.1.6. Variable Bin Width

The approach discussed above makes it immediately clear how one should construct a histogram with unequal bin widths. It is sometimes useful to use a narrower bin width for shorter intervals than for long ones (there are usually more short intervals, and the pdf changes most rapidly in this region). Thus, if the ordinate is specified as, for example, frequency per 2 ms, then the height of the ordinate for a bin width of 2 ms (say the bin for 6 to 8 ms) is the actual number of observations found to fall within this bin. However, if the shorter intervals are plotted with a bin width of 1 ms rather than 2 ms, then the height of the ordinate for the the 1-ms-wide bins should be twice the number actually observed to fall into the bin. Thus, the area still represents the actual number observed. The plotted function is still as given in equation 37 above, but d is now interpreted as the base width of the bins, i.e., 2 ms in this example, because the ordinate is the frequency per 2 ms bin.

5.1.7. Measurement of P_{open}

One often wishes to measure the probability that a channel is open from a single-channel record. This quantity is usually denoted P_{open} and is sometimes called the "open probability." It is undesirable to refer to P_{open} as the probability of opening, because this sounds like a rate constant (probability of opening in a short time interval; see Chapter 18, this volume), which is not what is intended.

Measurements of P_{open} are useful as an empirical index of the activity in a record, though the overall P_{open} for a whole record will often be so distorted by long sojourns in desensitized or inactivated states as to be uninterpretable. More fundamentally, if it is possible to identify the parts of the record when channels are desensitized, then measurements of P_{open} on the remaining sections provide the best means of constructing equilibrium concentration–response curves (e.g., Colquhoun and Ogden, 1988). Such P_{open} curves have the advantages over other methods that (1) they are corrected for desensitization, (2) they measure response on an *absolute* scale (the maximum possible response is known in advance to be 1), and (3) they allow direct inspection of the channels that underlie the response so there can be little doubt about their identity and homogeneity (see Section 5.9 for tests of homogeneity).

In a record that is in the steady state, P_{open} is simply the average fraction of time spent in the open state. An absolute value for P_{open} can, however, be measured only from a record that contains only one individual channel (or from a section of a record, such as a burst or

cluster, where only one channel is active; see Sections 5.6 and 5.9). However, for the purposes of assessment of stability (Section 5.1.1), this is not really important.

When all of the open and shut times have been measured, P_{open} can be calculated as total open time divided by total length of the record. For records where there is essentially only one open level, this is the same thing as the average current level throughout the record, divided by the open-channel current level. In this case, the best method to measure P_{open} is to integrate the record (with an analogue integrator circuit or digitally). This is a good method in principle because the record is filtered, and linear filters do not affect the *area* of the response, only its shape, so integration should be unaffected by the imperfect resolution of open and shut times. Use of digital integration is equivalent to the use of point-amplitude histograms to measure P_{open}, as described in Section 5.3.2. It is important to notice, however, that integration will be satisfactory *only* as long as adequate allowance can be made for the drift in the baseline (shut) level that occurs in most real records.

When the system is not in a steady state, P_{open} will be a function of time and can no longer be defined as the average fraction of time spent in the open state. This is the case, for example, following a voltage or concentration jump or during a synaptic current. In such cases, $P_{open}(t)$ must be measured by repeating the jump many times and measuring the fraction of occasions when the channel is open at time t.

5.2. Missed Events: Imposition of a Consistent Time Resolution

Unless the mean length of an opening is very long compared with the minimum resolvable duration, it is inevitable that some short openings will remain undetected. Similarly, some short shuttings will also be missed. Methods for making appropriate allowances or corrections for such missed events are considered briefly in Section 6.11 and in rather more detail in Chapter 18 (this volume). In this section we discuss only the aspects of the problem that require action to be taken *before* histograms are constructed.

5.2.1. Definition of Resolution

When the single-channel record is scanned to fit the time of each opening and shutting, as discussed in Sections 3 and 4, the usual procedure would be to fit every detectable opening and gap (shut time). The length of opening (or gap) considered "detectable" will depend on the sort of detection method used. For the threshold-crossing analysis described in Section 4.1, the minimum length is set by T_d, although observed durations up to about twice this value are biased and need to be corrected (e.g., with equation 21) before insertion into a histogram. With time-course fitting, the minimum length is not clearly defined and will certainly depend on the details of the method that is used, on who the operator is, and, quite possibly, on how tired he or she is. This will not matter too much as long as care is taken to fit everything that might possibly be an opening or shutting, so that when a realistic resolution is subsequently imposed (Section 5.2.3), it can be said with certainty that events longer than this chosen resolution will not have been omitted during the fitting process.

5.2.2. Effects of Missed Events

Consider, for example, the distribution of the open time when there is a substantial proportion of undetected short gaps; openings will appear to be longer than they actually

are, because two (or more) openings separated by an undetected gap will be counted as a single opening (the measured open times are, therefore, more properly referred to as *apparent* open times).

When the histogram of shut or of open times is plotted, the frequency will tend to fall off for very short durations, below which some or all events are too short to be detected. Thus, the distribution may appear to have a peak. One way to deal with this is to look at the histogram and decide on a duration above which it is thought that all openings will be detected and accurately measured; only observations that are longer than this minimum time are used for the fitting process. There is, of course, a large arbitrary element in this decision (and it is also always possible that the open time distribution really does go through a maximum; see Chapter 18, this volume). Nevertheless, if the value chosen is on the safe side, this method may seem to be satisfactory. But it is actually fundamentally inconsistent, as becomes clear when we consider the effect of the open-time resolution on the shut-time distribution.

One way in which inconsistency arises becomes obvious when we consider fitting of *shut* times. If we look at the histogram and see that it has a peak near 100 μs but falls off for shorter shut times, we may decide, quite reasonably, that it is safe to fit (see Section 6) all shut times longer than, say, 140 μs. However, the shut times shorter than 140 μs are still present in the data, and even though they have just been deemed to be too short to be reliable, they will still be regarded as separating two openings, and will therefore, despite their unreliability, shorten the apparent open time to a lower value than it would have if the short gaps had not been detected at all. And, of course, an exactly analogous inconsistency in measurement of apparent shut times can arise when short openings are partially missed.

Another sort of inconsistency will arise if the criterion for the gap length that is detectable does not remain exactly the same throughout the analysis. If it is not constant, the apparent lengths of openings will vary with time, so the distribution of the measured open times will be distorted even if *all* the openings are long compared with the minimum resolvable duration.

A third reason why it is important to know about the resolution is encountered when, for example, measurements of open times are made at different membrane potentials. The resolution for, say, brief shuttings, will be worse when the single-channel currents are smaller (potentials closer to the reversal potential), so more of them will be missed. The apparent open times will therefore appear to be longer at potentials near the reversal potential, even if the true open time does not depend on membrane potential at all.

Finally, the methods that have been developed recently for making corrections for missed brief events almost all require that the resolution of the data be known and consistent. In other words, if the resolution is stated to be 100 μs, then we must be as sure as possible that no events shorter than this are present in the data, and that all events longer than 100 μs have been detected.

For all of these reasons, it is important that the resolution (the shortest event fitted) be stated in published work; without knowing the resolution, it is impossible for other authors to compare their results for quantities as mean "apparent open time" (though this rarely stops them from trying).

5.2.3. Imposition of Resolution

One way to avoid the inconsistencies just described is to impose a resolution on the data retrospectively (Colquhoun and Sakmann, 1985). In the analysis of the original experi-

mental record, every event is fitted even if it is so short that its reality is dubious. While this is done, a judgment is made as to the shortest duration (t_{res} say) that can be trusted (the value of t_{res} may not be the same for open times and for shut times). Again, this is quite subjective; a value on the safe side should be chosen. The most important criterion for the choice of t_{res} is that it should be chosen so that it ensures a sufficiently low false-event rate, e.g., below 10^{-8} s^{-1} (see Section 3.3).

When the analysis is completed, and the idealized record is stored (see Section 4), the chosen value of t_{res} can be specified and the idealized record revised as follows:

1. All open times that are shorter than t_{res} must be removed. Shut times that are separated by openings shorter than t_{res} are treated as single shut periods. The lengths of all such shut times are concatenated (together with the lengths of intervening short openings) and inserted in the revised data record as single shut times.
2. Similarly, all shut times that are shorter than t_{res} must be removed. If the two openings that are separated by the short gap have both got the same amplitude, then the two open times are concatenated (together with the intervening shut time) and inserted into the revised record as a single opening. If the two openings have different amplitudes, they are inserted into the revised record as two openings with a direct transition from the first open level to the second. This procedure entails deciding exactly what "the same amplitude" means. Some criterion must be specified, which will depend on what amplitude difference is deemed large enough to be detectable; for example, amplitudes that are separated by less than 10% of the full amplitude might be deemed "the same."

In this way a new idealized record, with consistent time resolution throughout, is produced, and it is this that is used for subsequent construction of histograms and fitting. The new record cannot, of course, contain any openings (or gaps) shorter than t_{res}, so the histograms start at this point. As long as the original idealized record is kept, it is easy to repeat the fitting with a different resolution if necessary.

It may be noticed that, for example, imposition of a 50-μs resolution on a perfect record, followed by imposition of 100-μs resolution, will not necessarily give exactly the same result as imposition of 100-μs resolution directly on the perfect record. To the extent that the data we start with are never perfectly resolved, this approach does not give precisely the required results, but it is, nevertheless, the best that can be done.

5.2.4. Resolution, Sublevels, and Fit Range

It must be remembered that events (openings or shuttings) may be *detected* with certainty in the single-channel record even when their duration is shorter than the risetime (T_r) of the recording system. However, their duration must be at least $2T_r$ before their amplitude can be measured accurately (see Section 4). If, for example, it is desired to construct a distribution of the apparent times but to include in the distribution only those open times that are sufficiently long for their amplitudes to be known, then only openings longer than $2T_r$ or $2.5T_r$ can be used. However, this does *not* mean that the resolution of $2T_r$ should be imposed on the data. If this resolution were imposed on the shut times, many brief shuttings, which are nevertheless long enough to be detected with certainty, would be excluded, thus causing the apparent open times to be longer and causing unnecessary error in the estimation of the open time. The resolution that is imposed should depend on what can be *detected* reliably (i.e., distinguished from random noise), but, in the case just described, the range of values that are used for *fitting* should exclude values shorter than $2T_r$. When conditional distributions

are used for maximum-likelihood fitting, as described below, there is no problem in fitting only those observations that lie within any specified range.

The distinction between resolution and fit range does not cause too many problems when there are no subconductance levels in the record. In this case, any deflection toward the baseline must represent a complete shutting. But when the channel shows subconductance levels, the problem is more difficult, and it may be desirable to impose different resolutions and/or different filtering for different sorts of analysis.

For example, Howe *et al.* (1991) describe procedures for analysis of NMDA-type glutamate-activated channels that showed conductance levels of 30 pS, 40 pS, and 50 pS. A resolution that produces an acceptable false-event rate for the 50-pS openings may result in an unacceptably high false-event rate for smaller openings in the same record. For analysis of amplitudes, the results were treated as described above; the resolution was set to produce an acceptable false-event rate for 50-pS events, but events shorter than $2.5T_r$ were excluded from fitting. For distributions of shut times the open-time resolution was set to give an acceptable false-event rate for 50-pS openings, but the shut-time resolution was set to ensure that events described as shuttings were unlikely to be transitions from 50 pS to 40 or 30 pS or transitions from 40 pS to 30 pS. To achieve this, the resolution was set to duration w, such that events are counted as shuttings only if they are seen to reach a level safely (say 2 standard deviations) below the 30-pS level. This can be achieved by solving for w (e.g., by bisection)

$$\frac{A_{50} - (A_{30} - 2s_{30})}{A_{50}} = \text{erf}(0.886w/T_r)$$

The right-hand side of this equation gives the fraction of its maximum amplitude attained by a rectangular pulse of length w (see equations 8 and 12). On the left hand side, A_{50} and A_{30} are the absolute current amplitudes for the 50-pS and 30-pS openings, and s_{30} is the standard deviation of the 30-pS currents (these values being obtained from fitting of amplitude histograms). For further details, see Howe *et al.* (1991).

5.3. The Amplitude Distribution

Single-channel current amplitudes are interesting for two main reasons. First, the ways the amplitude varies with ionic composition of the bathing medium and with membrane potential are important for the study of ion permeation mechanisms. Second, amplitude measurements are often a useful way to characterize channel types, e.g., types with different subunit compositions or with mutations.

It has been shown in Section 4.2.3 that the amplitude of a channel opening can be measured accurately only if the duration of the opening is at least twice the risetime (T_r) of the recording system. Amplitude measurements should, therefore, be included in the amplitude histogram only when the opening is longer than some specified length such as $2T_r$ or $2.5T_r$. This can, of course, be done properly only if the amplitude is estimated separately for every opening, and there are, unfortunately, still many analysis programs in use that cannot do this.

Often there will be more than one channel in the patch of membrane from which the recording is made, and in this case, more than one channel may be open at the same time, so that current amplitudes that are integer multiples of the single-channel current are seen. This question is discussed further in Section 5.4.

5.3.1. How Variable Are Single-Channel Amplitudes?

The amplitudes of single-channel currents are, in some cases at least, very consistent. For example, Fig. 17 shows a distribution of amplitudes measured from adult rat endplate nicotinic acetylcholine receptors. It has been fitted (arbitrarily) with a Gaussian curve and shows a mean of 6.62 pA and a standard deviation of 0.12 pA (i.e., 1.8% of the mean). The variability from one opening to the next of the same ion channel or of different channels in the patch seems to be very small, possibly no greater than the error in the fitting of the amplitude. In this case the amplitude is not an inherently random variable like the open time but is, for practical purposes, a more or less fixed quantity.

However, channel amplitudes more commonly are not exactly constant. It seems that just about every sort of channel shows extra open-channel noise; i.e., the current record is somewhat noisier when the channel is open than when it is shut (e.g., Sigworth, 1985, 1986). If it is assumed that the excess open-channel noise is independent of the baseline noise, so their variances are additive, the root mean square excess noise, s_{excess}, can be estimated as

$$s_{excess} = (s^2_{open} - s^2_{shut})^{0.5} \tag{39}$$

The extent of this excess open-channel noise varies greatly from one sort of channel to another; it is very small for adult frog muscle nicotinic receptors (D. Colquhoun, unpublished

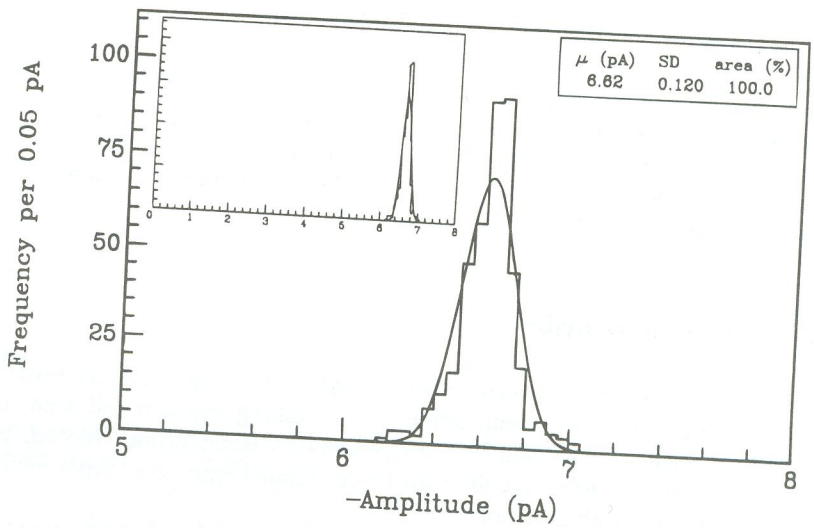


Figure 17. An example of the distribution of the fitted amplitudes of single-channel currents; amplitudes were defined by eye, by means of a cursor on the computer screen. Unpublished data of D. C. Ogden and N. K. Mulrine; channel openings elicited by 100 nM acetylcholine in cell-attached patch on adult (200-g) rat endplate in extracellular solution (with 20 mM K^+ and 1 mM Ca^{2+}), at resting potential −80 mV. After a resolution of 50 μs was imposed for both openings and shuttings, there were 1100 resolved intervals, and the histogram was constructed from 433 amplitudes of openings that were longer than 2 risetimes. The continuous curve is a Gaussian distribution, which was fitted to the data by the method of maximum likelihood; it has a mean of 6.62 pA and a standard deviation of 0.12 pA. The main display covers only the range from 5 pA to 8 pA in order to show clearly that the observed distribution has a sharper peak and broader tails than the Gaussian curve, as predicted in Section 5.3.3 and Appendix 2. The inset shows the same distribution plotted over the range 0–8 pA to show that there is only one narrow peak in the distribution.

data) but very large for some neuronal nicotinic receptors (e.g., Mathie *et al.,* 1991). It is likely that the phenomenon is intrinsic to the receptor protein; it appears in recombinant receptors and can be strongly influenced by small mutations. It seems likely that it can be regarded as resulting from fluctuations in channel structure that produce small changes in conductance or from entry into subconductance states that are short-lived and/or differ only slightly from the main conductance level. The appearance of extra open-channel noise can also be mimicked by frequent and brief channel blockages (e.g., Ogden and Colquhoun, 1985).

It is common for more than one conductance level to appear in single-channel recordings. One, probably quite common, reason for this is heterogeneity of the channels in the membrane patch. In addition, though, it has become apparent that most types of ion channel have more than one conductance level. For some types these conductance sublevels are rare, but for others they are quite common. For example, the NMDA-type glutamate receptors all show this phenomenon clearly, as illustrated in Figs. 12 and 18. These channels have a 50-pS main level and a briefer 40-pS sublevel. It is not known whether such sublevels have any functional importance (though it seems unlikely), but they are certainly useful for characterizing subunit combinations (Stern *et al.,* 1992).

In this case of NMDA receptors, the 50-pS and 40-pS peaks are quite clear and reproducible from experiment to experiment. There is, however, some question as to whether all "50-pS" openings have exactly the same conductance (apart from random measurement errors). There is some reason to suspect that they may not.

The various methods that are used for investigation of amplitudes are discussed next.

5.3.2. Point-Amplitude Histograms

The simplest procedure is to make a histogram of the values of the individual digitized data points (after subtracting the baseline value, so the shut channel appears with zero amplitude). This is often known as a *point-amplitude histogram* to distinguish it from histograms formed from fitted amplitudes (see Section 5.3.3). There will be a lot of points in such histograms, but the points are not independent, so the large number of points does not necessarily imply high precision. In order for the sample points in filtered data to be approximately independent, they would need to be about one risetime (T_r) apart, but the sample rate is normally a good deal higher than this. For statistical purposes, the "effective number of points" could taken roughly as (sample duration)/T_r.

The relative areas of the peaks in a point-amplitude histogram represent the number of data points, i.e., the length of time spent, at each amplitude level (cf. next section). The areas of the peaks can therefore be used to estimate the fraction of time for which the channel is open, i.e., the probability of being open (P_{open}), as long as all data points are included (see also Section 5.1.7).

5.3.2a. ***The All-Point-Amplitude Histogram.*** The crudest method is simply to make the histogram directly from all points in the data record. In fact, this is the *only* method that is available in many commercial programs. The main problem is that the method obviously depends on the baseline remaining exactly constant throughout the record. This is rarely true, so in practice it is possible to use the method only on relatively short stretches of data for which the baseline can be checked carefully. Alternatively, if both baseline and open levels have been fitted, as illustrated in Fig. 12, all the data points in the region that has been fitted (and approved) can be entered into the histogram; this provides excellent compensation for baseline changes, but the results cannot be used to estimate P_{open} because many shut points are omitted.

An example of an all-point histogram constructed in the latter way is shown in Fig. 18B and C. The peaks for the shut level and for the main (about 5-pA) open level are obvious. However, there is a smear of points between the two (the data points that lie in the transition regions between open and shut), and this partially obscures the small peak that corresponds to the sublevel at about 4 pA; this is shown on an enlarged scale in Fig. 18C. This smearing can be reduced as follows.

5.3.2b. ***Open-Point and Shut-Point Amplitude Histograms.*** Once transitions have been located by one of the methods described in Section 4, then it becomes possible to exclude data points that lie on the transitions from one conductance level to another. Knowledge of the step-response function of the recording system allows the transition period to be defined accurately. An example is shown in Fig. 12; only those data points that correspond to the flat sections of the fitted curve (i.e., areas where no transitions were detected) are entered into the histogram. The open-point amplitude histogram in Fig. 18E was constructed in this way. Most of the smearing has gone, and the rather small 4-pA component is more clearly defined than in the all-point histogram, as shown on an enlarged scale in Fig. 18F. And, since the baseline adjacent to the openings is fitted along with the openings, there should be no distortion caused by baseline drift.

The data points that correspond to shut periods are entered into a separate histogram, as for the open points. A shut-point histogram is shown in Fig. 18D; it is usually found, as in this case, that the shut-point histogram is fitted very well by a simple Gaussian curve (i.e., the baseline noise is Gaussian). Open-point histograms, on the other hand, may not be perfectly Gaussian because of such effects as undetected sublevel transitions or brief closures.

5.3.2c. ***Analysis of Flickery Block.*** The asymmetry in point-amplitude histograms contains information about the nature of open-channel noise. This information can be interpreted by use of either noise analysis (Sigworth, 1985, 1986; Ogden and Colquhoun, 1985; Heinemann and Sigworth, 1990) or the amplitude histogram itself (Yellen, 1984; Heinemann and Sigworth, 1991). These methods have been used, for example, to analyze rapid channel block. High concentrations of a low-affinity channel-blocking agent produce so-called "flickery noise." Because of the high concentration, blockages are frequent, and openings are short, and when blockages are so brief that they cannot be resolved easily in the single-channel record, the open channel appears to be very noisy and to have a reduced amplitude (see Chapter 18, this volume). Such flickery noise, when it happens to be in the right frequency range, will produce a characteristically shaped smear in the all-point amplitude histogram. If the blocking process is approximated as a two-state process, and we look at the channel only while it is open or blocked, the mechanism can be written thus

$$\text{open} \underset{k_{-B}}{\overset{k_{+B}x_B}{\rightleftharpoons}} \text{blocked}, \tag{40}$$

where k_{-B} is the dissociation rate constant for the blocker, k_{+B} is the association rate constant, and x_B is the blocker concentration, so $k_{+B}x_B$ is the transition rate from open to blocked state (per unit open time).

One approach, which works best for events that are close to being resolvable (mean duration comparable to the filter risetime), is based on the work of Fitzhugh (1983). This theory showed that, for data that have been filtered through a simple RC filter with time constant $\tau_f = RC$, the point-amplitude histogram should be described by the beta distribution. The beta distribution was used to analyse fast block by Yellen (1984). If we denote as y the

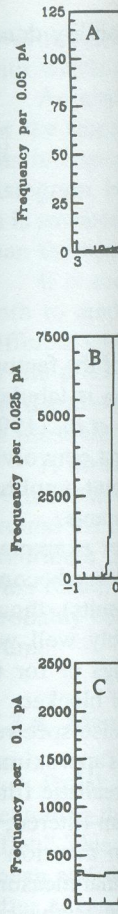


Figure 18. Exa
channels (same
shuttings and 4
0.5 pA. A: Distr
of less than two
6.0 pA) with a
values. The cor
seen in 1 mM
0.36 pA and 0.
amplitude of al
effects of basel
"40-pS" compo
inevitable in an
points that corr
for regions whe
the transition fr
deviation 0.12
demarcation of
is not perfect, t
fitted amplitude

ed in the latter way is shown in Fig.
e main (about 5-pA) open level are
the two (the data points that lie in the
artially obscures the small peak that
wn on an enlarged scale in Fig. 18C.

Histograms. Once transitions have
4, then it becomes possible to exclude
ance level to another. Knowledge of
s the transition period to be defined
data points that correspond to the
nsitions were detected) are entered
in Fig. 18E was constructed in this
ll 4-pA component is more clearly
enlarged scale in Fig. 18F. And,
g with the openings, there should

ntered into a separate histogram,
Fig. 18D; it is usually found, as
well by a simple Gaussian curve
s, on the other hand, may not be
level transitions or brief closures.
y in point-amplitude histograms
This information can be interpre-
; Ogden and Colquhoun, 1985;
itself (Yellen, 1984; Heinemann
ample, to analyze rapid channel
agent produce so-called "flick-
re frequent, and openings are
resolved easily in the single-
d to have a reduced amplitude
it happens to be in the right
ear in the all-point amplitude
state process, and we look at
can be written thus

(40)

he association rate constant,
rate from open to blocked

to being resolvable (mean
of Fitzhugh (1983). This
imple RC filter with time
ed by the beta distribution.
84). If we denote as y the

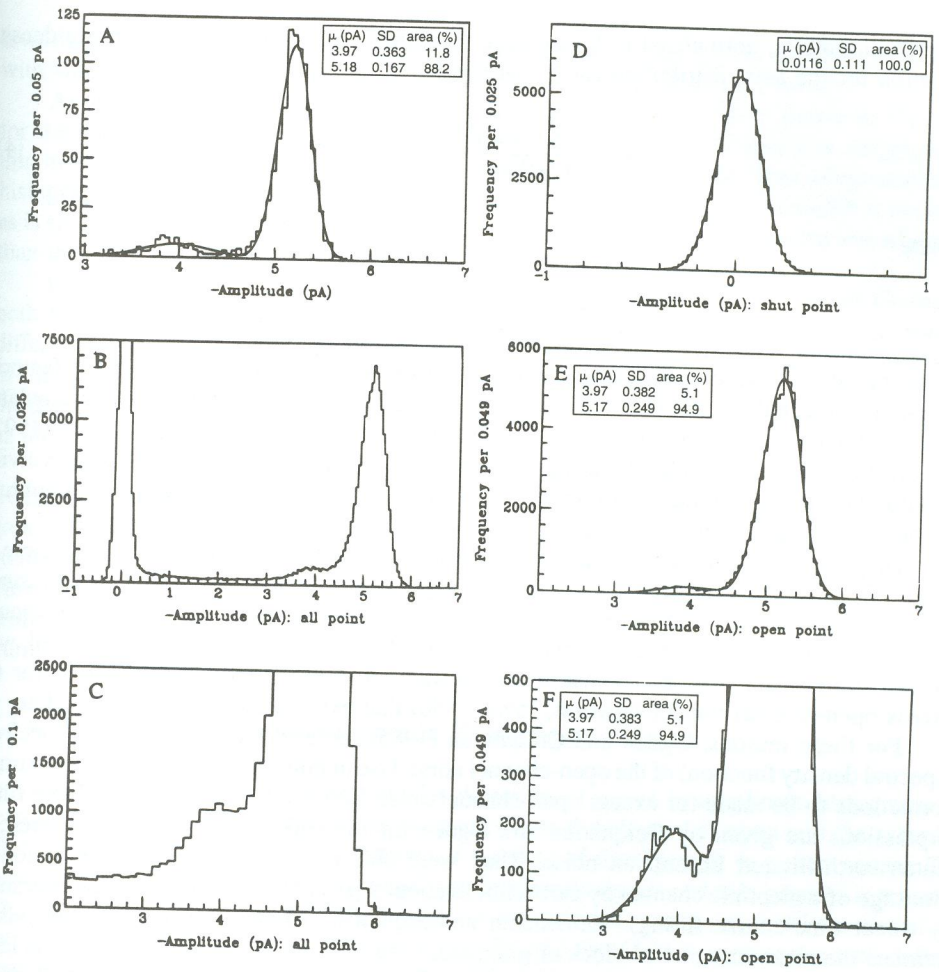


Figure 18. Examples of various sorts of amplitude histograms. Data were from a 10-s recording of NMDA channels (same data as were illustrated in Fig. 12, where details are given). Resolution was set to 30 μs for shuttings and 40 μs for openings, with concatenation of contiguous open levels that differed by less than 0.5 pA. A: Distribution of fitted amplitudes (of the type listed in legend of Fig. 12). Openings with a duration of less than two risetimes (332 μs) were excluded, which left 1049 amplitudes to be fitted (between 3.4 and 6.0 pA) with a mixture of two Gaussian distributions by maximum-likelihood method using the original values. The components had means of 3.97 pA and 5.18 pA (the usual "40-pS" and "50-pS" components seen in 1 mM Ca). The areas of the components were 11.8% and 88.2%, and the standard deviations were 0.36 pA and 0.17 pA, respectively. B and C: All-points amplitude histogram. This histogram shows the amplitude of all data points within the fitted range (solid line in Fig. 12). This ensures freedom from the effects of baseline drift but means that the relative area occupied by the shut points is arbitrary. The small "40-pS" component is shown on an enlarged scale in C; this also makes more obvious the smearing that is inevitable in an all-points histogram. D, E, and F: Separate open-point and shut-point histograms. The data points that correspond to the regions where the fitted curve (see Fig. 12) was flat were collected separately for regions where the channel was shut and where it was open. This eliminates the smeared points during the transition from shut to open. The shut-point histogram in D is well fitted with a single Gaussian (standard deviation 0.12 pA). The open-point histogram in E (and, on an enlarged scale, in F) shows much clearer demarcation of the subconductance level than the all-points histogram. The fit with two Gaussian components is not perfect, though the fitted means, 3.97 pA and 5.17 pA, are almost identical to those found from the fitted amplitudes in A.

current amplitude, normalised to lie between the values of 0 and 1, the probability density function for the beta distribution can be written as

$$f(y) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{(a-1)}(1-y)^{(b-1)} \tag{41}$$

where

$$a = k_{-B}\tau_f \quad \text{and} \quad b = k_{+B}x_B\tau_f.$$

In this result, $\Gamma(x)$ denotes the gamma function, which is a continuous version of the factorial function, such that $\Gamma(x+1) = x!$ when x is an integer. The gamma function is tabulated by Abramovitz and Stegun (1965) or can be computed as described by Press *et al.* (1992). This result does not include the background noise in the recording, but by first convolving it with the baseline noise distribution and then fitting the result to the point-amplitude histogram, one can obtain estimates of the blocking and unblocking rate constants.

The beta distribution method makes some undesirable assumptions. For example, a simple *RC* filter is never used in practice. For any more realistic filters, the theory becomes a great deal more complicated (A. Jalali and A. G. Hawkes, unpublished results), though Yellen (1984) showed that use of the beta distribution could work reasonably well with Bessel-filtered data. Second, it makes no allowance for spontaneous shuttings or for the excess open-channel noise (Sigworth, 1985, 1986) that exist in the absence of blocker.

For these reasons, Ogden and Colquhoun (1985) preferred to use the noise spectrum (spectral density function) of the open-channel noise. Use of noise analysis allows approximate corrections to be made for excess open-channel noise, and it allows use of a realistic filter; expressions are given by Colquhoun and Ogden for the variance of Gaussian-filtered, or Butterworth-filtered Lorentzian noise. They were able to estimate the mean duration of blockage of a nicotinic channel by carbachol as about 9 μs (which is similar to that measured by direct time-course fitting). Heinemann and Sigworth (1990) used a noise analysis to estimate the mean duration of block of gramicidin channels by Cs^+ as about 1 μs. This latter value was confirmed by Heinemann and Sigworth (1991) by inspection of the cumulants of the point-amplitude distribution. The cumulant method provides a method of analysis of point-amplitude histograms that is complementary to the beta-function approach. The beta function works best with events that are comparable with the filter risetime, but the cumulant method is better for events that are much shorter than the risetime (but are widely spaced).

5.3.3. Amplitude Histograms from Fitted Amplitudes

The other main method for display of single-channel amplitudes is to measure the amplitude of each opening separately and to make a histogram of the results. The amplitudes can be measured by placing a cursor on the data on the computer screen, by eye, or by using a least-squares fit to the data as illustrated in Fig. 12. In either case, the amplitude can be measured only for events that are longer than about two risetimes, as explained in Section 4.2.3. And in either case, the estimates are susceptible to bias resulting from undetected brief closures. The latter problem can be minimized by fitting all possible closures, even if they are so short that they will eventually be eliminated when a safe resolution is imposed (see Section 5.2.3). The histogram has only one value for each opening, so if more than one open

level is present, the relative areas of the components will represent the relative *frequencies* with which the levels occur (rather than the relative time spent at each level).

An example of an amplitude histogram constructed in this way is shown in Fig 18A, for the same data that were used to illustrate the point-amplitude histograms. In principle, this histogram should show components even more clearly than the open-point-amplitude histogram, because smearing between transitions from one open level to another is avoided, as is smearing between open and shut levels. It is, on the other hand, somewhat less objective than the open-point-amplitude histogram.

It is usual to fit (as described in Section 6.8.3) a Gaussian or a mixture of Gaussians both to amplitude histograms and to point-amplitude histograms in an attempt to resolve different conductance levels. In fact, this may not be appropriate in either case (as discussed in the preceding section for point-amplitude histograms). In the case of fitted amplitudes, the distribution often shows a sharper peak and broader tails than is expected for a Gaussian, as illustrated in Fig. 18A or, particularly clearly, in Fig. 17.

A distribution of this sort is to be expected because the amplitude values are obtained from events of variable duration. The long events give the most precise estimates and cluster around the true value to give the sharp peak. Short events give values with more scatter and contribute to the tails. The distribution that would be expected is derived in Appendix 2. This result, although preferable to Gaussian fits, has not yet been used much in practice, probably because of the inconvenience involved in determining the background noise spectrum.

5.3.4. Mean Low-Variance Amplitude Histograms

Patlak (1988) suggested a method for detection of peaks in amplitude histograms by searching the digitized data record for sections where the channel is open and the record is "flat." This is done by looking at sections of the data of fixed length (e.g., ten points). The mean and standard deviation of each such section is calculated, and this process is repeated after advancing the start of the data section by, for example, one point, until the end of the data is reached. A data section is deemed to be flat and therefore to represent a well-defined conductance level if its standard deviation is less then some specified multiple (e.g., 0.5 to 2) of the standard deviation of the baseline noise. All sections that have a larger standard deviation than this are rejected, and a histogram is constructed of the mean amplitudes of the remaining sections. Three different values have to be specified to construct the histogram (the section length, the number of points to advance, and the threshold standard deviation), so a variety of histograms can be produced. This method may work well on some sorts of data, especially if the conductance levels are reasonably long-lived. However, use of sufficient filtering will make long-lived conductance levels obvious by any method of analysis. It is much harder to distinguish subconductance states that are short-lived. If the data sections are made shorter, their increased scatter means the histogram is more scattered, so peaks become hard to distinguish. And if the threshold standard deviation for inclusion is made lower, the number of points in the histogram is reduced, so again peaks become hard to distinguish with certainty.

5.3.5. Subconductance Transition Frequencies

When there is more than one open-channel conductance level, it may be of interest to measure the frequency of transitions from one open level to another (and from each open

level to the shut level). This provides another way to characterize quantitatively different receptors or subunit combinations (Howe *et al.*, 1991; Stern *et al.*, 1992). It can also provide useful information about reaction mechanisms, and it allows a test of the principle of microscopic reversibility (see Chapter 18, this volume). For example, in the data shown in Fig. 18, there are components with means of about 4 pA and 5 pA (i.e., conductances of about 40 pS and 50 pS). When amplitudes have been fitted to each opening (Section 5.3.3), it is simple to categorize each transition in the idealized record as $0 \rightarrow 50$ pS, $50 \rightarrow 0$ pS, $0 \rightarrow 40$ pS, $40 \rightarrow 0$ pS, $40 \rightarrow 50$ pS or $50 \rightarrow 40$ pS. This cannot, of course, be done with programs that produce only point-amplitude histograms.

Calculation of a Critical Amplitude

The amplitude components almost always overlap to some extent, so the classification of openings (into 40-pS and 50-pS classes in the above example) will not be entirely unambiguous. A critical amplitude, A_c, that minimizes the total number of amplitudes misclassified was used by Howe *et al.* (1991). This number is proportional to

$$n_{\text{mis}} = a_1 \int_{A_c}^{\infty} f_1(A)dA + a_2 \int_0^{A_c} f_2(A)dA$$

$$= 0.5\{a_1[1 - \text{erf}(u_1/\sqrt{2})] + a_2[1 - \text{erf}(u_2/\sqrt{2})]\} \tag{42}$$

where f_1 and f_2 are the Gaussian densities for the components with smaller and larger means, respectively; a_1 and a_2 are proportional to the areas of these components; erf represents the error function (see Appendix 3.3); and u_1 and u_2 are standard normal deviates; i.e., $u_1 = |(A - \mu_1)/\sigma_1|$, $u_2 = |(A - \mu_2)/\sigma_2|$, where μ_1, σ_1 and μ_2, σ_2 are the means and standard deviations of the components. This is at a minimum when

$$(a_1/\sigma_1)e^{-u_1^2/2} = (a_2/\sigma_2)e^{-u_2^2/2}$$

Thus, A_c may be found by solving the quadratic equation

$$aA_c^2 + bA_c + c = 0 \tag{43}$$

where the coefficients are defined as

$$a = (1/\sigma_2^2) - (1/\sigma_1^2),$$

$$b = 2[(\mu_1/\sigma_1^2) - (\mu_2/\sigma_2^2)],$$

$$c = (\mu_2^2/\sigma_2^2) - (\mu_1^2/\sigma_1^2) - 2\ln[(a_2/\sigma_2)/(a_1/\sigma_1)]. \tag{44}$$

5.4. The Open and Shut Lifetime Distributions

There are only two directly observable types of distribution, the distribution of open times and the distribution of shut times or gaps (i.e., of the durations of the intervals between openings). Although the open times are an obvious focus of attention, the shut times are equally if not more informative (see Chapter 18, this volume). Usually it is sensible to look

characterize quantitatively different
tern *et al.*, 1992). It can also provide
lows a test of the principle of micro-
example, in the data shown in Fig.
d 5 pA (i.e., conductances of about
each opening (Section 5.3.3), it is
rd as 0 → 50 pS, 50 → 0 pS, 0 →
s cannot, of course, be done with

some extent, so the classification
e example) will not be entirely
tal number of amplitudes misclas-
portional to

$\text{erf}(u_2/\sqrt{2})]\}$ (42)

with smaller and larger means,
components; erf represents the
rd normal deviates; i.e., $u_1 =$
are the means and standard

(43)

$\tau_1)]$. (44)

the distribution of open
of the intervals between
ion, the shut times are
ly it is sensible to look

at the shut-time distribution first, because it is this that dictates whether or not it is feasible to divide the openings into bursts.

It is preferable to refer to these distributions as those of *apparent* open times and *apparent* shut times because the effects of undetected shuttings and openings, respectively, mean that the results will rarely be accurate (see Sections 5.2 and 6.11 and Chapter 18, this volume). For example, if some shut times are too short to be resolved, then the measured openings will be too long, because some actually consist of two or more openings separated by unresolved gaps. The shut times may also be too long if they contain brief undetected openings. However the word "apparent" will, for brevity, be dropped when the intention is clear from the context.

Both distributions are usually fitted by mixtures of exponentials, as in equation 30. The number of components in the open-time distribution should be equal to the number of open states, and the number of components in the shut-time distribution should be equal to the number of shut states. It is, of course, always possible that some of the components will be too small or too fast to be detected, so the distributions can provide only a lower bound for the numbers of states. Although these distributions are much more susceptible to errors resulting from missed events than are distributions such as that of the total open time per burst (see below), it is remarkable that such errors should not much affect the *number* of components that are found, even when the time constants of the components are quite wrong (see Section 12 of Chapter 18, this volume, Hawkes *et al.* 1992).

When the patch contains more than one channel, even when no multiple openings are seen, there is no way to be sure whether or not a particular opening originates from the same channel as the preceding opening. This complicates the interpretation of the results (see Chapter 18, this volume). In cases in which the openings are observed to occur in bursts, there is often reason to think that all of the openings in one burst may originate from the same channel, even if the next burst originates from a different channel, so the gaps within bursts may be easier to interpret. It is therefore usually interesting to analyze the characteristics of bursts of openings when it is possible to do so. Distributions that are relevant to this case are considered in Section 5.5.

5.4.1. Multiple Openings

If the experimental record has periods when more than one channel is open, measurement of apparent open lifetimes becomes more difficult. Such records are useful for averaging to simulate a relaxation or for calculation of the noise from the patch recording. They may also be useful for estimating the number of active channels in the patch (see Chapter 18, this volume) and for testing for the mutual independence of channels. In general, however, records with multiple openings are unsuitable for looking at distributions of open times and shut times because, if two channels are open, there is no way of telling, when one of them closes, whether the one that closes is that which opened first or that which opened second.

Although it is possible to recover open- and shut-time distributions from records with multiple openings (Jackson, 1985), it is generally desirable to use records that have only one channel open at a time or only very few multiple openings. In order to use the method of Jackson (1985), the number of active channels must be known, and in most cases this is difficult to estimate accurately (see Section 8 of Chapter 18, this volume), and this method cannot cope with subconductance levels, which almost all channels show to some extent.

When there are only a few multiple openings in a record, one way to deal with them is to omit all the openings in the group where multiple openings occur and to measure the

lifetimes of only the single openings before and after this group. The time between these openings is not a valid shut time and must be marked as "unusable" in the idealized list of shut times so that it can be excluded from the shut-time distribution. This procedure tends to select against long openings, so the open times thus measured will be slightly too short on average. An alternative procedure would be to take the length of the group of multiple openings as a single open time, which would make the open times too long on average. If there are enough multiple openings in the record that the bias could be substantial, then both of these methods could be used; if the two methods give results that disagree by enough to matter, then the number of multiple openings is too large to allow any simple analysis.

5.4.2. Distributions of Open Times Conditional on Amplitude

When there is more than one conductance level, it will usually be interesting to look separately at open times for each level. For example, in the data shown in Fig. 18 there are components with means of about 4 pA and 5 pA (i.e., conductances of about 40 pS and 50 pS). When amplitudes have been fitted to each opening (Section 5.3.3), it is simple to go through each opening and select the openings whose amplitudes lie in a specified range. The histogram is then plotted using the durations of these openings. A method for calculating an optimum critical amplitude that minimizes the number of misclassified amplitudes has been given above, in Section 5.3.5.

It is, of course, necessary to exclude openings that are too short for their amplitudes to be well defined. This is done by excluding from *fitting* (see Section 6.8.1) all values below $t_{min} = 2T_r$ or $2.5T_r$, rather than by imposing a low resolution on the data, as described in Section 5.2.4. Such analyses obviously can not be done with computer programs that do not fit an amplitude to every opening but rely only on all-point amplitude histograms.

5.5. Burst Distributions

5.5.1 Definition of Bursts

In extreme cases, it will be obvious to a casual observer that openings are occurring in groups, separated by long silent periods, rather than at random (exponentially distributed) intervals. For example, Colquhoun and Sakmann (1985) observed groups of channel openings separated by very short shut periods of average duration around 40 μs, even though the agonist concentration was so low that these groups occurred, on average, at intervals of the order of 500 ms (i.e., 10^4 times longer). Empirically speaking, openings will appear to be grouped into bursts whenever the distribution of all shut times requires two (or more) exponentials to fit it. If the time constants for the exponentials are very different, as in the above example, the bursts will be very obvious, and it will usually be quite clear whether any particular shut period should be classified as being within a burst or between bursts. If the time constants differ by less than a factor of 100 or so, the distinction becomes progressively more ambiguous.

Burst characteristics can be rigorously defined in at least two different ways. These two definitions will be, for practical purposes, equivalent in cases (such as the example given above) in which the bursts are very obvious, but in general, they are different. The definitions are as follows.

1. A burst of openings can be defined empirically as any series of openings separated

by gaps that are all less than a specified length (t_{crit}, say). In the example given above, we might take $t_{crit} = 0.4$ ms; the probability that a gap with a mean duration of 50 μs will be longer than 0.4 ms is about 0.3 per 1000, and the probability that a gap with a mean duration of 500 ms is less than 0.4 ms is about 0.8 per 1000. Thus, there is little chance that a gap would be wrongly classified in this case. A suitable value for t_{crit} must be chosen by inspection of the distribution of all shut periods before burst analysis is attempted.

2. A gap within a burst can be defined, for a particular mechanism, as a sojourn in a particular (short-lived) state (or set of states), for example, the blocked state in the case of a simple ion channel-blocker mechanism (see Section 4 of Chapter 18, this volume). Gaps between bursts are then similarly defined as sojourns in a different (long-lived) state or set of states. This definition was adopted by Colquhoun and Hawkes (1982; see Chapters 18 and 20, this volume). Unlike the first definition, it depends on an interpretation of the observations in terms of mechanism. Conversely, though, it allows inferences about mechanism from the observations; it connects the theory with the observations. On the other hand, unlike the first definition, it is not an algorithm that can be automatically and empirically applied to a set of data regardless of subsequent interpretation.

Choice of the Critical Shut Time for Definition of Bursts

There is no unique criterion for the optimum way to divide an experimental record into bursts. At least three methods have been proposed.

Suppose that we wish to find a value of t_{crit} that lies between two components of the shut-time distribution. The slower component has, say, an area a_s and mean t_s, and the faster component is specified by a_f and τ_f (see equation 30).

Jackson *et al.* (1983) proposed that t_{crit} should be defined as the shut-time duration that minimizes the *total number* of misclassified intervals. This criterion involves solving for t_{crit} the equation

$$\frac{a_f}{\tau_f} e^{-t_{crit}/\tau_f} = \frac{a_s}{\tau_s} e^{-t_{crit}/\tau_s} \tag{45}$$

The criterion proposed by Magleby and Pallotta (1983) and by Clapham and Neher (1984) is to choose t_{crit} so that *equal numbers* of short and long intervals are misclassified. This involves solving for t_{crit} the equation

$$a_f e^{-t_{crit}/\tau_f} = a_s(1 - e^{-t_{crit}/\tau_s}). \tag{46}$$

A third approach is to choose t_{crit} so that *equal proportions* of short and long intervals are misclassified (Colquhoun and Sakmann, 1985). In this case, t_{crit} is given by solving

$$e^{-t_{crit}/\tau_f} = 1 - e^{-t_{crit}/\tau_s} \tag{47}$$

None of these three equations can be solved explicitly, but the value of t_{crit} can be found easily by numerical solution by, for example, the bisection method (Press *et al.*, 1992) with τ_f and τ_s as the initial guesses between which t_{crit} must lie.

The three methods defined by equations 45–47 all give different values for t_{crit}, though 46 and 47 will be the same when $a_s = a_f$. When the areas for short and long intervals differ greatly, the first two methods (especially the first) may result in misclassification of a large proportion of the rarer type of interval, and so it may sometimes be felt to be more appropriate

to use the third method, despite the fact that it does not minimize the total number of misclassifications.

When the time constants, τ_f and τ_s are very different, as in the example above, it will make very little difference which of the methods is used. But the difference that is needed is often underestimated. If the record contains N shut times (and N open times) the number of bursts that are found will be N times the probability that a shut time is greater than t_{crit}. The latter probability is, from equation 36,

$$P(\text{shut time} > t_{crit}) = \Sigma a_i e^{-t_{crit}/\tau_i} \tag{48}$$

In the case of the two-component shut-time distribution,

$$P(\text{shut time} > t_{crit}) = a_f e^{-t_{crit}/\tau_f} + a_s e^{-t_{crit}/\tau_s}$$

so, if $t_{crit} >> \tau_f$, the first term will be very small, and if $t_{crit} << \tau_s$, then the second term will be approximately a_s, so the number of bursts located will be Na_s for any value of t_{crit} that satisfies these criteria. Nevertheless, equation 48 shows that the number of bursts found decreases monotonically as t_{crit} is increased. There is no genuine plateau where it becomes independent of t_{crit}.

Consider, for example, the case where τ_s is 100 times longer than τ_f; e.g., $\tau_f = 1$ ms, and $\tau_s = 100$ ms. When $a_f = a_s = 0.5$, the three methods in equations 45–47 give, respectively, $t_{crit} = 4.65$ ms, 3.40 ms, and 3.40 ms. The total number misclassified per 100 openings is, respectively, 2.75, 3.34, and 3.34, but the first method misclassifies 4.5% of long openings and 0.95% of short openings, whereas the last misclassifies 3.34% of both. When there are more short openings than long (i.e., many openings per burst), say $a_f = 0.9$, $a_s = 0.1$, the results are the same for the last method, but 45 and 46 give $t_{crit} = 6.87$ ms and 5.18 ms respectively, and equation 45 gives the total number misclassified per 100 as only 0.757, though 6.6% of long openings and 0.10% of short are misclassified.

If, however, τ_s is only 10 times longer than τ_f, e.g., $\tau_f = 1$ ms and $\tau_s = 10$ ms, then, when $a_f = a_s = 0.5$, the three methods give $t_{crit} = 1.80$ ms, 1.80 ms, and 2.56 ms, respectively, but even equation 45 misclassifies 15.2 shut times per 100, with 22.6% of long shut times being misclassified. Clearly, a factor of 10 is not big enough. This is apparent immediately from the fact that 16.5% of intervals with a mean length of 1 ms are greater than $t_{crit} = 1.80$ ms, and 16.5% of intervals with a mean length of 10 ms are shorter than 1.80 ms.

Once bursts have been defined, many sorts of distribution can be constructed from the idealized record, some of which are now listed.

5.5.2. The Distribution of the Number of Openings per Burst

We simply count the number of apparent openings (r, say) in each burst. Unlike the other distributions to be considered, this is a discontinuous variable; it can take only the integer values 1,2, . . . , ∞. This number will, of course, be underestimated if some gaps are too short to be resolved (see Sections 5.2 and 6.11 and Chapter 18, this volume). If there is only one sort of open state, the number of openings per burst is expected to follow a geometric distribution, i.e.,

$$P(r) = (1 - \rho)\rho^{r-1} \tag{49}$$

which decreases with r (because $\rho < 1$). The mean number of openings per burst is

$$\mu = 1/(1 - \rho) \tag{50}$$

Further details are given in Sections 6.1 and 6.8. Notice that $P(r)$ decreases by a constant factor (ρ) each time r is incremented by 1. This property is characteristic of exponential curves, and the geometric distribution is in fact the discrete equivalent of the exponential distribution encountered elsewhere. When the mean becomes large, the distribution approximates the exponential distribution with mean μ, namely, $\mu^{-1}e^{-r/\mu}$.

In general, the distribution will be a mixture of several such geometric terms; the number of terms will often be equal to the number of open states but may be fewer in principle (apart from the problem that not all components may be detectable). The question of the expected number of components is quite complex and is discussed in Section 13.4 of Chapter 18 (this volume).

5.5.3. The Distribution of Burst Length

This is the length of time from the beginning of the first opening of a burst to the end of the last opening. Clearly, it will be relatively unaffected by the presence of short unresolved gaps, compared with the distributions of open times and of number of openings per burst. The distribution should be described by a mixture of exponentials, as in equation 30, under the usual assumptions. The number of exponential components is, in principle, quite large, being equal to the number of open states plus the number of short-lived shut states (see Chapter 18, this volume; Colquhoun and Hawkes, 1982). In practice it is unlikely that all components will be resolved, and under some circumstances the burst length distribution may be well-approximated by a single exponential, as described in Section 5.3 of Chapter 18 (this volume).

5.5.4. The Distribution of the Total Open Time per Burst

This is the total length of all the openings in each burst. It is also relatively insensitive to undetected brief openings or shuttings (shuttings that are brief enough to be missed will cause only a small error in measuring the total open time). This distribution should also be described by a mixture of exponentials, as in equation 30. It is, in principle, simpler than the distribution of burst length, because the number of components is expected to be equal to the number of open states (Chapter 18, this volume; Colquhoun and Hawkes, 1982). This, together with the fact that it is less sensitive to missed events than the distribution of apparent open times, makes it the best distribution to look at in order to make inferences about the (minimum) number of open states. The distribution of the total open time per burst is also of interest because it is predicted, surprisingly, that it will not be affected by a simple channel blocker (see Chapter 18, this volume). This prediction provides a useful way of investigating blocking mechanisms (Neher and Steinbach, 1978; Neher, 1983; Colquhoun and Ogden, 1985; Johnson and Ascher, 1990).

The distribution of the total shut time per burst may also be of interest for some sorts of interpretation (Colquhoun and Hawkes, 1982).

5.6. Cluster Distributions

Sakmann *et al.* (1980) observed that bursts of openings could themselves be grouped together into clusters of bursts with long gaps between clusters. They were looking at nicotinic

channels with high agonist concentrations, and the long silent periods between clusters occurred when all the ion channels in the patch were in long-lived desensitized states. In records of this sort it is often possible to say, with a high degree of certainty, that all of the openings in one cluster originate from the same individual ion channel. All of the shut times within a cluster can therefore be interpreted in terms of mechanism, even when the number of channels in the patch is not known (see Section 8 of Chapter 18, this volume). Such clusters are also useful for measurement of the probability that a channel is open (P_{open}), as described in Section 5.1.7.

Another case in which clusters of bursts (and superclusters of clusters) have been observed is the NMDA-type glutamate receptor (Gibb and Colquhoun, 1991, 1992). Measurements at very low agonist concentrations allow resolution of this unusually complex structure if the individual channel activations and subdivision of the record into bursts of openings and into clusters of bursts should aid in the interpretation of such records. The relevant theory has been given by Colquhoun and Hawkes (1982). This can, of course, be done only when the time constants of the shut-time distribution are sufficiently well separated (see Section 5.1). The mean gap between clusters should preferably be at least 100 times greater than the mean gap between bursts (within a cluster); and the latter should preferably be 100 times greater than the shut times within a burst.

Of course, we are quite free to treat the whole cluster as a long burst by an appropriate choice of t_{crit} (see Section 5.5.1); these bursts can then be analyzed like any other (they will have a rather complex distribution of gaps within bursts). Equally, we can ignore the clustering and analyze the individual bursts as above (the distribution of gaps between bursts would then be rather complex).

When the record is divided into clusters of bursts, a large number of different sorts of distributions can then be constructed, for example, the length of the kth burst in a cluster and the distribution of gaps between bursts within clusters; further details are given by Colquhoun and Hawkes (1982).

5.7. Measurement and Display of Correlations

Certain types of mechanism can give rise to correlations between the length of one opening and the next or between the length of an opening and that of the following shut time. When this happens, the correlation will gradually die out over successive openings: there will be a smaller correlation between the length of an opening and the length of the next but one opening (described as a correlation with lag = 2), and so on for increasing lags. Such correlations have been observed for both nicotinic and NMDA receptors. Measurements of correlation can give information about mechanisms, in particular information about how states are connected, that cannot be found in any other way. The origin and interpretation of correlations are discussed in Section 10 of Chapter 18 (this volume), where appropriate references will be found. We shall discuss here the ways in which correlations may be measured and displayed.

5.7.1. Correlation Coefficients and Runs Test

Perhaps the simplest way to test for correlations is to use a *runs test,* as employed by Colquhoun and Sakmann (1985). To do this, open times (or shut times or burst lengths, etc.) that are shorter than some specified length (e.g., 1 ms) are represented as 0, and values

longer than this length are represented as 1. We then ask whether runs of consecutive 0 values (or of consecutive 1 values) occur with the frequency expected for independent events. If, for example, long openings tend to occur together, this will produce long runs of 1 values. Say there are n_0 zero values, n_1 unity values, and $n = n_0 + n_1$ values altogether. The number of runs, N_r say, in the data is then counted, a run being defined as a contiguous section of the series that consists entirely of (one or more) 0 values or entirely of 1 values (thus 110001 has three runs). If the series is random, then the mean and variance of N_r will be

$$E(N_r) = \frac{2n_0n_1}{n} + 1 \qquad \text{var}(N_r) = \frac{2n_0n_1(2n_0n_1 - n)}{n^2(n-1)} \tag{51}$$

The test statistic

$$z = \frac{N_r - E(N_r)}{[\text{var}(N_r)]^{1/2}} \tag{52}$$

will have an approximately Gaussian distribution with zero mean and unit standard deviation, so a value of $|z|$ larger than about 2 is unlikely to occur by chance.

The extent of correlation for any specified lag m can be calculated as a correlation coefficient, r_m. If the observations (e.g., open times, shut times, burst lengths, etc.) are denoted $t_1, t_2, \ldots, t_n$, with mean $\bar{t}$, then the correlation coefficient is calculated as

$$r_m = \frac{\sum_{i=1}^{i=n-m} (t_i - \bar{t})(t_{i+m} - \bar{t})}{\sum_{i=1}^{i=n} (t_i - \bar{t})^2} \tag{53}$$

5.7.2. Distributions Conditional on Length of Adjacent Event

The calculations in the last section give no visual impression of the strength of correlations, but various graphical displays that do so can be made. For example, the distribution of the length of openings conditional on the length of adjacent shut time can be constructed. Examples of such conditional distributions are shown in Chapter 18, this volume, (Section 10, Fig. 13). If, as in these examples, short openings tend to occur next to long shuttings, then the distribution of open times, conditional on the open time being next to a long shutting, will show an excess of short openings (relative to the overall open-time distribution). In order to construct such a conditional distribution from experimental data, it is necessary to specify a range of shut times rather than a single value. For example, to construct a distribution of open times conditional on the adjacent shut time being between 0.05 and 0.3 ms, simply locate all the open times that are adjacent to shut times that fall in this range and plot the histogram of these openings.

A more synoptic view can be obtained by restricting attention to the mean open times rather than looking at their distribution. Define several shut-time ranges and then plot the mean open time (for openings that are adjacent to shut times in each range) against the midpoint of the range. It will generally be best to center these shut-time ranges around the time constants of the shut-time distribution. The mean open time may also be plotted against

the *mean* of the shut times in the range rather than against the midpoint of the range. An example is shown in Chapter 18 (this volume, Fig. 12). The mean open time decreases as the adjacent shut time increases.

A third way to display correlation information is to construct a two-dimensional dependency plot (Magleby and Song, 1992). This plot is explained and illustrated in Chapter 18 (this volume, Section 10).

5.7.3. Distribution of Open Time Conditional on Position within the Burst

The distributions of quantities such as (1) the length of the kth opening in a burst or (2) the length of the kth opening in a burst for bursts that have exactly r openings are potentially informative when there are correlations in the data. If these distributions differ for different values of k (or of r), these variations can be tested against the predictions of specific mechanisms, which can be calculated as described by Colquhoun and Hawkes, 1982; Chapter 20, this volume). Such distributions are, however, likely to be rather sensitive to undetected brief events (see Section 6.11 below; Section 12 of Chapter 18, this volume). Their potential has yet to be exploited.

5.8. Distributions following a Jump: Open Times, Shut Times, and Bursts

It is often of interest to measure single-channel currents following a rapid (step-like) change of membrane potential or of ligand concentration (a *voltage jump* or *concentration jump*). The principles underlying such measurements are discussed and exemplified in Section 11 of Chapter 18 (this volume).

Notice that application of a rectangular pulse (of membrane potential or of ligand concentration) is actually *two* concentration jumps. In terms of macroscopic current, the first step is sometimes referred to as the "on-relaxation," and the second, when the stimulus is returned (usually) to the prejump condition, is referred to as the "off-relaxation." In the context of voltage-activated channels (but, for no particular reason, not for agonist-activated channels), the off-relaxation is often referred to as a "tail current"; it is probably rather unhelpful, though harmless, to use a separate term for an off-jump, since it does not differ in principle from an on-jump. Sometimes attention is focused mainly on the on-relaxation (e.g., when a step depolarization opens a voltage-activated channel); sometimes the main focus is more on the off-relaxation (e.g., the events following a brief pulse of agonist applied to an agonist-activated channel).

The distribution of the latency until the first opening occurs is of crucial importance for understanding topics such as the shape of synaptic currents or the mechanism of inactivation of sodium channels (see Chapter 18, this volume). In principle it is easy to measure it from experimental records. The main problem in practice is that it cannot be interpreted unless there is only one channel in the patch (or at least a known number of channels). This is often hard to achieve.

Even when the channel shows no correlations, the distribution of first latencies is expected to differ from that of other shut times (see Chapter 18, this volume), though in this case the distributions of all subsequent shut and open times should be the same as those at equilibrium. When the channel shows correlations, the distributions (and hence means) of the first, second, . . . open time, and shut time, after the jump may differ from their equilibrium

values. If the channel also shows correlations between burst lengths, then the distributions of the first, second, etc. burst length following the jump will also differ. After a sufficient number of openings has occurred, the equilibrium distributions will eventually be attained. Further details and examples can be found in Chapter 18 (this volume, Sections 10 and 11).

5.8.1 Delays in the Recording System

When first latencies are being measured, it is obviously very important that we know precisely when the step was applied (i.e., where $t = 0$ lies on the experimental record).

Voltage Jumps

In the case of voltage jumps, this problem has been discussed in detail by Sigworth and Zhou (1992). It is important to compensate properly for the large capacitative current artifact that accompanies a voltage jump applied with the patch clamp. Methods for doing this are discussed in Chapter 7 (this volume) and by Sigworth and Zhou (1992). The voltage jump may not be applied to the patch at the precise moment that the command pulse is applied. This can happen because vagaries of the relative timing of DAC outputs and ADC inputs: these depend on the characteristics of the computer's real-time interface and on precisely how it is programmed. Delays may also occur when the command pulse is filtered (to reduce its maximum rate of rise). The true $t = 0$ point on the record can be estimated by measuring the time from when the command pulse starts to the midpoint of the instantaneous current (the current that flows "instantly" through channels that are already open when the potential changes). Alternatively, the capacity compensation can be slightly misadjusted, and then one can measure the time to the peak of the resulting capacitative current. These procedures are illustrated by Sigworth and Zhou (1992).

Concentration Jumps

In the case of concentration jumps, delays may be much greater than for voltage jumps. Typically, a jump is applied to an outside-out patch by moving (by means of a piezoelectric device) a theta glass pipette from which two solutions flow, so the interface between the solutions moves across the patch. Delays arise primarily because of the time taken for the command pulse to be translated into movement of the piezo and the time taken for the solution leaving the theta glass to reach the patch. The delay can be measured as follows. Break the patch at the end of the experiment and flow a hypotonic solution through one side of the application pipette; then measure the time from application of a command pulse to the piezo to the appearance of a junction response. It is obviously important that the relative position of patch and application pipette remains the same throughout. It is still better if the measurement of delay can be made with the patch intact, as it is during the experiment proper. This may be possible, for example, by applying a step change in potassium concentration while a potassium-permeable channel is open (the channel opening itself can be used to trigger the command pulse to the piezo). This method was used by Colquhoun *et al.* (1992) to estimate the rate at which concentration changes at the patch surface; it is also an ideal method to measure delay (as long as an appropriate channel can be found).

There will also be a delay in the current-measurement pathway, essentially all of which is caused by filtering. An eight-pole Bessel filter introduces a delay (in seconds) of $0.51/f_c$,

where f_c is the -3 dB frequency in Hertz. For example, a 1-kHz filter introduces a delay of 510 μs (Sigworth and Zhou, 1992).

The fitting of the results of jump experiments is considered later, in Section 6.13.

5.9. Tests for Heterogeneity

It is, unfortunately, quite common for more than one sort of channel to be in the patch of membrane from which a recording is made. This may be the case not only with native receptors but also with recombinant channels expressed in oocytes (e.g., Gibb *et al.*, 1990); injection of a defined set of subunit RNAs does not necessarily guarantee that a single well-defined sort of channel will be produced (see also Edmonds *et al.*, 1995a,b). This sort of heterogeneity will make distributions confusing and serious kinetic analysis impossible. It is, therefore, important to know when it is present.

One criterion that has been used for agonist-activated channels is based on P_{open} measurements (see Section 5.1.7). At high agonist concentrations, when the probability of the channel being open is high, openings appear in long clusters separated by even longer desensitized periods (Sakmann *et al.*, 1980; see also Section 5.6). Because all of the openings in one cluster are likely to arise from the same individual channel, a value of P_{open} can be measured from each cluster (by integration or by measuring individual open and shut times; see Colquhoun and Ogden, 1988, for example). The next cluster may arise from a different channel, but it should give, within sampling error, the same value for P_{open} if all the channels in the patch are identical.

An excellent method for assessing whether the P_{open} values (or open times or shut times, etc.) vary to a greater extent than expected from sampling error was proposed by Patlak *et al.* (1986). They used a randomization test (an elementary account of the principles of randomization tests is given by Colquhoun, 1971). This method has been used, for example, by Mathie *et al.* (1991) and by Newland *et al.* (1991). Suppose that measurements are made on N clusters of openings, and n_i is the number of openings in the ith cluster. The observed scatter of the measurements, S_{obs}, can be measured as

$$S_{obs} = \sum_{i=1}^{N} n_i(y_i - \bar{y})^2 \tag{54}$$

where y_i represents the measurement of interest (e.g., P_{open} or mean open time or mean shut time) for the ith cluster. The probability of observing a value of S_{obs} (or larger) on the null hypothesis that the clusters are homogeneous, can then be found as follows. Take all the measured open and shut times from all the clusters as a single group and select at random from them N groups of n_i values. Then calculate the scatter from these artificially generated clusters, using equation 54 in exactly the same way as was done for the real measurements; this will produce a value that may be denoted S_{ran}. This randomization procedure is then repeated many times (e.g., 1000 or more). A histogram can be constructed from the values of S_{ran} so generated. The fraction of cases in which S_{ran} exceeds the observed value, S_{obs}, is the required probability. If it is very small, then it is unlikely that the null hypothesis was correct, and it must be supposed that the channels are heterogeneous.

6. The Fitting of Distributions

6.1. The Nature of the Problem

The term *fitting* means the process of finding the values of the constants in some specified equation that produce the best fit of that equation to the experimental data. This definition implies that one must (1) decide on an appropriate equation to fit to the data, (2) decide what the term "best" means, and (3) find an algorithm that can then find the best fit. It is perhaps worth noting that the process of fitting involves thinking in a somewhat inverted way. Normally, one thinks of the data as being variable and the parameters in an equation (e.g., the slope and intercept of a straight line) as being constants. During the process of fitting, though, the data are constant (whatever we happened to observe), but the parameters ("constants") are varied to make the equation fit the observations.

There are two quite different approaches to fitting, which may be called (1) empirical fitting and (2) fitting a mechanism directly. In the former case, exponentials (or geometrics) are fitted without necessarily specifying any particular mechanism; the parameters are the time constants and areas of the exponential components. In the latter approach, the parameters to be fitted are not the time constants of the exponentials but the underlying rate constants in a specified mechanism. The former approach is by far the most common, and it will be discussed next. The direct fitting of mechanisms requires consideration of missed events and will be discussed later, in Section 6.12.

6.1.1. Empirical Fitting of Exponentials

In practice, this usually means fitting a mixture of exponential distributions to data that consist of a list of time intervals (e.g., a list of apparent open times, shut times, or burst lengths, found as described earlier). Similarly, a mixture of geometric distributions may be used to fit the number of openings per burst, etc. This process is not entirely empirical, however, because there is good reason to expect that these may be appropriate equations. Any "memoryless" reaction mechanism is expected to result in observations that can be described by exponentials (or geometrics), as described in Chapter 18 (this volume), so they are obviously sensible things to fit. There is, of course, no guarantee that they will fit adequately. For example, (1) the effect of limited time resolution will, in principle, result in nonexponential distributions (e.g., Section 6.11, Chapter 18, this volume; Hawkes *et al.*, 1990, 1992), or (2) the transition rates may not be constant, e.g., because membrane potential or ligand concentration are not constant, or (3) the mechanism may be genuinely non-Markovian. These topics are discussed at greater length in Chapter 18 (this volume).

The general form for a mixture of exponential densities has already been given in equation 30. If a_i represents the area of the ith component, and τ_i is its mean or time constant, then, when there are k components,

$$f(t) = a_1\tau_1^{-1}e^{-t/\tau} + a_2\tau_2^{-1}e^{-t/\tau_2} + \cdots$$

$$= \sum_{i=1}^{k} a_i\tau_i^{-1}e^{-t/\tau_i} \tag{55}$$

The areas add up to unity, i.e., $\Sigma a_i = 1$, and the overall mean duration is $\Sigma a_i\tau_i$. Although it was stated above that the areas are proportional, roughly speaking, to the number of events

in each component, it must be emphasized that, in general, the areas and time constants (means) of the components have no separate physical significance. An approximate physical interpretation of the components may be possible in particular cases (some examples are given in Chapter 18, this volume), but they must be demonstrated separately in each case. The density is sometimes written in the alternative form

$$f(t) = \sum_{i=1}^{k} w_i e^{-t/\tau_i} \tag{56}$$

where the coefficients w_i are the amplitudes (dimensions s^{-1}) of the components at $t = 0$. Clearly, they are related to the areas thus:

$$w_i = a_i/\tau_i. \tag{57}$$

The cumulative exponential distributions have already been given in equations 35 and 36.

6.1.2. Empirical Fitting of Geometrics

The general form for a mixture of geometric distributions (see Section 5.5.2) with k components, is

$$P(r) = \sum_{i=1}^{k} a_i(1 - \rho_i)\rho_i^{r-1}, \qquad r = 1, 2, \ldots, \infty \tag{58}$$

where a_i is the area of the ith component, and ρ_i is a dimensionless parameter ($\rho_i < 1$) (see Chapters 18 and 20, this volume). Alternatively, this can be written as

$$P(r) = \sum_{i=1}^{k} w_i \rho_i^{r-1} \tag{59}$$

where the coefficients w_i are the relative amplitudes, at $r = 1$, of the components. The area and amplitudes are related by

$$a_i = w_i/(1 - \rho_i) \tag{60}$$

The "means", μ_i, of the individual components (which, as for exponentials, will not generally have any separate physical significance) are

$$\mu_i = 1/(1 - \rho_i) \tag{61}$$

and the overall mean is

$$\mu = \sum_{i=1}^{k} a_i \mu_i \tag{62}$$

Thus, yet another general form for a mixture of geometric distributions is

$$P(r) = \sum_{i=1}^{k} a_i \mu_i^{-1}(1 - \mu_i^{-1})^{r-1} \tag{63}$$

Under certain circumstances (see Section 13.4 of Chapter 18, this volume), it is predicted that there will be a component with $\rho = 0$, i.e., from equation 62, a component with a mean of exactly one opening per burst. Such a component will contribute to $P(1)$ only.

The cumulative form of the geometric distribution, i.e., the probability that we observe n or more (e.g., the probability of observing n or more openings per burst) is

$$P(r \geq n) = \sum_{i=1}^{k} a_i \rho_i^{n-1} \tag{64}$$

6.1.3. The Number of Parameters to Be Estimated

In the cases of both exponential and geometric distributions there are $2k - 1$ parameters, the values of which must be estimated from the data by the fitting process. For exponentials there are k different time constants, τ_i (or rate constants, $\lambda_i = 1/\tau_i$) and $k - 1$ values for the areas, a_i (the areas must sum to 1, so estimation of $k - 1$ values defines the kth value). For geometrics, the parameters could be k values of μ_i (or of ρ_i), plus $k - 1$ values for the areas, a_i. Sometimes it may be desirable not to estimate all of these parameters from the data but to fix the values of one or more of them (they might, for example, be fixed at values that have already been determined from earlier experiments). This should improve the precision of the remaining parameters that are estimated from the data.

It will always be sensible to constrain the values of the time constants, τ_i, to be positive when fitting exponentials. Negative values are obviously impossible, so the fitting routine should be prevented from trying negative values. Similarly, in fitting geometrics, the values of μ_i should be constrained to be not less than 1 (or, if fitting ρ_i, the ρ_i values should be constrained to lie between 0 and 1). When fitting steady-state distributions, the areas, a_i, of the components are expected to be nonnegative too, so it may help the fitting process if they too are constrained. However, some sorts of distribution (for example that of the shut time preceding the first opening after a jump) may well have one or more negative areas; in such cases it is important that the program *not* constrain areas to be positive (see Sections 7 and 11 of Chapter 18, this volume).

6.2. Criteria for the Best Fit

The usual approach is to define a measure of the goodness of fit (or of the badness of fit) of the fitted distribution to the experimental observations. The parameter values are then chosen to maximize the goodness of fit (or to minimize the badness of fit). Different measures of goodness of fit will give different estimates of the parameters from the same data.

For conventional curve fitting (e.g., to ordinary graphs or to macroscopic currents), the weighted least-squares criterion is usually supposed (and in some cases has been shown) to be the best method. In such cases the distribution of the observations is almost always

unknown. In the single-channel context, though, the problem is rather different. The distribution of the observations *is* known—it is what is being fitted. It is, therefore, possible to do better by using the maximum-likelihood approach.

The likelihood function provides a measure of goodness of fit and is discussed in Sections 6.5–6.9. Other, less good, methods appear in the literature, e.g., the χ^2 statistic (which provides one measure of badness of fit and is discussed in Section 6.4). Still worse, one can find some wholly inappropriate use of least-squares criteria, or even "curve stripping" on semilogarithmic plots, but they are not worth discussion here.

All fitting methods will give much the same results if the amount of data is very large and the fit very close, but this is rarely the case in the real world. The maximum-likelihood method is undoubtedly preferable to any other for the purposes of fitting distributions, and the speed of computers is now such that there is no reason to use any other method.

6.2.1. How Many Components Should Be Fitted?

If a specified mechanism is being fitted (see Section 6.12), the mechanism dictates the number of exponential components, so there is no problem. But when exponentials are being fitted without reference to a mechanism, it is often difficult to decide how many components should be fitted to the observations. For example, in Fig. 15 the shut-time data are shown fitted with both a two-exponential fit and a three-exponential fit. The fastest and slowest components are obvious, but the intermediate component (with $\tau = 1.31$ ms) has only 3.7% of the area and could easily be missed, especially if the log display were not used and the histogram that reveals this component most clearly (Fig. 15*B*) were not inspected. It could also be missed easily if the amount of data were smaller.

There are three ways in which to judge the number of components that are needed: (1) visual inspection of the histograms—e.g., in Fig. 15 the need for the third component is pretty convincing when the appropriate display is inspected. (2) By checking the reproducibility of the time constants and areas from one experiment to another (if they are not reasonably reproducible you are probably trying to fit too many components). And (3) statistical tests - the question can be asked 'is there a statistically-significant improvement in the fit when an extra component is added?' The second of these methods is by far the most reliable.

The Statistical Approach

The statistical approach is easy to apply when the fitting is done by the method of maximum likelihood (see below). This and related questions are discussed by Horn (1987). Denote as L the maximum value for the log(likelihood), i.e., the value evaluated with the best-fit parameters, $L(\hat{\theta})$ (see Sections 6.5–6.8). Suppose that the same data are fitted twice. First we fit k_1 components (and hence $n_1 = 2k_1 - 1$ parameters), yielding a maximum value for the log-likelihood of L_1. Next we fit the same data again, but with more components (say k_2 components and hence $n_2 = 2k_2 - 1$ parameters); this time the maximum value for the log-likelihood is L_2. With more free parameters to adjust, the second fit is bound to be better (so $L_2 > L_1$), but is it *significantly* better or not? The extent of the improvement in fit can be measured by

$$R = L_2 - L_2$$

where R is the *log likelihood ratio,* i.e., the logarithm of the ratio of the two likelihoods, defined such that the ratio is greater than 1 ($R > 0$). It can be shown (e.g., Rao, 1973) that,

if the correct number of components were k_1, then $2R$ would have (for large samples) a χ^2 distribution with $n = n_2 - n_1$ degrees of freedom. Thus, by obtaining the probability corresponding to $2R$ (by computation or from a χ^2 table), it is possible to judge whether the second fit is significantly better. The P value so found is the (approximate) probability that fitting with k_2 components would produce, by chance, an improvement in fit equal to (or greater than) that observed, if in fact the fit with k_1 components were correct. If P is sufficiently small, it would be concluded that chance alone is unlikely to account for the observed improvement, so the larger number of components is justified.

The Criterion of Reproducibility

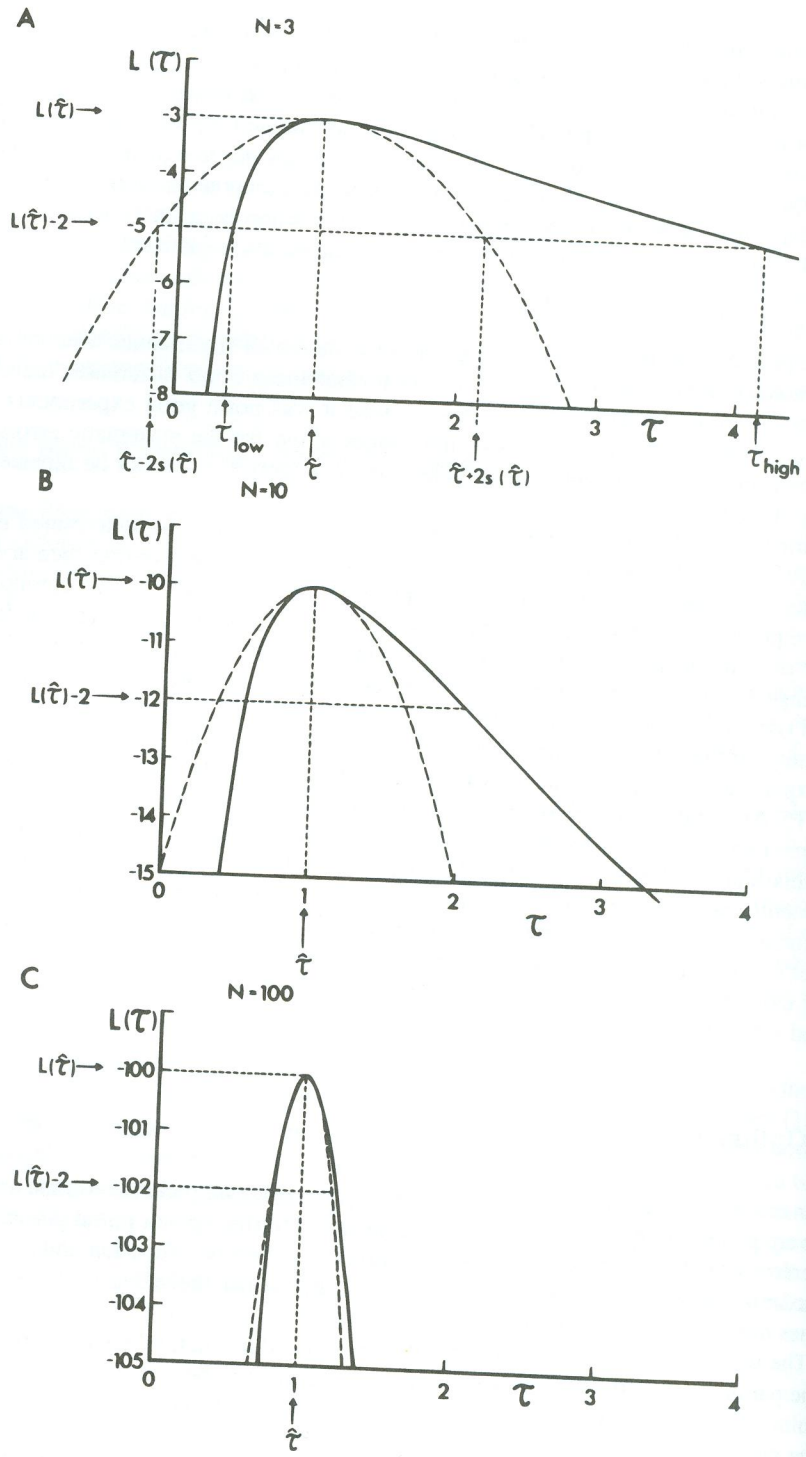
The problems with the statistical approach are, as for all significance tests, of two sorts. First, a nonsignificant difference does not mean that there is no difference, merely that a difference could not be detected (possibly because it was not a good experiment). Second, the test copes only with random errors and cannot allow for the systematic errors that are so common in real experiments. Nevertheless, if the experiment cannot be repeated, this is probably the best approach.

Normally, though, a distribution (such as that in Fig. 15) is not determined only once but many (or at least several) times in separate experiments. The question then arises about what should be done if some experiments appear to be fitted well by two components but others require three. This question shows the inadequacy of the statistical approach. The number of components that are required is dictated by the mechanism involved and does not change from one experiment to another (as long as they are all done with the same channel type and are not invalidated by heterogeneity of the channels). However, the amount, and quality, of the data, and hence one's ability to *detect* components, may vary considerably from experiment to experiment. This is illustrated nicely by the history of the data shown in Fig. 15. At first distributions of this sort were usually fitted with two components. However, it became apparent that quite often the data needed three components, as in the case shown. Once this had become quite convincing, the earlier data sets were all refitted with three components, *whether or not this produced a significant improvement* in any individual experiment. The results showed that, within reasonable limits, the time constant and area of all three components were reproducible from one experiment to another. This is the strongest sort of evidence for the need for three components, and it is the procedure that should be adopted whenever possible.

6.3. Optimizing Methods

In order to begin, one should obtain a good optimizing computer subroutine or procedure. These are general-purpose programs that are designed to find (given initial guesses for them) the parameter values that minimize (or maximize) any specified function and so can be used to maximize the likelihood (or, equivalently, to minimize the negative likelihood—most routines are designed to mimimize).

The user has to supply only a subroutine or procedure that, when supplied with values for the parameters, will calculate a value for the quantity to be minimized (e.g., a value for the minus log-likelihood; see below). The minimization subroutine then adjusts the values for the parameters, and for each set of parameters it calls the user's routine to see how well the parameters fit the data. It is not expected that the user's routine will itself change the parameter values (though it can be useful to have it do so, as described below).


A
N=3
L(τ)
L(τ̂)
L(τ̂)-2
-3
-4
-5
-6
-7
-8
0
1
2
3
4
τ
τ low
τ high
τ̂-2s(τ̂)
τ̂
τ̂+2s(τ̂)
B
N=10
L(τ)
L(τ̂)
L(τ̂)-2
-10
-11
-12
-13
-14
-15
0
1
2
3
4
τ
τ̂
C
N=100
L(τ)
L(τ̂)
L(τ̂)-2
-100
-101
-102
-103
-104
-105
0
1
2
3
4
τ
τ̂

The fitting process is shown graphically in Fig. 19 in the case where there is only one parameter to be estimated. We simply find the value of the parameter that corresponds to the maximum log-likelihood. In the case where two parameters are to be fitted, we would need a three-dimensional version of this graph, with the possible values of the two parameters on the x and y axes and the value of the likelihood that corresponds to each pair of parameter values sticking out of the paper on the z axis. This sort of graph is often shown as a contour map in two dimensions, with parameter values on x and y axes and the corresponding likelihood values marked on contours. The contour map portrays, in geographic terms, a hill, and the problem is now simply to find the top of the hill; this is the maximum likelihood, and the pair of parameters that are the coordinates of the maximum are the *maximum-likelihood estimates* of the parameters. Since optimizing subroutines usually minimize functions, the more common geographic analogy is that we are searching for the bottommost point in a valley. These graphical analogies are usually an excellent way to picture what is happening (though in ill-behaved cases it is possible for contour lines to cross each other, which is not allowed in ordinary maps; e.g., see Colquhoun, 1971, Fig. 12.8.2).

There are very many programs available. They may be found in many standard libraries (such as NAG or IMSL) or by inquiry from your local computer center. The main choice lies between simple search methods and more complicated gradient methods. A much faster, noniterative method for fitting macroscopic exponential curves by use of Chebyshev polynomials is available, but it is inappropriate for fitting distributions. Even for macroscopic exponentials, it cannot be recommended until such time as the properties of the estimates it provides (in comparison with least-squares estimates) have been fully explored.

6.3.1. Simple Search Methods

The simple search methods look for the bottom of a valley by trying various sets of parameter values and simply noting whether one set of values is better than the previous set. "Better" means "further down the valley"; i.e., the user-defined subroutine produces a smaller value of the quantity to be minimized.

An advantage of search methods is that they usually converge (approach the bottom of the valley) reliably, even with poor initial guesses or when the function is ill behaved. These properties can be quite important. It is a considerable advantage in practice to be able to give rather rough initial guesses (it takes time to find good guesses). Even more important is the ability of these methods to cope with any sort of constraint on the values being fitted. For example, in fitting the time course of single-channel openings, as described in Section

---

Figure 19. The log-likelihood, $L(\tau)$, of a particular value of the time constant, τ, plotted against τ (continuous line) for the case of a simple exponential distribution (from equation 71). Graphs are given for samples of size $n = 3$ (A), $n = 10$ (B), and $n = 100$ (C). The dashed line shows the quadratic curve that has the same curvature at the maximum as $L(\tau)$, namely $Q(\tau) = Q(\hat{\tau}) - (\tau - \hat{\tau})^2/2s^2$ where $s = \hat{\tau}/\sqrt{n}$ (see equation 80). The curves have been drawn for the case $\hat{\tau} = 1$, and the abscissa can be interpreted as $\tau/\hat{\tau}$. In graph A, the definition of standard deviations and likelihood intervals is illustrated for the case of $m = 2$ unit likelihood intervals and the corresponding ± 2-standard-deviation intervals (see Table I and Section 6.7.2). A horizontal dashed line is drawn two units below the maximum, i.e., at $L(\hat{\tau}) - 2$. The points at which this intersects the continuous line give the lower and upper limits (τ_{low} and τ_{high}) for $\hat{\tau}$. The points of intersection with the dashed line give the 2-standard-deviation limits, $\hat{\tau} \pm 2s(\hat{\tau})$. For large samples, the dashed and continuous lines become similar, so the two approaches to error specification give similar results (see also Table I).

4.2, it is desirable to constrain the amplitude of a short opening to be the same as that of the nearest opening that is long enough to have a well-defined amplitude. But as the parameter estimates are adjusted, what is considered "short" and "long" may change, so the function being fitted changes as the fitting progresses, and this function may itself change the parameter values. A similar sort of thing occurs in fitting distributions. If the $k - 1$ areas being fitted are adjusted by the minimization routine so that they add up to 1 or more, and it is desired to prevent the kth area being negative (this is not *always* desirable—see Section 6.1.3), then the function that is being minimized can scale all the area values down so they add up to, say, 0.99, and return the altered values to the minimization routine. Such tricks are very useful, but gradient methods tend to take grave exception to them, whereas search methods, which care only about whether the function is reduced or not, carry on quite happily. Search methods also take little computer memory (though this is rarely critical with modern computers). On the other hand, search methods are usually rather slow, especially in the later stages of convergence when high precision is demanded.

Simple search methods include *patternsearch* (see Colquhoun, 1971), and the *simplex* method (Nelder and Mead, 1965; O'Neill, 1971; Hill, 1978; Press *et al.,* 1992). Care is needed because there are many versions of *simplex* in circulation, some of which are not very good. The version given by Press *et al.* (1992), called AMOEBA, appears to be quite satisfactory; the version they give is somewhat inconvenient to use as it stands, so an example is given, in Appendix 3 (Section A3.4), of a small subroutine that may conveniently be used to call AMOEBA. The program as it stands is rather minimal; it can be improved, for example, by adding code (1) to print out the progress of the iterations, (2) to abort the program from the keyboard if it appears to be stuck, (3) to test the convergence by the parameter step size rather than by the reduction in the function, (4) to keep track of the absolute minimum encountered (which may sometimes be better than the final result), and (5) to restart the minimization if a local search after convergence suggests that further improvement is possible. A particularly valuable addition is code to allow the values of specified parameters to be fixed (e.g., at values determined from other experiments) rather than estimated. This can be achieved by defining the parameter array (*theta,* in Section A3.4) to contain all of the parameters (so it can be used for calls to the function or for printing the current parameter values), but defining a second array from which the fixed values are omitted for use by *simplex* when it is adjusting the parameter values.

6.3.2. Gradient Methods

There are many types of gradient methods (see Press *et al.,* 1992, for a brief survey). They have in common the characteristic that, given a set of parameter values that define a point on the surface of the value, they calculate the slope of the surface at that point and use this value to work out the next set of parameter values to try. For example, they may work out the direction of steepest descent and follow this path in the hope that it is the fastest way to the bottom of the valley.

Gradient methods fall into two main categories, as far as the user is concerned. One category requires only that the user supply a subroutine to calculate, for a specified set of parameter values, the value of the function to be minimized, exactly as for search methods. The other category requires that, in addition, the user supply a subroutine to calculate the first derivatives of the function to be minimized. The latter type allows gradients to be calculated faster, but is much less flexible for the user, because for each function that is to be fitted the user must differentiate it algebraically and write a subroutine to evaluate these derivatives, which may be quite complicated.

The gradient methods usually take fewer iterations to converge and so may be much faster than search methods. On the other hand, they often converge less reliably and require better initial guesses, and it may be difficult or impossible to impose the required constraints on the fit with this sort of method, as exemplified above.

6.4. The Minimum-χ^2 Method

This method is really obsolete, but it will be described here because it has been quite widely used in the past and will give satisfactory results if the data are good enough. In order for this method to be used, the observations must first be grouped into a histogram. The data for the fitting are the frequencies of the observations in each bin. Thus, the parameter values will depend, to some extent, on the bin widths that are chosen to display the histogram (this is not the case with the maximum-likelihood method). The observed number of values in the jth bin will be denoted f_j^{obs}. The χ^2 statistic is a measure of the deviation of this observed value from the fitted (or calculated or expected) frequency. The value of the expected frequency depends, of course, on the values of the parameters (time constants, etc.) that are chosen, so we shall denote it $f_j(\theta)$ where θ represents the values of all the parameters. For example, when fitting two exponentials, the parameters could be τ_1, τ_2, and a_1, so $\theta = [\tau_1 \ \tau_2 \ a_1]$ (this is, in the notation of the appendix to Chapter 20, this volume, a *vector,* but it can be read here as a set of parameter values). The expected frequency is calculated from the equation for the distribution (e.g., equation 55), which, as discussed in Section 5.1.5, is approximately proportional to the frequencies if the bin width is not too wide. The values of the parameters are adjusted (by the optimizing program) to minimize χ^2, i.e., to minimize the badness of fit.

The χ^2 statistic is defined as

$$\chi^2 = \sum_{j=1}^{n_{\text{bin}}} \frac{[f_j^{\text{obs}} - f_j(\theta)]^2}{f_j(\theta)} \tag{65}$$

where n_{bin} is the number of bins in the histogram. Notice that, as in any fitting procedure, the data are treated as constants, and the parameters are treated as variables.

This method can be regarded as a sort of weighted least-squares approach; the denominator would be an estimate of the variance (reciprocal weight) of the numerator for a Poisson-distributed variable (the observed frequency in a given bin should be multinomially distributed, and this may approximate a Poisson distribution).

The fact that the denominator depends on the values of the parameters slows down the fitting procedure, and sometimes a modified χ^2 method has been used in which the observed, rather than the expected, values are used in the denominator. In other words, the parameters are chosen to minimize

$$\sum_{j=1}^{n_{\text{bin}}} \frac{[f_j^{\text{obs}} - f_j(\theta)]^2}{f_j^{\text{obs}}}. \tag{66}$$

If an observed value, f_j^{obs}, is zero, it must be replaced by unity (or by an average value over nearby bins) to avoid division by zero.

The χ^2 criterion, though reasonable, is arbitrary. It is also clear that, in principle, some

information must be lost when the original time intervals are grouped into bins. For example, observations of 1.1 msec and 1.9 msec are treated as though they were both 1.5 msec if they are pooled into a bin from 1 to 2 msec. There is another, more natural way to fit the results that does not involve these disadvantages, namely, the method of maximum likelihood. This method also allows sensible estimates of error for the fitted parameters and is described next.

6.5. The Method of Maximum Likelihood: Background

When we have done an experiment and wish to choose the best values of the parameters, it seems sensible to ask what values of the parameters are, in the light of our data, the most probable. Although this may appear an innocent enough question, it has given rise to fierce debate for over three centuries. The debate still continues. The essential argument is about whether it is proper to talk about the probability of a hypothesis at all. If we measure durations of ion channel openings, we imagine that there is some real true value of the mean open time. Suppose our observed mean is 8 ms, and the true mean (which is never known of course) is 10 ms. The probability of the hypothesis that the true mean is 10 ms is unity; the probability that it is anything else (including 8 ms) is zero. Therefore, one cannot speak of the probability that the parameters have particular values (not, at least, if we wish to retain the familiar frequency interpretation of probability). Most people now think that the best way to circumvent this problem is to speak not of the probability of a hypothesis (given some data) but of the probability of getting the data (given an hypothesis). This latter probability was first used to measure the plausibility of hypotheses by Bernoulli in 1777; it was greatly developed and popularized by R. A. Fisher, who termed it *likelihood* from 1921 onwards.

The probability of observing the data, given a hypothesis, is just an ordinary probability distribution if the hypothesis is regarded as fixed and the data as varying. However, when we regard the data as fixed (as they are when we wish to analyze a particular experiment) and the hypothesis as varying, then this quantity no longer behaves like a probability, and we term it *the likelihood of the hypothesis.* In summary, denoting likelihood by Lik,

$$\text{Prob}[\text{data}|\text{hypothesis}] \equiv \text{Lik}[\text{hypothesis}|\text{data}] \tag{67}$$

In this expression, the vertical bar stands for "given" (see Section 2 of Chapter 18, this volume).

The method of maximum likelihood consists of varying the values of the parameters (the hypothesis) so as to maximize expression 67. Thus, we choose the parameter values that maximize the probability of observing our data.

This approach can be justified in two ways. The likelihood advocate would simply say that if you wish to use parameter values (such as minimum χ^2 values) that make the data less probable than his, then it is for you to justify your apparently perverse decision (see Edwards, 1972). Another approach is to examine closely the statistical properties of the method, which are, in most cases, at least as good as those of any other approach (see Rao, 1973).

Of course, in order to calculate the probability of getting the data, given some hypothetical parameter values, we need to know what probability distribution the observations follow. In most experimental work, this is not known with any certainty, so, although the method of least squares is actually the same as the method of maximum likelihood if errors follow a Gaussian distribution, the former term is usually used because knowledge of the distribution

is uncertain. However, with data of the sort we are discussing here, we do know about the distribution. It is the very thing that we look at and wish to fit; this is why maximum likelihood is a natural procedure to adopt.

6.6. Maximum Likelihood for a Simple Exponential Distribution

These ideas can most easily be made clear by discussing data that follow a simple exponential distribution before going on to more general cases.

The data consist, say, of n time intervals, which we can denote $t_1, t_2 \ldots, t_n$. These are fixed, and this list provides the data on which fitting is based. Histogram frequencies are not used, and the values obtained are quite independent of the bin width(s) that are chosen for the histogram. It is still necessary to construct a histogram in order to display the final results of the fit, and the appearence of the histogram will vary to some extent according to the bin width(s) that are chosen, but the fitted line will not.

What, given some hypothetical value of the time constant τ, is the probability of making these observations; in other words, what is the likelihood of this value of τ? The time values are (in principle) continuous variables, so we must use probability densities rather than probabilities (but this does not matter much because we only need something that is directly proportional to the likelihood). The simple exponential distribution can be written

$$f(t) = \tau^{-1} e^{-t/\tau} \qquad 0 < t < \infty, \tag{68}$$

so the probability (density) of making the first observation t_1 is

$$f(t_1) = \tau^{-1} e^{-t_1/\tau} \tag{69}$$

The probability of making all the observations (t_1 *and* t_2 *and* . . . *and* t_n) is, if the observations are independent, simply proportional to the product of the separate probability densities, and this is the likelihood of the specified value of τ. Thus,

$$\text{Lik}(\tau) = f(t_1) f(t_2) \cdots f(t_n). \tag{70}$$

It is more convenient to work with the logarithm of this quantity (so we get sums rather than products), and this log-likelihood is denoted $L(\tau)$. From equations 69 and 70, it is simply

$$L(\tau) = \sum_{i=1}^{n} \ln f(t_i) = n \ln(\tau^{-1}) - \tau^{-1} \sum_{i=1}^{n} t_i \tag{71}$$

This log-likelihood must, of course, reach its maximum at the same value of τ as does the likelihood (equation 70) itself. When $L(\tau)$ is plotted against various possible values of τ, it produces a curve like those shown in Fig. 19 (continuous lines). This curve summarizes all of the information that the data contain about τ. The curve goes through a maximum and the value of τ at the maximum is the value that makes the data most probable. It is the maximum-likelihood estimate (denoted $\hat{\tau}$) of the unknown true value of τ, i.e., of the true mean lifetime.

The position of the maximum can easily be found analytically in this simple case by differentiating equation 71 with respect to τ and equating the result to zero. This gives

$$\hat{\tau} = \sum_{i=1}^{n} t_i/n = \bar{t} \tag{72}$$

Not surprisingly, the estimate is simply the arithmetic mean of the observations.

With the help of equation 72, $L(\tau)$ from equation 71 can be written in a form that shows that the decline of the graph on either side of the maximum depends only on the ratio, $\tau/\hat{\tau}$, namely,

$$L(\tau) = L(\hat{\tau}) - n[\ln(\tau/\hat{\tau}) + (\tau/\hat{\tau})^{-1} - 1] \tag{73}$$

Consider next the (usual) case in which resolution is limited. Suppose that it is impossible to measure reliably any intervals (t_i values) less than some specified amount, t_{min} (see Sections 5.2, 6.8, and 6.11). Our observations are restricted to the range t_{min} to infinity. Therefore, rather than the simple exponential pdf in equation 68, we need the conditional pdf for t given that it is greater than t_{min}. To obtain this, we divide by the probability that an observation is greater than t_{min} (see Section 2 in Chapter 18, this volume), which, from equation 36, is simply $\exp(-t_{min}/\tau)$. This gives

$$f(t) = \frac{\tau^{-1}e^{-t/\tau}}{P(t > t_{min})} = \frac{\tau^{-1}e^{-t/\tau}}{e^{-t_{min}/\tau}} = \tau^{-1}e^{-(t-t_{min})/\tau} \tag{74}$$

The log-likelihood is therefore

$$L(\tau) = \sum_{i=1}^{n} \ln f(t_i) = n \ln(\tau^{-1}) - \tau^{-1} \sum_{i=1}^{n} (t_i - t_{min}) \tag{75}$$

Differentiating and equating to zero gives the maximum-likelihood estimate of the mean lifetime as

$$\hat{\tau} = \bar{t} - t_{min} \tag{76}$$

i.e., we subtract the lower limit t_{min} from the mean of the observations. This relationship was used by Neher and Steinbach (1978); it is generalized in Section 6.8. The same relationship can be obtained by noting that for an exponentially distributed variable with mean τ, the mean of all observations longer than t_{min} is simply $\tau + t_{min}$ (see the more general result following equation 85).

Once estimates of the parameters of the pdf have been found, we can estimate the true number (N) of observations, which includes those that have been missed because they are less than t_{min}. This is is done simply by dividing the observed number, n, by the probability that an observation is greater than t_{min}, i.e.,

$$N = \frac{n}{e^{-t_{min}/\tau}} \tag{77}$$

This is generalized below, in equation 87. The expected frequency in a bin between t and $t + \Delta t$ is then simply

$$N(e^{-t/\hat{\tau}} - e^{-(t+\Delta t)/\hat{\tau}}) \tag{78}$$

This can be compared directly with the observed frequency (see Section 5.1).

In these cases there was no need for iterative computer optimization because the maximum-likelihood estimates could easily be calculated explicitly from equation 72 or 76. This cannot be done in general (see Section 6.8).

Non-independent Observations

The multiplication in equation 70 is correct only if the observations are independent. This is not always true. It is quite common, for example, for open times to be correlated; in the case of the muscle nicotinic receptor a long opening tends to be followed by another long opening. The question of correlations is discussed in more detail in Sections 5.7 and 5.8 and particularly in Chapter 18 (this volume, Sections 10, 11, and 13). When such correlations are present, the estimates obtained by the methods described here will not be genuine maximum-likelihood estimates, and errors calculated for the estimates will, to some extent, be erroneous. The effect of correlations on the fitting process has never been investigated in detail. It seems unlikely that the effects will be serious, and the bias of the estimates is unlikely to be worse than that of genuine maximum-likelihood estimates.

6.7. Errors of Estimates: The Simple Exponential Case

Once an estimate ($\hat{\tau}$) of the mean lifetime is obtained, it is natural to ask how accurate this estimate is likely to be. Estimates of error calculated from within a single experiment are notoriously unreliable and overoptimistic. The only reasonable estimate of error is found by repeating the whole experiment several times. Nevertheless, internal error estimates may be useful as a warning when an attempt is made to extract more information than the data contain, or in cases where repetition of the experiment is impossible. Two ways of estimating errors follow naturally from the maximum-likelihood approach. They are discussed next for the simple exponential case and generalized below. (It should be noted that these are not the only ways in which errors can be assessed; there is no general agreement about how this should be done in nonlinear problems.)

6.7.1. Approximate Standard Deviations

The first approach is to attach some sort of standard deviation to the estimate, $\hat{\tau}$, that has been found. A standard approach is to calculate the observed information by differentiating $-L(\tau)$ twice and then substituting $\hat{\tau}$ for τ. From equation 71 or 75 we obtain

$$-\left[\frac{\partial^2 L(\tau)}{\partial \tau^2}\right]_{\tau=\hat{\tau}} = \frac{n}{\hat{\tau}^2} \approx \frac{1}{\text{var}(\hat{\tau})}. \tag{79}$$

The quantity in equation 79 has a simple interpretation. The second derivative measures the curvature of the graph (e.g., Fig. 19) near the maximum. If it is small, the graph is flat; i.e., the likelihood is rather insensitive to the exact value of τ; therefore, $\hat{\tau}$ is rather ill defined and has a large standard deviation. The reciprocal of expression 79 provides an estimate of

the variance of $\hat{\tau}$, and its square root is an estimate of the standard deviation of $\hat{\tau}$, denoted $s(\hat{\tau})$. Thus, we obtain

$$s(\hat{\tau}) \simeq \hat{\tau}/\sqrt{n} \tag{80}$$

It should be noted that the validity of this estimate of error depends entirely on the assumption that the observations really do come from a population described by a single exponential pdf, so that we are fitting the right thing. Insofar as this will never be exactly true, the estimate is optimistic (or even meaningless).

The standard deviation for $\hat{\tau}$ found above, as for any standard deviation, can be interpreted in terms of a confidence interval only if we know the distribution of $\hat{\tau}$ (i.e., what the distribution of $\hat{\tau}$ values would be if we had many such estimates). If we suppose that $\hat{\tau}$ has a Gaussian distribution, which, from the central limit theorem, will be approximately true when the number of observations is large, then an approximate 95% confidence interval for $\hat{\tau}$ might be calculated as $\hat{\tau} \pm 2$ standard deviations; i.e.,

$$\hat{\tau} \pm 2s(\hat{\tau}). \tag{81}$$

The imperfection of this approach can easily be illustrated by an extreme example. Suppose we have only three observations, and their mean indicates that $\hat{\tau} = 2$ ms. Then the standard deviation of the mean is estimated as $2/\sqrt{3} = 1.15$ ms. Now calculate a confidence interval for $\hat{\tau}$ by taking two standard deviations on either side of $\hat{\tau}$, i.e., 2 ± 2.3 ms or -0.3 ms to $+4.3$ ms. According to this calculation, a value of $\tau = -0.3$ ms for the true mean lifetime is compatible with the observations, although it is obvious that all negative values are actually quite impossible. One way of looking at the reason for this silly result is that intervals calculated in this way are necessarily symmetrical (the Gaussian distribution is symmetrical), but more realistic error limits, such as those described in the next section, will not generally be symmetrical.

This example may be thought not to matter much because we never use such small numbers of observations. However, in some cases, we do wish to calculate the mean of quite small numbers. Consider, for example, the "intermediate shut times" (with $\tau = 1.31$ ms) in Fig. 15. Their mean length is of interest, but even in a long experiment, not many values can be observed, so absurdities like that just illustrated can easily occur in practice. They can be avoided by the method described in Section 6.7.2.

Standard Deviations and Standard Errors

Since the time intervals, t_i, follow a simple exponential distribution in this case, the standard deviation of the individual observations should, on average, be equal to the mean lifetime (e.g., Colquhoun, 1971); i.e., $s(t_i) = \bar{t}$. The standard deviation of the mean of n lifetimes, often known as the standard error of the mean, is calculated as $s(t_i)/\sqrt{n}$, which, since $\hat{\tau} = \bar{t}$ in this case, is just the result obtained in equation 80, but here it was obtained *via* the rather general method of equation 79. When quantities like that in equation 80 are obtained, it is often asked whether they are standard deviations or standard errors. This question is based on a common misunderstanding, because these are not two separate things. In fact, there is only one sort of measure of variability involved, and that is the standard deviation. This measure can be applied to any sort of variable quantity, as an index of how variable it is. It can be applied, for example, to a set of measured time intervals, t_i, and it will measure how much they vary for one interval to another. It can equally be applied to the mean of n lifetimes to measure how much repeated measurements of such means vary.

Or it can be applied to a time constant of a distribution (a τ value, equation 30), as a measure of how much repeated measurements of that τ value will vary. The standard error, a term that perhaps causes more misunderstanding than any other in elementary statistics, is not a separate sort of thing but is merely a piece of jargon standing for "the standard deviation of the mean of n observations" or, more generally, for "the (predicted) standard deviation of any quantity derived from the raw observations." The term standard error of the mean is still worse—it is not only misleading but also tautologous. The valid distinction is not between standard deviation and standard error but between (1) standard deviations that are estimated directly from a set of replicate observations (e.g., a set of measurements of individual lifetimes), the scatter of which can be directly observed, and (2) standard deviations that are calculated indirectly (e.g., standard deviation of the mean, or the standard deviation of a τ value) when we have actually got only one value (for the mean or for τ). In order to understand what the standard deviation means in the latter cases, we need to consider the standard deviation as a measure of how scattered the values would be *if* the quantity in question (the mean, or the τ value) were repeatedly estimated under identical conditions.

6.7.2. Likelihood Intervals

The second approach to estimation of errors, the calculation of likelihood intervals, overcomes these problems. This is quite easy in the case of simple exponentials (but uses quite a lot of computer time in more complex cases; see Section 6.9). The method is simply illustrated by the graph of the log-likelihood function, $L(\tau)$, against τ shown in Fig. 19. The maximum on the graph is at $\tau = \hat{\tau}$, so it is $L(\hat{\tau})$. If a horizontal line is drawn at a fixed distance, m $\log_e$ units, below the maximum, it intersects the graph at two points, one below $\hat{\tau}$ and one above $\hat{\tau}$.

The values of τ at these intersection points, τ_{low} and τ_{high} say, are, more formally, the (two) solutions of

$$L(\tau) = L(\hat{\tau}) - m \tag{82}$$

The values of τ_{low} and τ_{high} are clearly both less likely than $\hat{\tau}$ to the same extent (m $\log_e$-likelihood units), so it seems that they are good candidates to provide limits for the uncertainty in $\hat{\tau}$. They are called m-unit likelihood intervals or support intervals (see Edwards, 1972).

Conventional confidence intervals have an exact probability associated with them, but this is generally not possible in nonlinear problems of the sort that we have. Consider, however, a Gaussian variable with mean μ. In this case, the curve $L(\mu)$ has a simple quadratic form with constant curvature, from equation 79, and $\hat{\mu}$ is simply the arithmetic mean. In this case, the m-unit likelihood interval is just the conventional confidence interval defined as μ plus or minus $(2m)^{1/2}$ standard deviations, i.e.,

$$\hat{\mu} \pm (2m)^{1/2} s(\hat{\mu}) \tag{83}$$

Thus, there is a correspondence between $m = 0.5$ limits and one-standard-deviation limits; similarly, there is a correspondence between $m = 2$ limits and two-standard-deviation limits, and between $m = 4.5$ limits and three-standard-deviation limits.

The likelihood curves for a simple exponential distribution from equation 71 are plotted as continuous lines in Fig. 19 for samples of size $n = 3$, $n = 10$, and $n = 100$. The dashed curves in Fig. 19 show the corresponding quadratic curves that are implicitly assumed in

the calculation of the approximate standard deviations. The values for error limits are tabulated in Table I (which is, like Fig. 19, normalized to unit value of $\hat{\tau}$). Thus, to return to the example that follows equation 81 with $\hat{\tau} = 2$ ms and $n = 3$, the two-unit likelihood interval, from Table I, is seen to be 2×0.379 to 2×4.16, i.e., 0.758 ms to 8.32 ms. These limits are unsymmetrical (from $\hat{\tau} - 1.242$ ms to $\hat{\tau} + 6.32$ ms), and are far more realistic limits than $\hat{\tau} - 2.3$ ms to $\hat{\tau} + 2.3$ ms, which were found from the "approximate standard deviation" approach.

It is clear from Fig. 19 and Table I that in the simple exponential case, approximate limits from equation 81 are quite satisfactory for samples of 100 or more, for which $\hat{\tau}$ has a nearly Gaussian distribution.

6.8. Maximum-Likelihood Estimates: The General Case

The case of a single exponential distribution has been discussed in Sections 6.6 and 6.7. The results given there generalize easily to any number of components.

In general terms, we denote the values of the parameters to be estimated (θ_1, θ_2, ...) by the symbol θ and denote the jth observation as y_j, so the n data values are y_1, y_2, ..., y_n. The probability (density) of a particular observation, y_1 say, given some trial values of the parameters, θ, is denoted $f(y_1|\theta)$. The probability of observing all of our particular data values is, for the specified θ, proportional to the product of all the individual probabilities (densities). This is, by definition, the likelihood of θ for our particular data. As before, we prefer to work with the logarithm of this quantity, which is

$$L(\theta) = \sum_{j=1}^{n} \ln f(y_j|\theta) \tag{84}$$

This can be calculated as soon as we specify the distribution explicitly. An optimizing computer routine can then find the values of the parameters that maximize $L(\theta)$; these are the maximum-likelihood estimates, and they are collectively denoted $\hat{\theta}$.

Table I. Likelihood Intervals[a] and Standard Deviations

	Sample size (n)			
$s(\hat{\tau}) = 1/\sqrt{n}$	3 0.577	10 0.316	100 0.100	Approximate probability[b]
$m = 0.5$	0.591, 1.89	0.741, 1.40	0.906, 1.11	
$\hat{\tau} \pm s(\hat{\tau})$	0.423, 1.58	0.684, 1.32	0.900, 1.10	0.68
$m = 2$	0.379, 4.16	0.564, 2.03	0.824, 1.23	
$\hat{\tau} \pm 2s(\hat{\tau})$	−0.155, 2.16	0.368, 1.63	0.800, 1.20	0.95
$m = 4.5$	0.260, 11.1	0.441, 3.08	0.751, 1.37	
$\hat{\tau} \pm 3s(\hat{\tau})$	−0.732, 2.73	0.051, 1.95	0.700, 1.30	0.997

[a]Comparison of m-unit likelihood intervals (from equation 82) and corresponding intervals calculated from approximate standard deviations (equation 81) for three sample sizes. The numbers in the table can be obtained from the graphs in Fig. 19-19, or by solving the equations. The numbers given are the limits on either side of $\hat{\tau} = 1.0$; they should be multiplied by the observed value of $\hat{\tau}$.

[b]This probability is based on the normal deviate by which the standard deviations are multiplied; use of Student's t statistic would give a better approximation.

This procedure can be made clearer if it is illustrated by the three most common sorts of distribution.

6.8.1. Mixtures of Exponentials

Distributions that have the form of a mixture of a number (k) of exponential densities are the most common; they have already been defined in Section 6.1. The parameters in this case are the time constants, $\tau_1, \tau_2, \ldots, \tau_k$, and the relative areas, $a_1, a_2, \ldots, a_{k-1}$. Alternatively, we could estimate the τ_i and the amplitudes w_i, or we could estimate the rate constants λ_i and the areas, a_i. It makes no difference which of these ways we choose, because, for example, $\hat{\tau}_i = 1/\hat{\lambda}_i$, so we get the same result whether the distribution is written in terms of rate constants or of time constants. However, the areas (a_i) are likely to be more nearly independent of the time constants than are the amplitudes (see also Section 6.10), so convergence may be easier if areas are estimated.

The distribution can be written, if we choose the time constants and areas as parameters, as in equation 55. Notice again that there are not $2k$ parameters but $2k - 1$, because the areas must add up to unity, as in equation 31.

Limiting the Fitted Range

In practice, limited frequency resolution means that nothing shorter than $t_{\min}$ can be measured; this limitation can be incorporated into the fitting procedure, as described in Section 6.6. Sometimes we may wish to exclude values below some $t_{\min}$ value that is greater than the resolution. We may also sometimes wish to exclude from the fit all values that are longer than some specified length $t_{\max}$ (e.g., to exclude a small number of exceptionally large values). Therefore, we need, in general, the conditional pdf, given that all the observations are between $t_{\min}$ and $t_{\max}$. This is given by

$$f(t) = \frac{\sum_{i=1}^{k} a_i \tau_i^{-1} e^{-t/\tau_i}}{P(t_{\min} < t < t_{\max})} \qquad (t_{\min} < t < t_{\max}), \tag{85}$$

which is a generalization of equation 74. The mean value of such censored observations is

$$E(t) = \frac{\sum_{i=1}^{k} a_i [(t_{\min} + \tau_i)e^{-t_{\min}/\tau_i} - (t_{\max} + \tau_i)e^{-t_{\max}/\tau_i}]}{P(t_{\min} < t < t_{\max})}$$

The denominator in these results is simply the probability that an observation with the distribution in equation 55 lies between $t_{\min}$ and $t_{\max}$, namely, from equation 36,

$$P(t_{\min} < t < t_{\max}) = \sum_{i=1}^{k} a_i (e^{-t_{\min}/\tau_i} - e^{-t_{\max}/\tau_i}) \tag{86}$$

The observations consist of n measured time intervals $t_1, t_2, \ldots, t_n$. Equation 85 can be evaluated for each of these in turn using some particular trial values (θ) of the parameters.

The logarithms of these values are added to give, from equation 84, the value of $L(\theta)$. The optimizing program then adjusts the parameter values so as to maximize $L(\theta)$. The values of parameters that do this are the maximum-likelihood estimates $\hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{a}_1, \hat{a}_2, \ldots$ An estimate of the true number of observations, N (including those shorter than t_{min} or longer than t_{max}), can then be obtained from the observed number, n, as in equation 77:

$$N = \frac{n}{P(t_{min} < t < t_{max})} \tag{87}$$

where the denominator is as given by equation 86, with $\hat{\tau}_i$, $\hat{a}_i$ substituted for τ_i, a_i. A numerical example is considered in Section 6.10.

6.8.2. Mixtures of Geometric Distributions

In general, the distribution of the number of openings per burst, and similar quantities, is expected to be a mixture of one or more (k, say) geometric distributions of the sort defined already in equation 58 (see also Chapter 18, this volume). The distribution gives the probability of observing r (openings per burst, for example), and it can be written in a number of different ways. Alternative forms are given in equations 58, 59, and 63. In general, we may wish to include in the fitting process only those observed values that are between r_{min} and r_{max} inclusive. Thus, as in the exponential case, we need the conditional distribution, which, from equation 58, is:

$$P(r) = \frac{\sum_{i=1}^{k} a_i(1 - \rho_i)\rho_i^{-1}}{P(r_{min} \le r \le r_{max})}, \qquad (r_{min} \le r \le r_{max}) \tag{88}$$

From equation 64, the denominator is given by

$$P(r_{min} \le r \le r_{max}) = \sum_{i=1}^{k} a_i(\rho_i^{r_{min}-1} - \rho_i^{r_{max}}) \tag{89}$$

The data consist of a series of n observations of the variable r, which we can denote $r_1, r_2, \ldots, r_n$. These might be, for example, the number of openings observed in n different bursts. The probability of observing all of these values is given by the product of the $P(r_j)$ values, so the log-likelihood is

$$L(\theta) = \sum_{j=1}^{n} \ln P(r_j) \tag{90}$$

where $P(r_j)$ is calculated from equation 88 for particular values of the parameters (a_i and p_i), which are collectively denoted θ. The optimizing program adjusts the values of these parameters until $L(\theta)$ is maximized, as usual. If there is only a single component, there is only one parameter, and if all observations are included ($r_{min} = 1$, $r_{max} = \infty$), then $L(\theta)$ can be maximized analytically in this case. This gives the maximum-likelihood estimate of the

mean, $\hat{\mu}$, simply as $\bar{r}$, the arithmetic mean of the observations, and hence, from equation 61, $\hat{\rho} = 1 - (1/\hat{\mu}) = 1 - (1/\bar{r})$.

An estimate of the true number of observations, N (including those below r_{min} or greater than r_{max}), can then be obtained from the observed number, n, thus:

$$N = \frac{n}{P(r_{min} \leq r \leq r_{max})} \tag{91}$$

where the denominator is given by equation 89, with $\hat{\rho}_i$, $\hat{a}_i$ substituted for ρ_i, a_i.

6.8.3. Mixtures of Gaussian Distributions

The principles are exactly the same as in the other cases. Suppose that the variable y (usually a single-channel amplitude measurement in the present context) has a Gaussian distribution with mean μ and standard deviation σ. Its probability density function is

$$f(y) = \frac{1}{\sigma(2\pi)^{1/2}} e^{-u^2/2} \tag{92}$$

where

$$u = \frac{(y - \mu)}{\sigma} \tag{93}$$

is the "standard Gaussian deviate."

A mixture of k Gaussians is, therefore,

$$f(y) = \sum_{i=1}^{k} a_i f_i(y) \tag{94}$$

where $f_i(y)$ represents the Gaussian in equation 92 with mean μ_i and standard deviation σ_i, and a_i are the relative areas of the components.

The cumulative form of the Gaussian distribution, the probability that y is less than some specified value, y_1 say, is the integral of $f(y)$,

$$P(y \leq y_1) = \int_{y=-\infty}^{y_1} f(y)dy \tag{95}$$

Unlike the other cumulative distributions given above, this one cannot be written in an explicit form. However, it is easy to calculate values for it in a computer program, since all mathematical function libraries contain routines to calculate values of the error function, erf(x) (see also Appendix 3). The cumulative Gaussian distribution is simply related to the error function, thus:

$$P(y \leq y_1) = 0.5[1 + \text{erf}(u_1/\sqrt{2})] \tag{96}$$

where $u_1 = (y_1 - \mu)/\sigma$.

We shall often want to fit constants over a restricted range of values, excluding values below y_{min} and values greater than y_{max}. Again, we need the distribution of y conditional on y being between y_{min} and y_{max}. This is given by dividing $f(y)$, from equations 92 and 94, by $P(y_{min} < y < y_{max})$, which, from equation 96, can be calculated as

$$P(y_{min} < y < y_{max}) = 0.5 \sum_{i=1}^{k} a_i [\text{erf}(u_i^{max}/\sqrt{2}) - \text{erf}(u_i^{min}/\sqrt{2})] \tag{97}$$

where

$$u_i^{max} = \frac{(y_{max} - \mu_i)}{\sigma_i} \quad \text{and} \quad u_i^{min} = \frac{(y_{min} - \mu_i)}{\sigma_i} \tag{98}$$

The distribution of y, conditional on y being between y_{min} and y_{max}, is therefore

$$f(y|y_{min} < y < y_{max}) = \frac{f(y)}{P(y_{min} < y < y_{max})} \tag{99}$$

where $f(y)$ is given by equation 94 and the denominator is given by equation 97.

The data consist of a series of n observations of the variable y (e.g., channel amplitudes), which we can denote $y_1, y_2, \ldots, y_n$. The probability of observing all of these values is given by the product of the $f(y_j)$ values, so the log-likelihood is

$$L(\theta) = \sum_{j=1}^{n} \ln f(y_j) \tag{100}$$

where $f(y_j)$ is calculated from equation 99 for particular values of the parameters (a_i, μ_i, and σ_i), which are collectively denoted θ. In the case of Gaussian fits, there are $3k - 1$ parameters to be estimated. In cases where components overlap too much for all of these parameters to be estimated successfully, it may be helpful to constrain the standard deviation to be the same for all k components. In this case, there will be $2k$ parameters to be estimated, namely, k values of the means (μ_i), $k - 1$ values for the areas (a_i), and one value of σ. The optimizing program adjusts the values of these parameters until $L(\theta)$ is maximized, as usual.

An estimate of the true number of observations, N (including those below y_{min} or greater than y_{max}), can then be obtained from the observed number, n, as before, from

$$N = \frac{n}{P(y_{min} < y < y_{max})} \tag{101}$$

where the denominator is given by equation 97 with the maximum-likelihood values substituted for the parameters.

6.8.4. Binned Maximum-Likelihood Fits

The full maximum-likelihood fitting method is quite fast enough for it to be feasible, on a fast PC, to fit up to, say, five exponential components to several thousand intervals.

With more components or more data (or a slow computer), the full fit may become inconveniently slow. If a faster method is really necessary, the binned maximum likelihood method (Sigworth and Sine, 1987) should be used. In this method we use, to calculate the likelihood, not the probability (density) of observing a particular interval (given a set of parameter values) but, rather, the probability that our particular bin frequencies will be as observed. The values for the fitted parameters will, therefore, no longer be independent of how the bin boundaries are chosen. However, it has been shown, for logarithmically binned data (see Section 5.1.3), that the results are likely to be close to those from the full maximum-likelihood fit if at least 8–16 bins per decade are used (Sigworth and Sine, 1987).

The quantity to be maximized, the "binned log-likelihood," can be written in the form

$$L(\theta) = \sum_{j=1}^{n_{\text{bin}}} n_j \ln\left[\frac{F(t_{j+1}) - F(t_j)}{P(t_{\min} < t < t_{\max})}\right] \tag{102}$$

where the number of terms summed is now the number of bins, n_{bin} (rather than the total number of intervals), n_j is the number of observations in the jth bin, and t_j is the lower boundary of the jth bin. The numerator of this expression uses the cumulative distribution, $F(t)$, as given in equation 35 or 36 to calculate the probability, for the specified parameter values, θ, that an observation lies in the jth bin. The denominator, which was defined in equation 86, gives the probability that an observation is within the fitted range, $t_{\min}$ to $t_{\max}$, the values of which must, in this case, correspond to bin boundaries.

6.9. Errors of Estimation in the General Case

The treatment in Section 6.7 can be generalized with the help of matrix notation, so that the two sorts of error calculation can be calculated for distributions with any number of parameters. An introduction to this notation is given in Chapter 20 (this volume). Further details can be found in Box and Coutie (1956), Beale (1960), Bliss and James (1966), Edwards (1972), and Colquhoun (1979). The following procedures are reasonable approaches to the specification of errors, but they are not unique.

6.9.1. Approximate Standard Deviations

Denote the parameters, ν in number, as $\theta = (\theta_1, \theta_2, \ldots, \theta_\nu)$. The analogue of equation 79 is the observed information matrix, $\mathbf{I}(\theta)$, which is a $\nu \times \nu$ matrix with elements

$$-\left(\frac{\partial^2 L(\theta)}{\partial\theta_i \partial\theta_j}\right)_{\theta=\hat{\theta}} \tag{103}$$

This form is known as a Hessian matrix. The inverse of this matrix gives the covariance matrix, $\mathbf{C}(\theta)$, of the parameter estimates, so

$$\mathbf{C}(\theta) \simeq \mathbf{I}(\theta)^{-1} \tag{104}$$

The elements of this matrix may be denoted $\text{cov}(\theta_i,\theta_j)$. The diagonal elements of $\mathbf{C}(\theta)$ give estimates of the variances of the parameter estimates, θ_i. Thus,

$$\text{var}(\theta_i) = \text{cov}(\theta_i,\theta_i) \tag{105}$$

The square root of this gives the approximate standard deviation of the parameter estimate, θ_i. The off-diagonal elements ($i = j$) give the covariances of these estimates. These measure the tendency of the estimate of θ_i to be large if the estimate of θ_j happens to be large (see Section 6.10 for examples). This tendency is more conveniently expressed as a correlation coefficient, r_{ij}, between the two estimates; this can be calculated as

$$r_{ij} = \frac{\text{cov}(\theta_i,\theta_j)}{[\text{var}(\theta_i)\text{var}(\theta_j)]^{1/2}} \tag{106}$$

It has been noted that if we fit the sum of k exponentials, only $k - 1$ areas ($a_1, \ldots, a_{k-1}$, say) are estimated. The area, a_k, for the kth component follows immediately from the fact that the total area for the pdf is unity:

$$a_k = 1 - \sum_{i=1}^{k-1} a_i. \tag{107}$$

A standard deviation can be attached to a_k by the relationship

$$\text{var}(a_k) = \sum_{i=1}^{k-1} \text{var}(a_i) + 2 \sum_{i=1}^{k-1} \sum_{\substack{j=1 \\ j<i}}^{k-1} \text{cov}(a_i,a_j) \tag{108}$$

The right-hand side of this equation is simply the sum of all the elements in those rows and columns of $\mathbf{C}(\theta)$ that refer to the $k - 1$ estimates of areas. If there are only two components, it reduces to $\text{var}(a_2) = \text{var}(a_1)$. For three components it reduces to $\text{var}(a_3) = \text{var}(a_1) + \text{var}(a_2) + 2\text{cov}(a_1,a_2)$.

Explicit algebraic derivation of equation 103 or 104 would be a formidable task in all but the simplest cases, but, fortunately, it is not necessary. The second derivatives in equation 103 can be estimated by standard numerical methods as long as we have a subroutine to calculate $L(\theta)$ for specified values of the parameters. The Hessian so found can be inverted numerically by means of a matrix-inversion routine (see Chapter 20, this volume) to give the covariance matrix according to equation 104.

6.9.2. Likelihood Intervals and Likelihood Regions

Likelihood intervals can also be calculated in the general case, and this is probably one of the best ways of expressing errors for the parameters taken one at a time. In principle, it would be better to calculate a joint likelihood region for all k parameters, but such a k-dimensional region cannot be simply represented when $k > 3$. An example of a joint likelihood

for the case where two parameters are estimated is shown in Fig. 20 (see also Colquhoun, 1979). The graph shows a contour for $L(\theta) = L(\hat{\theta}) - 2$, so any pair of parameter values, θ_1 and θ_2, that lie on this contour are 2 log-likelihood units less likely than the best estimates, $\hat{\theta}_1$ and $\hat{\theta}_2$. The obliqueness of the contour shows that the estimates of θ_1 and θ_2 are positively correlated in this case; i.e., if both θ_1 and θ_2 were decreased, or both were increased, the fit would be little worse; i.e., $L(\theta)$ would be reduced only slightly. This may be compared with the effect of increasing θ_1 and decreasing θ_2 (or vice versa); this would cause the fit to become much worse. The tangents to the contour are also shown in Fig. 20; they define (see text) 2-unit limits for $\hat{\theta}_1$ and $\hat{\theta}_2$ *separately.* When the parameter estimates are correlated, as in this example, these limits for the individual parameters are, in a sense, pessimistic: if, for example, the true value of θ_1 were actually near $\hat{\theta}_1^{low}$, the correlation makes it improbable that the true value of θ_2 would be near $\hat{\theta}_2^{high}$. In order properly to take into account the correlation between the parameter estimates, a joint likelihood region (the contour in Fig. 20) is preferable. Points outside this region define pairs of θ_1, θ_2 values that are unlikely.

The numerical calculations that are needed to calculate likelihood regions or intervals take a good deal longer than those for the approximate standard deviations but are perfectly feasible on fast personal computers. The principle is very simple. The m-unit likelihood limits (see Section 6.7 for explanation of this term) for a particular parameter, θ_1, say, are defined as the values of θ_1 such that, if θ_1 is held constant at that value, and the likelihood $L(\theta)$ is maximized again, allowing all the parameters except θ_1 to vary freely, then the maximum value of $L(\theta)$ that can be attained is $L(\hat{\theta}) - m$; i.e., it is m units less than the true maximum, $L(\hat{\theta})$, which is attained when all of the parameters are allowed to vary.

In order to calculate the lower or upper limit for θ_1, iterative procedures are used. An initial guess is made, and the minimization is performed with θ_1 fixed at this value; if the maximum attained is not $L(\hat{\theta}) - m$, then the whole process is repeated by any standard iterative method (e.g., bisection or Newton–Raphson). This process is illustrated graphically

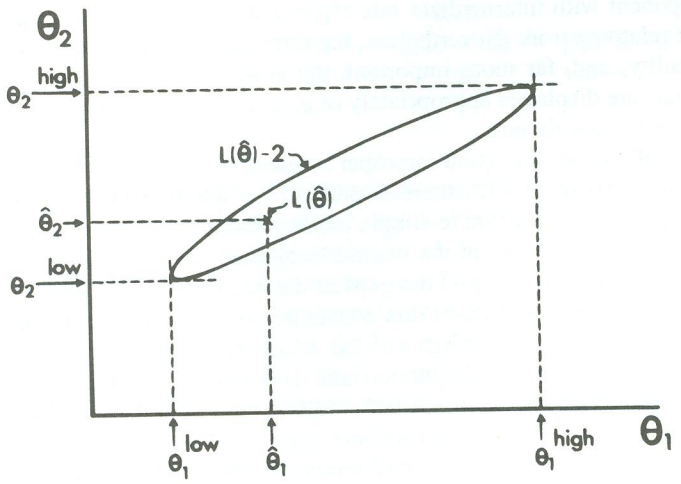


Figure 20. Schematic illustration of a joint likelihood region and of likelihood limits for the separate parameters in a case in which there are two parameters, θ_1 and θ_2, so $\theta = (\theta_1\ \theta_2)$. The graph shows a contour map of $L(\theta)$ with the peak of the hill, $L(\hat{\theta})$, marked with a cross; this corresponds with the maximum-likelihood estimates, $\hat{\theta}_1$ and $\hat{\theta}_2$, of the two parameters, as shown. The contour for $L(\hat{\theta}) - 2$ is shown. Further explanation is given in the text.

for the case when there are two parameters in Fig. 20. A numerical example is illustrated in Fig. 21.

If there are two components, the likelihood limits for a_2 are simply unity minus the limits for a_1. If there are more than two components (cf. equations 107 and 108 above), then in order to find limits for a_k it is necessary to refit the whole curve, so that a_k becomes one of the parameters that is estimated rather than the one inferred from the fact that the total area is unity.

6.10. Numerical Example of Fitting of Exponentials

The simultaneous fit of a triple-exponential pdf can be illustrated by data on shut times that were obtained with a low concentration (100 nM) of suberyldicholine (*R. temporaria*, cutaneous pectoris endplate $E_m = -123$ mV, 10°C). The results are similar to those shown in Fig. 15. The total number of openings fitted was 1021, but after imposition (see Section 5.2) of a minimum resolvable time of 50 μs (for both openings and gaps) and elimination of a few shut times that were unusable (because, for example, they contained ambiguous openings or simultaneous openings of more than one channel), the number of shut times to be fitted was 934. It was decided (see Section 5.2) to fit all durations between $t_{min} = 50$ μs and $t_{max} = 2000$ s, a total of 931 shut times.

The estimates of the time constants for the three components, found by maximizing $L(\theta)$ from equation 84 with $f(t)$ given by equation 85 were $\hat{\tau}_1 = 45.2$ μs $\hat{\tau}_2 = 1.28$ ms and $\hat{\tau}_3 = 440$ ms. The areas under the pdf accounted for by these components were, respectively, 74.0% (i.e. $\hat{a}_1 = 0.740$), 2.3% ($\hat{a}_2 = 0.023$), and 23.7% ($\hat{a}_3 = 0.237$). The maximum value of $L(\theta)$ attained was $L(\hat{\theta}) = -2899.33$. The fitted curve resembles that shown in Fig. 15. This fit implies, from equations 86 and 87, that the true number of shut times is $N = 1860.0$, of which 931 are in the observed range (the data), 922.3 are shorter than 50 μs, and 4.7 are above 2000 s.

The component with intermediate rate ($\hat{\tau}_2 = 1.28$ ms) is quite small and, as expected, has the largest relative errors. Nevertheless, the error calculations below give no real reason to doubt its reality; and, far more important, the need for this component is visible to the eye when the data are displayed appropriately (e.g., as in Fig. 15B or D), and it is reproducible from experiment to experiment.

In general, of course, it is quite improper to speak of short gaps, intermediate gaps, and long gaps on the basis of this fit; there is one pdf (which happens to be described by the sum of three exponentials), not three simple exponential pdfs. At least, it is improper unless we *define* the term "short gaps" in the manner suggested below, in which case the problem arises only when we wish to interpret the gaps so defined in terms of dwell times in particular states or sets of states. In some cases this convenient terminology can be justified, but only insofar as separate physical meanings can be attached, as an approximation, to the three components (see, for example, Colquhoun and Hawkes 1982; Chapter 18, this volume). Insofar as such an interpretation is valid, the data suggest that there are $N_f = \hat{a}_1 N = 1376$ "short gaps," $N_m = \hat{a}_2 N = 42.8$ "intermediate gaps," and $N_S = \hat{a}_3 N = 440.8$ "long gaps." Of the "short gaps," only $N_f e^{-50/45.2} = 455$ would be above 50 μs and therefore detectable.

First consider the errors for these estimates found by the approximate standard deviation approach (see Section 6.9.1). The second derivatives in equation 103 were estimated numerically; reasonable numerical accuracy is obtained by incrementing the parameters by ±10% from the maximum-likelihood values given above or by incrementing each parameter by enough to decrease $L(\theta)$ by 0.1. This provides an estimate of the observed information matrix.

Table II. Analysis of the Triple-Exponential Fit to Shut Time Duration[a]

Parameter	ML estimate $\hat{\theta}$	Approx SD $s(\hat{\theta})$	Likelihood intervals $m = 0.5$	Likelihood intervals $m = 2$	$2s(\hat{\theta})$
τ_1 (μs)	45.2	2.4	42.9–47.7 (−2.3 to +2.5)	40.6–50.5 (−4.6 to +5.3)	4.8
$100a_1$ (%)	74.0	1.6	72.2–75.5 (−1.8 to +1.5)	70.5–77.1 (−3.5 to +3.1)	3.1
τ_2 (ms)	1.28	0.42	0.90–1.76 (−0.38 to +0.48)	0.67–2.45 (−0.61 to +1.17)	0.84
$100a_2$ (%)	2.29	0.43	1.93–2.74 (−0.36 to +0.45)	1.55–3.32 (−0.74 to +1.03)	0.86
τ_3 (ms)	440.0	24.0	418–466 (−22 to +26)	396–494 (−44 to +54)	48.0
$100a_2$ (5)	23.7	1.5	22.3–25.4 (−1.4 to +1.7)	20.8–27.0 (−2.9 to +3.3)	3.0

[a]The maximum likelihood estimate, $\hat{\theta}$, of each parameter is given, with its approximate standard deviation, $s(\hat{\theta})$. Likelihood intervals are given in the form of intervals, and also, in parentheses, in the form of the deviation from $\hat{\theta}$. This deviation may be compared with $s(\hat{\theta})$ for the $m = 0.5$ unit intervals, and with $2s(\hat{\theta})$ (which is listed in the last column) for the $m = 2$ unit intervals.

This is then inverted numerically by means of any standard matrix-inversion subroutine to give the covariance matrix (equation 104) as follows (it is symmetric, so only the lower part is given):

$$\operatorname{cov}(\theta) \simeq \begin{array}{c} \begin{matrix} \tau_1 & a_1 & \tau_2 & a_2 & \tau_3 \end{matrix} \\ \begin{bmatrix} 5.73 \times 10^{-6} & & & & \\ -2.20 \times 10^{-5} & 2.44 \times 10^{-4} & & & \\ 2.51 \times 10^{-4} & -1.97 \times 10^{-4} & 0.18 & & \\ 9.07 \times 10^{-8} & -1.88 \times 10^{-5} & -8.40 \times 10^{-5} & 1.81 \times 10^{-5} & \\ 1.50 \times 10^{-3} & -1.21 \times 10^{-2} & 1.28 & 3.40 \times 10^{-3} & 578.3 \end{bmatrix} \end{array} \begin{matrix} \tau_1 \\ a_1 \\ \tau_2 \\ a_2 \\ \tau_3 \end{matrix} \tag{109}$$

The diagonal elements of this give the approximate variances of the parameter estimates (the order of the parameters is shown above, and to the right of, the matrix). The square roots of these variances are the standard deviations of the estimates and are shown in Table II. For example, for $\hat{\tau}_2$ the standard deviation is $s(\hat{\tau}_2) = (0.18)^{1/2} = 0.42$. The standard deviation for the area of the slowest component ($\hat{a}_3$) is obtained from equation 108 as $\operatorname{var}(\hat{a}_3) = 2.44 \times 10^{-4} + 1.88 \times 10^{-5} + 2(-1.88 \times 10^{-5}) = 2.25 \times 10^{-4}$, so the standard deviation for $\hat{a}_3$ is $(2.25 \times 10^{-4})^{1/2} = 1.5 \times 10^{-2}$, or 1.5%, as shown in Table II.

The correlation matrix is found from equation 109 by means of equation 106. It is

$$\begin{array}{c} \begin{matrix} \tau_1 & a_1 & \tau_2 & a_2 & \tau_3 \end{matrix} \\ \begin{bmatrix} — & & & & \\ -0.59 & — & & & \\ 0.25 & -0.03 & — & & \\ 0.009 & -0.28 & -0.05 & — & \\ 0.03 & -0.03 & 0.13 & 0.03 & — \end{bmatrix} \end{array} \begin{matrix} \tau_1 \\ a_1 \\ \tau_2 \\ a_2 \\ \tau_3 \end{matrix} \tag{110}$$

The correlation coefficient, for example, between the estimates of τ_1 and a_1 is $r_{21} = -2.20\times10^{-5}/[5.73\times10^{-6})(2.44\times10^{-4})]^{1/2} = -0.59$. This modest correlation is the strongest found; it reflects the intuitively obvious fact that the fit would be almost as good if τ_1 were decreased and a_1 increased, or *vice versa*. In other words, a faster time constant for the fast component would not reduce the goodness of fit much if the area allocated to this component were simultaneously increased (this implies a considerable increase in the amplitude of the fast component, $w_1 = a_1/\tau_1$: see equation (57). This correlation is aggravated by the lack of observations below 50 μs. There is also a small negative correlation (−0.28) between $\hat{a}_1$ and $\hat{a}_2$ and a small positive correlation (+0.25) between $\hat{\tau}_1$ and $\hat{\tau}_2$. Apart from these, the estimates are virtually independent. The fact that the slow component is well separated from, and nearly independent of, the other components means that a rough estimate of the standard deviation of its time constant can be calculated (compare equation 79) as $\hat{\tau}/\sqrt{N_S} = 21$ ms, which is not far from the value of 24 ms given by the full calculation (see Table II). For the small intermediate component, this approximation is, however, very poor; it gives $s(\hat{\tau}_2) = 0.19$ ms, compared with 0.42 ms from the full calculation.

The fact that only modest correlations are found for this fit is a good sign; it implies that the parameters are well-defined. If, for example, a strong positive correlation were found between two parameters, this would mean that if both were increased the quality of the fit would be little affected. In other words, the ratio of the two parameters is well defined, but their separate values are dubious.

The likelihood intervals for $m = 0.5$ and $m = 2.0$ (see Sections 6.7 and 6.9) are given for each parameter in Table II. It can be seen that the former are not far from what is expected from the approximate standard deviations, even for the small intermediate component, in this example (which has quite a large number of observations). The difference between the two approaches is larger in the case of the two-unit intervals, especially for the small component; for example, $\hat{\tau}_2 = 1.28$ ms, and $\hat{\tau}_2 \pm 2s(\hat{\tau}_2)$ implies an interval about $\hat{\tau}_2$ of $\hat{\tau}_2 - 0.84$ to $\hat{\tau}_2 + 0.84$ ms, whereas the two-unit likelihood interval gives $\hat{\tau}_2 - 0.61$ to $\hat{\tau}_2 + 1.17$ ms. The estimation of the limits for $\hat{\tau}_2$ is illustrated in Fig. 21 (see also Sections 6.7 and 6.9).

6.11. Effects of Limited Time Resolution

Virtually all experimental records contain intervals that are too short to be detected or measured, and this can cause serious distortion of distributions of open times, shut times, and number of openings per burst. The effect of missing brief events will be much less on distributions such as those of the burst length or the total open time per burst, so one way of dealing with the problem is to present only these distributions.

The practical aspects of this problem have already been described in Section 5.2.

The question of making corrections for missed events can be dealt with in two ways. The first, and most common, case occurs when no specific mechanism is being postulated for the channel under investigation. In this case it may be possible to make approximate corrections for missed events retrospectively. This can be done only in the case that *either* short openings *or* short gaps, but not both, are missed to any substantial extent. Such approximate corrections can also be done only when the kinetics of the observations are relatively simple. For example, if the distribution of (apparent) open times has more than one exponential component, then such corrections become difficult (though not necessarily impossible). Methods for making this sort of approximate correction are discussed, for

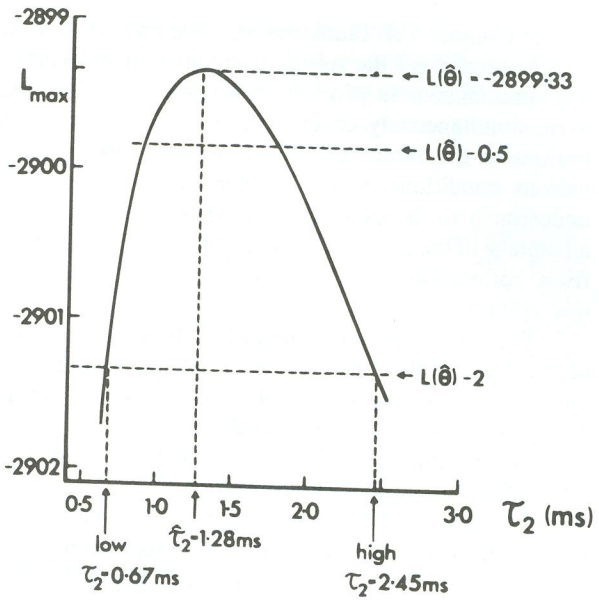


Figure 21. Estimation of likelihood intervals for τ_2 in the numerical example given in Section 6.10 (see Table II). The procedure is a generalization (to more than two parameters) of that illustrated in Fig. 20. The graph shows L_{max} plotted against τ_2, where L_{max} was found by holding τ_2 constant at the value shown on the abscissa and maximizing $L(\theta)$ with respect to the other four parameters (τ_1, a_1, a_2, and τ_3). The peak of the curve is, therefore, the overall maximum $L(\hat{\theta}) = -2899.33$ and corresponds to $\hat{\tau}_2 = 1.28$ ms. The values of τ_2 corresponding to $L_{max} = L(\hat{\theta}) - 2 = -2901.33$ are the 2-unit limits: 0.67 and 2.45 ms. The 0.5-unit limits can similarly be read off at $L(\hat{\theta}) - 0.5 = -2899.83$. In practice, it would be uneconomical to calculate this whole graph; the required points are found numerically by iteration (see text).

example, by Colquhoun and Sakmann (1985), and in Section 12 of Chapter 18 (this volume), and they are justified in more detail by Hawkes *et al.* (1992).

Exact corrections for missed events are possible only when a specific mechanism for channel operation is postulated. The methods that are available for doing exact corrections are discussed in Sections 12 and 13.7 of Chapter 18 (this volume). A particular benefit of these methods is that they have made it possible to fit reaction mechanisms directly to idealized data, as discussed next.

6.12. Direct Fitting of Mechanisms

The discussion so far has concerned the empirical fitting of exponentials (or geometrics) without specifying any particular reaction mechanism; the parameters to be fitted are the time constants and areas of the exponential components. Most investigations of reaction mechanisms have used such fits as the basis for a *post hoc* attempt to infer a mechanism. This procedure is less than ideal. One problem with it is that the information obtained from one sort of distribution may overlap strongly with that from another sort. For example, the distributions of burst length and of total open time per burst will be similar if the gaps within bursts are short (or rare). No method is known for combining the information from different sorts of fit in an optimal way to obtain the best idea about how well a specified mechanism fits the observations. Likewise, this approach makes it hard to compare two different putative mechanisms. Another problem with the *post hoc* approach is that, since each sort of distribution is fitted separately, the constraints on the relationship between them, which are implicit in the mechanism, are not taken into account.

Clearly, as mentioned in Section 6.1, it would be preferable to fit, as the adjustable parameters, not the time constants of the exponentials but the underlying rate constants in a specified mechanism (e.g., the values of k_{-1}, α_2, etc. in the mechanism specified in equation

110 of Chapter 18). Furthermore, since one set of values for these rate constants should be able to predict *all* the results from any sort of experiment, it is obviously preferable to do one simultaneous fit of all the observations that have been made. For example, it is desirable to fit, simultaneously, observations on steady-state records at several different agonist concentrations or membrane potentials, observations on channel openings following jumps under various conditions, and any other data that may have been obtained. Furthermore, it is undesirable to fit open times and shut times separately, because this procedure cannot take advantage of the information available from the sequence in which they occur (i.e., information from correlations—see Sections 5.7 and 5.8 above and Sections 10–13 of Chapter 18, this volume).

The sort of optimum approach to direct fitting just described was already well understood at the time of the first edition of this book (see Section 6.1.2 of Chapter 11 of the first edition), and attempts to implement direct fits had already been made (Horn and Lange, 1983). The problems were that the observations in the idealized record that are to be fitted suffer from omission of brief events and that retrospective corrections for missed events are not useful if a direct fit is to be attempted. Nothing very effective could be done until methods were devised to predict the distributions of what is *actually* observed rather than what would have been observed if time resolution had been perfect. Such methods now exist and are summarized in Sections 12 and 13.7 of Chapter 18 (this volume). They are now coming into use (e.g., Sine *et al.,* 1990). The approach is to calculate one value of the total likelihood from all the sets of data that are being fitted and to find the parameters that maximize this likelihood. The likelihood is calculated from the sequence of open and shut times rather than separately from each, so information from correlations is included in the fitting process. An example is given in Section 12.5 of Chapter 18 (this volume), and the general theory is summarized in Section 13.7 of Chapter 18 (this volume).

6.13. Fitting the Results after a Jump

The first problem is to get the results. Apart from the problem of estimating the number of channels, it is also the case that only one first latency can be measured for each jump, and it may be hard to get enough values in one experiment to make a decent-looking distribution. There will also be only one value per jump of each subsequent open and shut time if the first, second, etc. values differ (and this will not be known until their distributions have been looked at separately). It is perhaps for this reason that first latencies have often been displayed as cumulative distributions; the spurious appearance of precision that characterizes this sort of display (see Section 5.1.4) makes them look better than they are; this is highly undesirable.

Channel openings can be fitted by one of the methods already described, and a defined resolution can be imposed as described in Section 5.2 (this is especially desirable if the results are to be fitted with allowance for missed events). First latencies would then be corrected for recording delays (see preceding section). If the shut-time components are sufficiently well separated, it may be possible to define bursts of openings in the record. The theoretical distributions describing openings, shuttings, and bursts after a jump are given by Colquhoun and Hawkes (1987) in the case of a single channel and no missed events (see also Chapter 18, this volume). It is also possible to fit a mechanism directly, with allowance for missed events, as described for stationary records in Chapter 18 (this volume) and Section 6.11 (A. G. Hawkes, A. Jalali and D. Colquhoun, unpublished data).

When empirical mixtures of exponentials are being fitted to the first-latency distribution,

it should be remembered that the areas of some components may be negative, as explained and illustrated in Chapter 18 (this volume Section 11). It is therefore important to be sure that your fitting program does not constrain all the areas to be positive (see Section 6.1.3).

6.13.1. Latencies with *N* Channels

If more than one channel is present, the first latencies will, of course, appear to be shorter than they really are. In the case of the first-latency distribution (but not any of the others), it is relatively simple to correct the observations *if* the number of channels is known. When N independent channels are present, the observed first latency will be greater than t if the first latencies for all N individual channels are greater than t. Thus, from equation 36,

$$P(\text{all } N \text{ latencies} > t) = 1 - F_N(t) = [1 - F_1(t)]^N$$

where $F_1(t)$ is the probability, for one channel, that the latency is equal to or less than t (see Section 5.1.4), and the observed cumulative distribution provides an estimate of $F_N(t)$ (Aldrich *et al.*, 1983). The pdf of the first latency is the first derivative of $F_1(t)$, so if we denote the pdf for N channels as $f_N(t)$ we get (Colquhoun and Hawkes, 1987)

$$f_1(t) = \frac{f_N(t)}{N[1 - F(t)]^{N-1}}$$

6.13.2. Effect of Finite Sample Length

The rectangular pulse of voltage, or ligand concentration, will be of fixed finite length, and the length of the data record collected after the end of the pulse will usually also be of fixed length. There will, therefore, always be an incomplete interval at the end of each record; if the channel was shut at the end of the record, the length of the shutting is not known because the next opening has not been recorded, and conversely, if the channel is open at the end of the record, the length of this last opening is not known. But we do know, in either case, that the interval was *at least* as long as the bit of it that was observed. It is easy to take into account this information when doing maximum-likelihood fitting (with or without allowance for missed events). For all complete intervals of length t_i, the log-likelihood is found as $L = \Sigma \ln f(t_i)$ (see equation 84); the probability of observing an interval of length *at least* t is $1 - F(t)$, so a separate term, $\Sigma \ln[1 - F(t_i)]$, can be added to the log-likelihood for the incomplete intervals (of length t_i). We then maximize the sum of these two terms, which is the overall log-likelihood (Hoshi and Aldrich, 1988).

Appendix 1. Choice of the Threshold for Event Detection

The choice of the threshold setting that allows the detection of the briefest events was considered in Section 3.3. However, optimizing the detection of the shortest pulses is not necessarily the best strategy for detection of single-channel events, because one is interested in counting events of all widths. The ideal event detector would have a sharp transition at some width, w_{min}, such that events narrower than this would be missed but essentially all longer events would be counted. In practice, the transition, as seen in a graph of the probability

of detection as a function of w, is not necessarily sharpest when ϕ and f_c are chosen as described in Section 3.3.

Figure A1 demonstrates this property for pulses in the presence of $1 + f$ noise. The probability of detection p_{det} depends not only on w but also on ϕ, f_c, the channel amplitude, and the spectral characteristics. Part A of Fig. A1 corresponds to the case of a low channel amplitude (specifically, $A_0 = 0.22$ pA when the standard noise spectrum is assumed) in which events of width $w_{min} = 3$ ms or longer could be resolved. To construct each curve in the figure, a value of ϕ was first selected, and f_c was then chosen to give $\phi/\sigma_n = 5$. On the basis of these parameters, $p_{det}(w)$ was then estimated. The value of ϕ giving the best detection

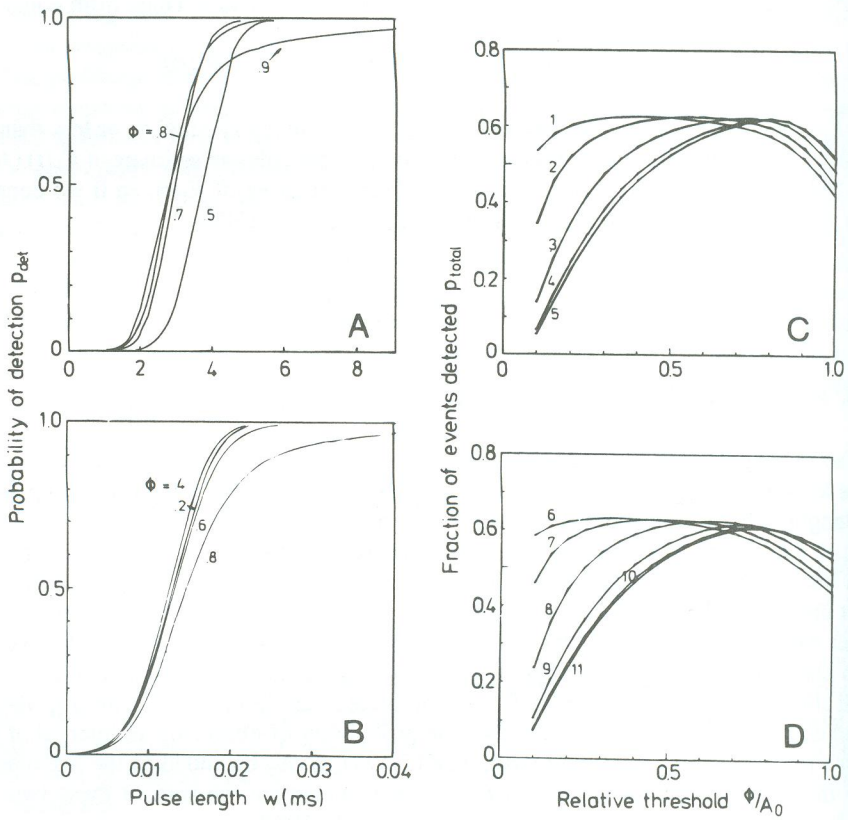


Figure A1. Performance of the event detector at various settings of the threshold ϕ. A and B: The probability of detection of isolated pulses of unit amplitude as a function of pulse width w. The parameters in A correspond to very small currents ($A_O = 0.22$ pA in $S_o = 10^{-30}$ A^2/Hz noise), giving $w_{min} \approx 3$ ms. At each value of ϕ, f_c was adjusted to give $\sigma_n = \phi/5$ to keep the false-event rate approximately constant. In B, the parameters correspond to relatively large currents ($A_o = 7.1$ pA in the same noise); much shorter events ($w_{min} \approx 13$ μs) can be detected. In this case, f_c was adjusted to keep $\phi/\sigma_n = 3$, corresponding to a higher false-event rate. C and D show the overall fraction, p_{total}, of pulses detected, given exponential distributions of pulse widths (equation A1). Each curve represents a different effective event amplitude, with the lowest-numbered curves corresponding to the largest amplitudes. Values for the amplitudes, time constants of the distribution, and other parameters are given in Table AI. In C, ϕ/σ_n was fixed at 5, whereas in D, $\phi/\sigma_n = 3$. The larger σ_n values in D cause the curves to be broadened and the optimum ϕ values to be slightly lower. Curves 4 and 6 were computed for the same conditions as in parts A and B, respectively.

of the shortest pulses was about 0.8 times A_0; when ϕ was reduced to 0.7 A_0, the transition moved to a slightly higher value of w but was steeper. On the other hand, increasing to 0.9 A_0 broadened the transition, so that whereas the very briefest pulses could be detected with higher probability, pulses even twice as long as w_{min} would be detected with only 80% probability. This sort of broadening of the transition region is very undesirable because it biases the selection of events in a way that can cause distortion of experimental lifetime distributions.

The broadening of the transition region is most severe when ϕ approaches A_0. This can be understood intuitively from the fact that if ϕ is near the full event amplitude, even moderately long events may fail to exceed ϕ when noise fluctuations are present. When ϕ is set lower (with f_c concurrently set lower), the longer events will have relatively larger peak amplitudes and will have a better chance of exceeding the threshold.

Figure A1B shows $p_{det}(w)$ curves for pulses of relatively large amplitude (A_0 = 7.1 pA in the standard noise spectrum; $\phi/\sigma_n = 3$). The minimum pulse width is about 13 μs in this case, and the optimum ϕ for detection of short pulses is much smaller, approximately 0.36 A_0. As the figure shows, the position and shape of the p_{det} curve depend only weakly on ϕ in the range 0.2 to 0.5 times A_0.

Parts C and D of Fig. A1 give a summary of the performance of an event detector in situations with various ratios of channel amplitude to background noise level. The quantity that is plotted here is the total fraction of events detected, p_{total}, out of an ensemble of pulse-shaped events having a probability density function $f(w)$ of widths,

$$p_{total} = \int_0^\infty p_{det}(w) f(w) dw \tag{A1}$$

where $f(w)$ was chosen to be exponential, $f(w) = (1/\tau) \exp(-w/\tau)$. For each curve, τ was fixed at the value $2w_{min}$; the actual values used are given in Table AI. The maximum values

Table A1. Parameters for the Curves in Fig. A1C,D[a]

					$\phi = 0.5\ A_0$		$\phi = 0.7\ A_0$	
Curve	ϕ/σ_n	$A_0^2/S_0 f_0$	A_0 (pA)	w_{min} (msec)	f_c (kHz)	p'_{total}	f_c (kHz)	p'_{total}
1	5	5000	7.1	0.023	7.62	1.00	10.94	0.96
2	5	500	2.2	0.089	2.00	1.00	3.02	0.99
3	5	50	0.71	0.43	0.375	0.95	0.641	1.00
4	5	5	0.22	3.84	0.046	0.87	0.086	0.98
5	5	0.5	0.07	38.4	0.0047	0.85	0.0092	0.98
6	3	5000	7.1	0.013	13.16	0.99	18.7	0.94
7	3	500	2.2	0.048	3.71	1.00	5.45	0.97
8	3	50	0.71	0.206	0.834	0.98	1.34	0.99
9	3	5	0.22	1.21	0.121	0.91	0.222	1.00
10	3	0.5	0.07	10.4	0.0129	0.88	0.025	1.00
11	3	0.05	0.02	104.0	0.0013	0.88	0.0025	1.00

[a]Part C was computed with $\phi/\sigma_n = 5$ (low false-event rate; curves 1–5), and D with $\phi/\sigma_n = 3$ (curves 6–11). Each curve represents a different value of the signal-to-noise parameter $A_0^2/S_0 f_0$, which corresponds to the given A_0 value in the standard case ($S_0 = 10^{-30}$ A^2/Hz, f_0 = 1 kHz, 1 + f spectrum). The w_{min} values give the effective minimum detectable pulse width. The distribution of pulse widths for calculating p_{total} was chosen to be exponential in each case, with the time constants $\tau = 2w_{min}$. For ϕ = 0.5 and 0.7, the corresponding f_c values and the relative detection efficiency $p'_{total} = p_{total}(\phi)/p_{total}(max)$ are given. The maximum value $p_{total}(max)$ was always within a few percent of $\exp(-w_{min}/\tau) = \exp(-1/2)$, the probability expected if only those events shorter than w_{min} were not detected.

of p_{total} computed in this way were near 0.6, which is to be expected since, if $p_{det}(w)$ were zero for $w < w_{min}$ and unity for all larger w, p_{total} would equal $\exp(-w_{min}/\tau) = 0.61$.

A comparison of the A_0 and w_{min} columns of Table AI shows the approximate limits of pulse detection, and the f_c columns show typical corresponding filter bandwidths. The choice of the ϕ/σ_n ratio equal to 3 instead of 5 allows pulses shorter by a factor of 2–3 to be detected, but at the cost of higher false-event rates. For large pulses ($A_0 > 1$ pA in this case), w_{min} decreases as $1/A_0$, whereas for smaller pulses, w_{min} varies as $1/A_0^2$. The A_0 values given correspond to the standard noise spectrum; for other $1 + f$ spectra, the dimensionless parameter $A_0^2/(S_0 f_0)$ is the appropriate measure for the signal-to-noise relationship, and w_{min} values should be scaled as $1/f_0$ for f_0 differing from 1 kHz.

Although this analysis has been quite complicated, the practical conclusions can be stated simply. First, for detecting channels of relatively low amplitude, implying that f_c must be set to be below f_0 (1 kHz in this example) to obtain a suitable background noise level, a good choice for ϕ is about $0.7A_0$. This is near the peaks of the corresponding p_{total} curves but is low enough to insure a sharp transition in the $p_{det}(w)$ curves. Second, for detecting larger channel events, for which f_c can be larger than f_0, the exact choice of ϕ is less critical, with the range 0.4 to $0.5A_0$ generally being best. The special case $\phi = 0.5\ A_0$ is of interest for event characterization. It can be seen from Fig. A1C and D that p_{total} is always at least 85% of its peak value when $\phi = 0.5\ A_0$ is chosen.

Appendix 2. The Expected Distribution of Fitted Amplitudes

We derive here the distribution of channel amplitudes that would be expected when amplitudes are estimated by averaging. Points are averaged over an interval w_a that lies within the "flat-top" portion of an event. This estimate, A, has an expected value (long-term average) equal to the true channel amplitude, A_0.

We assume that the background noise spectrum is flat and that the noise does not change appreciably when a channel opens. In this case, A has a variance that depends on w_a according to

$$\sigma_A^2(w_a) \approx S_0/w_a \tag{A2}$$

where S_0 is the (one-sided) spectral density. Strict equality holds in the limit when w_a is very large compared with the recording system risetime T_r, but the approximation is actually very good for all $w_a \geq T_r$. It is also a good approximation to the error in least-square fitting of the time course (Fig. 11B).

In practice, the background noise spectrum rises with frequency, but it is usually flat below 1 kHz. Since the frequencies that predominantly contribute to σ_A^2 are below $f = 1/2w_a$, for w_a on the order of 1 ms or larger the flat-spectrum assumption is usually justified, with S_0 being taken as the low-frequency spectral density.

Assuming that the baseline level is known exactly, σ_A^2 is the entire variance of the channel amplitude estimate. If we assume that the background noise is Gaussian distributed, the probability density of values of A for a given w_a is also Gaussian:

$$g_w(A;w_a) = \frac{1}{(2\pi)^{1/2}\sigma_A(w_a)} \exp\left[\frac{-(A - A_0)^2}{2\sigma_A^2(w_a)}\right] \tag{A3}$$

In practice, one does not want to hold the averaging interval constant but instead allows it to vary with the channel-open time, t_0. We assume the relationship

$$w_a = t_0 - t_m \qquad (t_0 \geq t_m) \tag{A4}$$

where t_m is the (fixed) length of an event that is "masked off" before averaging; this would typically be chosen to be between 1 and 2 risetimes in length to avoid any bias toward lower estimates as a result of the rising and falling edges of the pulse. Finally, we wish to ignore amplitude estimates from the briefest events by setting a lower limit w_{min} for averaging widths. The resulting pdf for the amplitude from an ensemble of events having random widths is then given by

$$g(A) = \int_{w_{min}}^{\infty} g_w(A;w_a) f(w_a) dw_a \tag{A5}$$

where $f(w_a)$ is the pdf of averaging widths. If t_0 is distributed according to a mixture of exponential densities, as in equation 30, then $f(w_a)$ is also multiexponential,

$$f(w_a) = \Sigma a_i \tau_i^{-1} e^{-w_a/\tau_i} \qquad w_a > 0 \tag{A6}$$

Substituting equations A6 and A3 into the integral A5 yields

$$g(A) = \frac{1}{(2\pi S_0)^{1/2}} \int_{w_{min}}^{\infty} w_a^{1/2} \exp\left[\frac{w_a(A - A_0)^2}{2S_0}\right] [\Sigma a_i \tau_i e^{-w_a/\tau_i}] dw_a \tag{A7}$$

It is helpful to change the variable of integration to $x_i = (w_a/\tau_i)^{1/2}$ and to introduce the definitions

$$\begin{aligned} x_{0i} &= (w_{min}/\tau_i)^{1/2} \\ \sigma_{0i} &= (S_0/\tau_i)^{1/2} \end{aligned} \tag{A8}$$

where x_{0i} is dimensionless and gives a measure of the spread of the distribution of w_a values, and σ_{0i} is the standard deviation of an amplitude estimate when $w_a = \tau_i$. Finally, we set $\delta_i = (A - A_0)/2^{1/2}\sigma_{0i}$ so that δ_i are the normalized deviations of A from its expected value. The integral can then be evaluated to yield

$$g(A) = \frac{1}{(2\pi)^{1/2}} \sum_{i=1}^{k} \frac{a_i}{\sigma_{0i}(1 + \delta_i^2)} \times \left\{ x_{0i} \exp[-x_{0i}^2(1 + \delta_i^2)] + \frac{\pi^{1/2}}{2(1 + \delta_i^2)^{1/2}} \operatorname{erfc}[x_{0i}(1 + \delta_i^2)^{1/2}] \right\} \tag{A9}$$

where erfc is the complementary error function. (A formula for numerically evaluating this function is given in Appendix 3).

Figure A2 shows plots of this distribution for various values of x_0 in the case where the open time has a simple exponential distribution with mean τ. Since σ_0 is kept constant, the figure demonstrates the effect of changing the duration limit w_{min} on the shape of the amplitude distribution obtained from a given set of single-channel events. When x_0 is larger than unity, the first term of equation A9 predominates, so that the distribution is essentially Gaussian in shape and has a standard deviation $\sigma_a \approx \sigma_0/x_0 = (S_0/w_{min})^{1/2}$. Large x_0 corresponds to the case in which w_{min} is large compared to τ, so that the distribution of w values dies off quickly beyond w_{min}. A nearly Gaussian amplitude distribution is therefore to be expected from the tightly clustered w_a values.

As x_0 decreases, the tails of the distribution become wider, and the distribution becomes distinctly non-Gaussian, but it remains symmetrical. To obtain the sharpest distribution, it is best to choose w_{min} (and therefore x_0) to be large. However, a high w_{min} value implies that fewer events will be counted in the amplitude histogram. A good compromise is to choose $w_{min} = \tau/2$, yielding $x_0^2 = 0.5$. This allows the fraction $\exp(-1/2) \simeq 0.6$ of the maximum number of events to be counted while yielding a distribution that is nearly indistinguishable from a Gaussian having the standard deviation $\sigma = 1.24\ \sigma_0$ (Fig. A2B).

Rather than computing the background noise power spectrum to determine S_0, it may be more convenient in practice to estimate σ_0^2 directly. This can be done by forming the averages of a large number of successive stretches, of length τ, of the background trace. The variance of these values can then be used directly as an estimate of σ_0^2.

Appendix 3. Numerical Techniques for Single-Channel Analysis

A3.1. A Digital Gaussian Filter

This digital filter forms output values y_i from input values x_i by forming a weighted sum

$$y_i = \sum_{j=-n}^{n} a_j x_{i-j} \tag{A10}$$

where the a_j are coefficients that sum to unity.

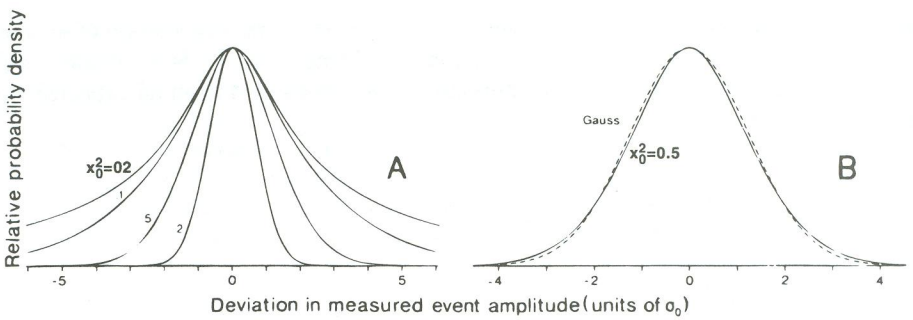


Figure A2. A: Plots of the function in equation A9 for various values of the parameter x_0^2, in the case where the open time has a simple exponential distribution. The plots were scaled to superimpose the peak values. B: Comparison of equation A9 with a Gaussian function. The parameter x_0 was chosen to be $1/\sqrt{2}$; the Gaussian function (dotted curve) was fitted by eye and had a standard deviation equal to $1.24\sigma_0$.

A continuous-time Gaussian filter is characterized by the width parameter or "standard deviation" σ_g of its impulse response, which is related to the cutoff frequency f_c according to (see equation 2)

$$\sigma_g = 0.1325/f_c \tag{A11}$$

Similarly, for a discrete filter, σ_g can be defined in units of sample intervals, in which case equation A11 holds if f_c is expressed in units of the sampling rate.

For a discrete Gaussian filter having width σ_g, the coefficients have the form

$$a_j = \frac{1}{\sqrt{2\pi}\sigma_g} \exp\left(\frac{-j^2}{2\sigma_g^2}\right) \tag{A12}$$

and the number of terms, n, is chosen so that the missing terms are negligible in size; in the implementation described here, n is chosen to be $4\sigma_g$.

If σ_g is relatively small, coefficients of the form of equation A12 sum to less than unity and yield a filter with wider bandwidth than f_c; these errors exceed 1% when σ_g is less than about 0.6. Since small σ_g corresponds to relatively light filtering, a suitable choice for the coefficients in this case is

$$\begin{aligned} a_1 &= \sigma_g^2/2 \\ a_0 &= 1 - 2a_1 \\ a_{-1} &= a_1 \end{aligned} \tag{A13}$$

so that each output value of the filter depends only on the corresponding input value and its two neighboring points. This simple filter function causes no problems with aliasing, provided the original data points are sampled at a sufficient rate, e.g., five times the cutoff frequency of Bessel-response prefiltering.

Filter procedures are presented in Fig. A3 for FORTRAN and in Fig. A4 for MODULA-2. The FORTRAN implementation operates on an array of integer input values and produces integer output; intermediate computations are however, performed in floating point. Note that because the number of coefficients n (this value is called NC in the FORTRAN subroutine, NumCoeffs in the MODULA-2 version) increases inversely as f_c, sufficient room in the coefficient array A should be provided for the smallest expected f_c value. For example, n = 53 for f_c = 0.01, but n = 5 for f_c = 0.1. The MODULA-2 implementation consists of two procedures, one to compute the coefficients and the other to perform the actual filtering. The latter, DoFilter, operates on real (floating-point) values and is capable of decimating the data, i.e., producing fewer output points than input points.

As an example of the use of these subroutines, suppose that we have a digitized record that was filtered with a Bessel filter at 2 kHz and sampled at a 10-kHz rate. To reduce the effective bandwidth to 1 kHz, the second filtering operation should have a cutoff frequency (see equation 4), of $(1 - 1/4)^{-1/2} = 1.15$ kHz. In calling the filter routine, the FC or Frequency variable should therefore be set to 0.115.

In both of the implementations shown, the evaluation of the sum (equation A10) is done only after checking that the input array bounds will not be exceeded; the result is that the values of the input points before the beginning and after the end of the input array are in effect assumed to be zero. Although the points in the middle of a long data array will not

```
      SUBROUTINE FILTER (IN, OUT, NP, FC)
C
C     Gaussian filter subroutine.  Accepts integer
C     data from the array IN, filters it with a -3db
C     frequency FC (in units of sampling frequency)
C     and returns the integer results in the OUT array.
C
      INTEGER IN(NP), OUT(NP)
      REAL A(54)
C           (Coefficient array.  54 terms are sufficient
C            for FC >= .01
C
C     -----First, calculate the coefficients-----
      SIGMA = 0.132505 / FC
      IF (SIGMA.LT. 0.62) GOTO 10
C
C        Standard gaussian coefficients.
C        NC is the number of coefficients not counting
C        the central one A(0).
      NC = INT( 4.0 * SIGMA )
      IF (NC .GT. 53) NC = 53
      B = -0.5 / ( SIGMA * SIGMA )
      A(1) = 1.0
      SUM = 0.5
C
      DO 5, I = 1, NC
      TEMP = EXP( (I*I) * B )
      A(I+1) = TEMP
      SUM = SUM + TEMP
 5    CONTINUE
C        Normalize the coefficients
      SUM = SUM * 2.0
      DO 7, I = 1, NC + 1
      A(I) = A(I) / SUM
 7    CONTINUE
      GOTO 20
C
C        Alternate routine for narrow impulse
C        response.  Only three terms are used.
10    A(2) = SIGMA * SIGMA / 2.0
      A(1) = 1.0 - 2.0 * A(2)
      NC = 1
C
C     -----Actual filtering is done here-----
20    DO 40, I = 1, NP
      JL = I - NC
      IF (JL .LT. 1) JL = 1
      JU = I + NC
      IF (JU .GT. NP) JU = NP
      SUM = 0.0
C
      DO 30, J = JL, JU
      K = IABS(J-I) + 1
      SUM = SUM + IN(J) * A(K)
30    CONTINUE
C
      OUT(I) = SUM
40    CONTINUE
      RETURN
      END
```

Figure A3

```
IMPLEMENTATION MODULE FilterReal;

FROM SYSTEM IMPORT ETOX; (* exp function *)
FROM InOut IMPORT WriteString, WriteInt, WriteLn;

CONST
  MaxFilterCoeffs = 220;

 (* Module global variables *)
VAR
  NumCoeffs : INTEGER;
  Coeffs    : ARRAY[0..MaxFilterCoeffs] OF REAL;


PROCEDURE SetGaussFilter ( Frequency: REAL );
 (* Load the filter coefficient values according to the cutoff
    frequency (in units of the sample frequency) given.
 *)

VAR
  sigma, b, sum   : REAL;
  i               : INTEGER;

BEGIN
  sigma:=0.132505/Frequency;
  IF sigma < 0.62 THEN   (* light filtering *)

    Coeffs[1] := sigma*sigma*0.5;
    Coeffs[0] := 1.0 - sigma*sigma;
    NumCoeffs:=1;

  ELSE  (* normal filtering *)

    NumCoeffs:= TRUNC(4.0 * sigma);
    IF NumCoeffs > MaxFilterCoeffs THEN
      WriteString ("FilterReal.SetGaussFilter: Too many coefficients:");
      WriteInt( NumCoeffs, 4 ); WriteLn;
      NumCoeffs:= MaxFilterCoeffs;
    END;
    b:= -1.0/(2.0*sigma*sigma);

      (* First make the sum for normalization *)
    sum:= 0.5;
    FOR i:=1 TO NumCoeffs DO
      sum:= sum + ETOX( b * FLOAT(i*i) );
    END;
    sum:= sum * 2.0;

      (* now compute the actual coefficients *)
    Coeffs[0]:= 1.0 / sum;
    FOR i:=1 TO NumCoeffs DO
      Coeffs[i]:= ETOX( FLOAT(i*i) * b ) / sum;
    END;
  END;
END SetGaussFilter;
```

Figure A4

be affected by this, the first and last n output values are reduced in magnitude by this truncation of the sum. This becomes an important issue when one wishes to filter a long digitized recording that does not fit into a single array of length N. The way to avoid the "edge effects" is to read overlapping segments of data into the input array and then to write out only the central $N - 2n$ points of the output array each time (with the exception of the first and last segments, where the initial and final "edges" should be written to preserve the total number of points).

```
PROCEDURE DoFilter( VAR Input, Output : ARRAY OF REAL;
                    NumInputPoints    : INTEGER;
                    Compression       : INTEGER);
  (* From the Input array, create a filtered Output that is
     decimated by Compression.  Thus the number of output points
     is equal to NumInputPoints DIV Compression.  SetGaussFilter
     must be called before this procedure to set up the filter coefficients.
  *)

VAR
  i0, i, j        : INTEGER;
  jmax, jmin      : INTEGER;
  sum             : REAL;

BEGIN
  FOR i0 := 0 TO (NumInputPoints DIV Compression) - 1 DO
    i := i0 * Compression;

     (* Make sure we stay within bounds of the Input array *)
    jmax := NumCoeffs;
    jmin := NumCoeffs;
    IF jmin > i THEN jmin := i END;
    IF jmax >= NumInputPoints - i THEN jmax := NumInputPoints - i - 1 END;

    sum := Coeffs[0] * Input[i];      (* Central point *)

    FOR j := 1 TO jmin DO                    (* Early points *)
      sum := sum + Coeffs[j] * Input[i-j];
    END;

    FOR j := 1 TO jmax DO                    (* Late points *)
      sum := sum + Coeffs[j] * Input[i+j];
    END;

     (* Assign the output value *)
    Output[i0] := sum;

  END;  (* FOR i0 *)
END DoFilter;

END FilterReal.
```

Figure A4. *Continued.*

A3.2. Cubic Spline Interpolation

A very useful interpolation technique for single-channel recording is the cubic spline, in which a cubic polynomial spans the interval between each pair of data points. A different polynomial is used for each interval, with coefficients chosen to match the function values as well as the first and second derivatives at the sample points. An introduction to the theory can be found in Hamming (1975). Briefly, we wish to find an interpolating function f whose values $f(1), f(2) \ldots$ match the data values $y_1, y_2, \ldots$ obtained at equally spaced sample times. Intermediate values $f(k + \theta)$ for θ between 0 and 1 are given by

$$f(k + \theta) = y_k\rho + y_{k+1}\theta + a_k(\rho^3 - \rho) + a_{k+1}(\theta^3 - \theta) \tag{A14}$$

where $\rho = 1 - \theta$. Before the interpolation is done, the coefficients a_k must be computed. They are specified by the system of equations

$$a_{k-1} + 4a_k + a_{k+1} = y_{k-1} - 2y_k + y_{k+1} \tag{A15}$$

which can be solved by Gaussian elimination.

```
      SUBROUTINE SPLINE (IN, OUT, A, N, NOUT)
C
C     This subroutine accepts N integer values from array
C     IN, interpolates them by the factor NE = NOUT/N
C     and returns the NOUT-NE+1 output points in the
C     array OUT.  The array A is used internally for
C     coefficients of the cubic term of the interpolating
C     polynomial.
C
      INTEGER IN(N), OUT(NOUT)
      REAL A(N)
C
      B = -1.0 / (2.0 + SQRT( 3.0 ))
      NE = NOUT / N
      NE1 = NE - 1
      E = NE
C
C        Form the coefficient array
C
      A(1) = 0.0
      A(N) = 0.0
      DO 10, I=2, N-1
        TEMP = 2 * IN(I) - IN(I-1) - IN(I+1)
        A(I) = B * (TEMP + A(I-1))
 10   CONTINUE
C
      DO 20, I=1, N-1
        J = N-I
        A(J) = A(J) + B * A(J+1)
 20   CONTINUE
C
C        Insert the original points into OUT
C
      DO 30, I=1, N
        K = NE*I - NE1
        OUT(K) = IN(I)
 30   CONTINUE
C
C        Handle the intermediate points
C
      DO 40, J=1, NE1
        P = J/E
        Q = 1.0 - P
        P3 = P * (P * P - 1.0)
        Q3 = Q * (Q * Q - 1.0)
        DO 40, I=1, N-1
          I1 = I+1
          K = NE * I + J - NE1
          OUT(K) = Q*IN(I) + P*IN(I1) + Q3*A(I) + P3*A(I1)
 40   CONTINUE
      RETURN
      END
```

Figure A5

```
IMPLEMENTATION MODULE SplineReal;

PROCEDURE Spline (VAR In, Work, Out: ARRAY OF REAL;
                      InNumber     : INTEGER;
                      Expansion    : INTEGER);

  (* From the InNumber input points, make (InNumber-1) * Expansion - 1
     output points, using cubic spline interpolation.  The output
     points Out[ Expansion * i] are equal to the corresponding input
     points In[i].
     The Work array must have at least InNumber elements.
  *)
  CONST
    c = 0.2674919;  (* equals 1 / ( 2 + sqr(3) ) *)

  VAR
    p,  q,
    p3, q3         : REAL;
    i,j,k,ini      : INTEGER;

  BEGIN
     (* Compute coefficients: forward calculation *)
    Work[0] := 0.0;
    FOR i := 1 TO InNumber-2 DO
      Work[i]:= c * ( In[i-1] - 2.0 * In[i] + In[i+1] - Work[i-1] );
    END;

     (* Back-substitution *)
    Work[InNumber-1] := -c * Work[InNumber-2];
    FOR i := InNumber-1 TO 1 BY -1 DO
      Work[i-1] := Work[i-1] - c * Work[i];
    END;

     (* Copy the original points *)
    j := 0;                  (* j is the output pointer *)
    FOR i:=0 TO InNumber-1 DO
      Out[j]:=In[i];
      INC(j, Expansion);  (* increment j by Expansion *)
    END;

     (* Compute the interpolated points *)
    FOR k:=1 TO Expansion-1 DO
      p:= FLOAT(k) / FLOAT(Expansion);
      q:= 1.0 - p;
      p3 := p * ( p * p - 1.0);
      q3 := q * ( q * q - 1.0);
      j  := k;
      ini:= 0;
      FOR i:=0 TO InNumber-2 DO
        Out[j]:= q * In[ini] + p * In[ini+1]
                 + q3 * Work[i] + p3 * Work[i+1];
        INC(ini);
        INC(j, Expansion);
      END;
    END;
  END Spline;

END SplineReal.
```

Figure A6

```
      subroutine AMOCALL(npar,nvert,simp,theta,stepfac,functol,funk,
     &     fmin,niter)
c Subroutine to simplify call of AMOEBA.FOR from Press et al. (1992)
c This subroutine uses the input values (see below) to:
c (1) set up the starting simplex in simp(21,20)
c (2) set the corresponding function values in fval(21)
c
c SIMP should be declared in calling program, e.g. as simp(21,20).
c      (SIMP is defined here, but because of problems in passing values
c      in 2-dimensional arrays with variable dimensions, it is simpler
c      to declare SIMP in the calling program)
c
c INPUT:
c     npar = number of parameters
c     nvert = npar+1
c     theta(npar) = initial guesses for parameters
c     stepfac = value to control initial step size, e.g. stepfac=0.1
c        starts with step size=0.1*initial guess.
c     functol = tolerance for convergence (should be set to machine
c        precision, or a bit larger -see Press et al.)
c     funk = name of subroutine that calculates the value to be
c        minimized
c
c OUTPUT:
c     theta = final values of parameters (in this version, set to the
c        parameters corresponding to the best vertex of final simplex).
c     fmin = corresponding minimum value for funk(theta)
c     niter = number of function evaluations done
c
      real simp(nvert,npar),fval(21),theta(npar),step(20)
      EXTERNAL funk
c
c     nvert=npar+1              ! # of vertices in simplex
```

Figure A7

The FORTRAN subroutine SPLINE (Fig. A5) accepts an integer array of N data values and fills a second integer array with the original points and interpolated values. The subroutine first computes the coefficients in an efficient manner that is equivalent to Gaussian elimination and backsubstitution. The N coefficients are kept in a real array A for further use if desired. The subroutine forces the second derivative of f to be zero at the first and last data points. This means that if a long record is to be interpolated in shorter segments, the segments

```
      do j=1,npar
        step(j)=stepfac*theta(j)
      enddo
      do j=1,npar
        simp(1,j)=theta(j)              !!start values=vertex #1
      enddo
      fval(1)=funk(theta)     !function value for these
      fac=(sqrt(float(nvert))-1.)/(float(npar)*sqrt(2.))
      do i=2,nvert
        do j=1,npar
          simp(i,j)=simp (1,j) + step(j)*fac
        enddo
        simp(i,i-1)=simp(1,i-1) + step(i-1)*(fac+1./sqrt(2.))
        do j=1,npar
          theta(j)=simp(i,j)       !copy paramters into theta (for funk)
        enddo
        fval(i)=funk(theta)           !function value for these
      enddo
c
      call AMOEBA(simp,fval,nvert,npar,npar,functol,funk,niter)
c
c     Return the best vertex
      fmin=fval(1)
      do i=2,nvert
        if(fval(i).lt.fmin) then
          fmin=fval(i)
          do j=1,npar
            theta(j)=simp(i,j)
          enddo
        endif
      enddo
c
      RETURN
      end
```

Figure A7 *Continued.*

should have some overlap (e.g., ten data points) to allow smooth "splicing" of the interpolated segments.

A MODULA-2 implementation of the same algorithm is given in Fig. A6. Here the coefficients a_k are stored in the Work array while the input and output data are assumed to be in arrays of real values.

A3.3. Error Function Evaluation

For computations involving the step response of Gaussian filters, a numerical approximation for the error function is required. One of the simplest approximations for the complementary error function is

$$\text{erfc}(x) = (a_1 t + a_2 t^2 + a_3 t^3)\exp(-x^2) \tag{A16}$$

where $t = 1/(1 + px)$; $p = 0.47047$; $a_1 = 0.3480242$; $a_2 = -0.0958798$; $a_3 = 0.7478556$; and where x is restricted to positive values. The error in this approximation is less than 2.5×10^{-5}.

The error function itself can be evaluated as

$$\text{erf}(x) = 1 - \text{erfc}(x)$$

and for negative values of x,

$$\text{erf}(x) = -\text{erf}(-x)$$

The formula in equation A16 is from Hastings (1955), which also contains more exact formulas. These formulas can also be found in Abramovitz and Stegun (1964), p. 299.

A3.4. A Calling Routine for AMOEBA

The subroutine (in FORTRAN) designed to simplify calling of the simplex minimization routine by Press *et al.* (1992) is given in Fig. A7. This was discussed in Section 6.3.

References

Abramovitz, M., and Stegun, I. A., 1965, *Handbook of Mathematical Functions,* Dover Publications, New York.

Aldrich, R. W., Corey, D. P., and Stevens, C. F., 1983, A reinterpretation of mammalian sodium channel gating based on single channel recording, *Nature* **306:**436–441.

Beale, E. M. L., 1960, Confidence regions in non-linear estimation, *J. R. Statist. Soc.* **B22:**41–76.

Bliss, C. I., and James, A. T., 1966, Fitting the rectangular hyperbola, *Biometrics* **22:**573–602.

Box, G. E. P., and Coutie, G. A., 1956, Application of digital computers in the exploration of functional relationships, *Proc. IEEE* **103**(Part B, Suppl. 1):100–107.

Clapham, D. E., and Neher, E., 1984, Substance P reduces acetylcholine-induced currents in isolated bovine chromaffin cells, *J. Physiol.* **347:**255–277.

Colquhoun, D., 1971, *Lectures on Biostatistics,* Clarendon Press, Oxford.

Colquhoun, D., 1979, Critical analysis of numerical biological data, in: *Proceedings of the Sixth International CODATA Conference* (B. Dreyfus, ed.), pp. 113–120, Pergamon Press, Oxford.

Colquhoun, D., and Hawkes, A. G., 1982, On the stochastic properties of bursts of single ion channel openings and of clusters of bursts. *Phil. Trans. R. Soc. Lond.* [*Biol.*] **300:**1–59.

Colquhoun, D., and Hawkes, A. G., 1987, A note on correlations in single ion channel records, *Proc. R. Soc. Lond.* [*Biol.*] **230:**15–52.

Colquhoun, D., and Ogden, D. C., 1988, Activation of ion channels in the frog endplate by high concentrations of acetylcholine, *J. Physiol.* **395:**131–159.