Journal of Analytical Chemistry, Vol. 60, No. 7, 2005, pp. 614–618. Translated from Zhurnal Analiticheskoi Khimii, Vol. 60, No. 7, 2005, pp. 692–697.

ARTICLES

Automatic Baseline Construction Using Vertical Histograms

A. S. Korotkov

Vernadsky Institute of Geochemistry and Analytical Chemistry, Russian Academy of Sciences, ul. Kosygina 19, Moscow, 119991 Russia

Received April 30, 2004; in final form, October 4, 2004

Abstract—An algorithm is proposed for automatic baseline construction in the absence of *a priori* information about the shape of the experimental curve and the arrangement of basic portions in the experimental curve. The functioning of the algorithm is considered for the case when the basic portion of the curve is distorted with a parasitic signal of an intricate shape.

In solving many problems, the processing of the experimental data is reduced to the determination of peak areas above a baseline. In the manual selection of the baseline, as well as in the presence of *a priori* information about the portions of the experimental curve that belong to the background with certainty, the construction of the baseline presents no difficulties. However, in the automatic processing of the experimental data and in the absence of *a priori* information about the basic portions, the inaccurate determination of the baseline may introduce significant errors into the result of processing. The problem is usually complicated by the experimental curve in its background portion being distorted by both random noise and parasitic signals of intricate shape (an example is provided by thermal stabilization systems generating parasitic periodic signals).

EXPERIMENTAL

Let me first consider the case when the baseline is given by the function $b(t)$ = const, that is, by a horizontal line. Let me construct a histogram for point distribution in the experimental curve $f(t)$ (Fig. 1a) along the axis of ordinates (vertical histogram). It will evidently take the shape shown in Fig. 1b. It is evident that the baseline must pass through the principal maximum in the histogram.

A question arising while constructing a vertical histogram concerns the size of the intervals into which the ordinate should be divided. If the intervals are too large, points that actually do not belong to the baseline will be attributed to the histogram maximum, and the baseline will shift relative to its true position. If the intervals are too small, the number of points falling in a certain interval will be small and, therefore, statistically unreliable.

Let me consider the case of small intervals using a particular example. Let the background portion of an experimental curve include 300 points, the distorting signal be a normally distributed noise with a standard deviation (SD) of σ, and the axis of ordinates be divided into intervals equal to $\sigma/3$. Then, the shape of the histogram must be similar to that shown in Fig. 2a. However, the number of points in one or another interval is a random value; for our estimates, we may suppose that this

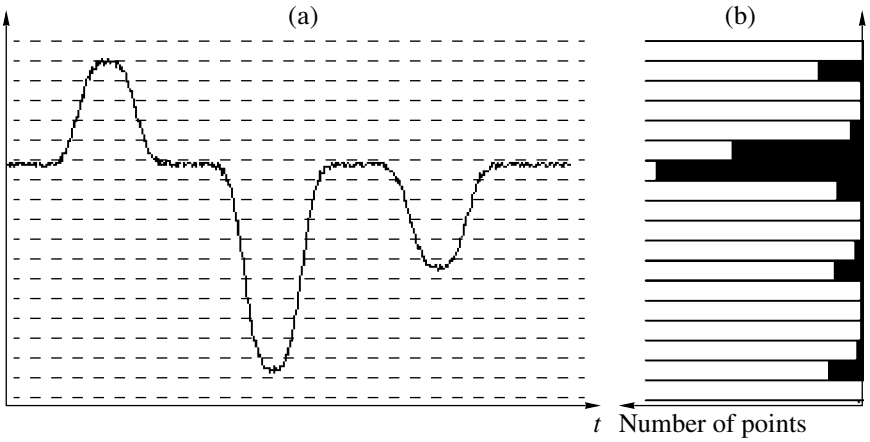

Fig. 1. (a) Experimental curve and (b) its vertical histogram.

1061-9348/05/6007-0614

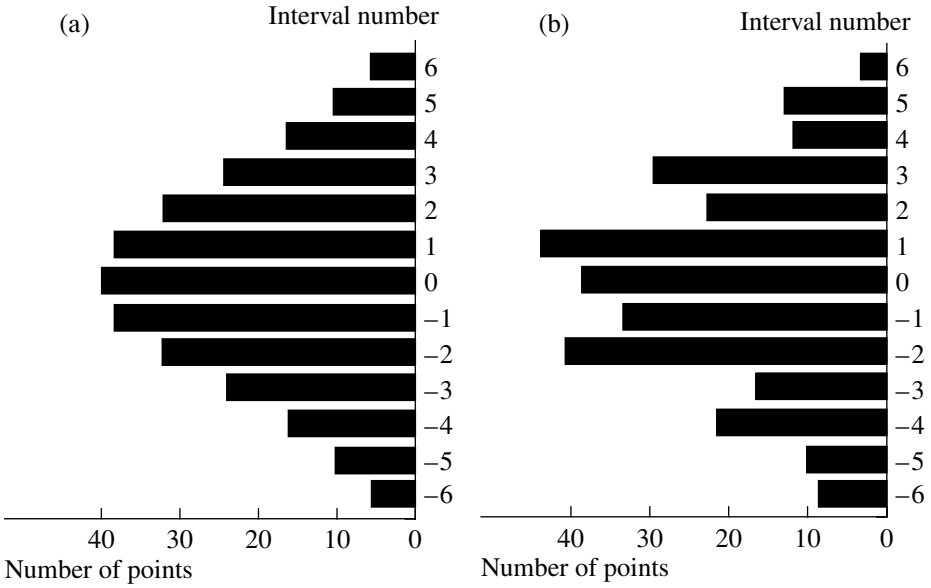

Fig. 2. (a) Mathematical expectation of the vertical histogram and (b) real histogram.

value obeys the Poisson distribution. Then, the number of points in the zero interval will vary from 27 to 51; in intervals 1 and –1, from 25 to 49; in intervals 2 and –2, from 21 to 43; and so on. As a result, a real histogram will be similar to that shown in Fig. 2b, and its maximum will be determined incorrectly with a high probability.

The solution is to divide the axis of ordinates into a relatively great number of small intervals and find a union of these intervals that includes the majority of baseline points without covering the informative portions of the experimental curve where possible. We will search for an optimum among the unions of neighboring intervals $g_i, g_{i+1}, \ldots, g_{i+n-1}$ for all possible i and n, where i is the beginning of the union interval, n is the number of histogram intervals in the union interval, and g_j is the number of points in the jth interval of the histogram. The optimality criterion for the union interval will be sought in the form of the function $F(i, n) = f(n\Delta)\sum_{j=1}^{n} g_{i+j-1}$, where Δ is the length of a unit interval in the histogram and $n\Delta$ is the length of the union interval. Function $f(n\Delta)$, which determines the character of the optimality criterion will be found from the condition that the distorting signal is normally distributed with a mean value $\bar{x}$ and SD = σ.

Changing from the histogram to the expectation function, we obtain $\sum_{j=1}^{n} g_{i+j-1} = M\int_{a}^{b} e^{-\frac{(x-\bar{x})^2}{2\sigma^2}}\,dx$, where M is a normalizing factor that does not affect the position of the maximum and a and b are the beginning and end of the union interval, respectively ($b - a = n\Delta$). Let us take $(b-a)^{-\gamma}$ as an analogue of the function $f(n\Delta)$. One can easily find that the function $F_\gamma(a, b) = (b-a)^{-\gamma}\int_{a}^{b} e^{-\frac{(x-\bar{x})^2}{2\sigma^2}}\,dx$ at $\gamma(k) = k\frac{e^{-\frac{1}{2}k^2}}{\int_{0}^{k} e^{-\frac{1}{2}x^2}\,dx}$ has only one maximum at $b = \bar{x} + k\sigma$, $a = \bar{x} - k\sigma$. Setting $k = 1.644$ (that is, the value at which $\frac{1}{\sqrt{2\pi}}\int_{-k}^{k} e^{-\frac{1}{2}x^2}\,dx = 0.9$), we can find that 90% of the baseline points will fall within the union interval (a, b) within which $F_{\gamma(1.644)}(a, b) = 0.377392$ attains a maximum. Note than $\gamma(1.644) = 0.377392$; therefore, the parameter $\gamma(k) = 0.5$ corresponds to $k = 1.4$ and 84% of the baseline points falling within the optimal interval.

Let me go back to histograms. Let me consider the optimum union interval as the interval within which the function $F(i, n) = (n\Delta)^{-0.377392}\sum_{j=1}^{n} g_{i+j-1}$ attains a maximum. This interval will include about 90% of the baseline points, which is quite sufficient for determining the baseline level and the variance of the distorting signal. It is clear that if the variance is calculated without taking into account 10% of the most strongly deviating points, it will be underestimated and, strictly speaking, one should introduce a correction.

Strictly speaking, the function $F_\gamma(i, n)$ depends on the interval length Δ, that is, on the partition of the Δ axis in constructing the histogram. The comprehen-

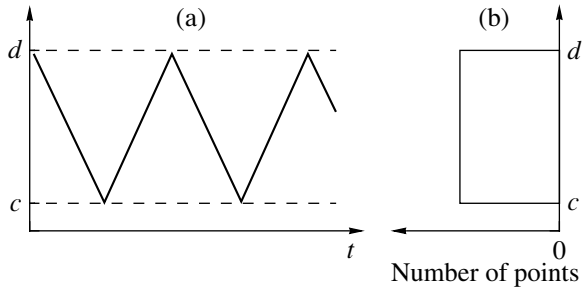

Fig. 3. (a) A saw-tooth distorting signal and (b) its vertical probability density function.

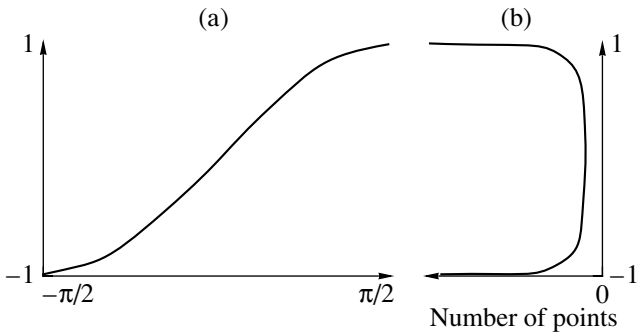

Fig. 4. (a) A portion of sinusoid and (b) its vertical probability density function.

sive analysis of this dependence is beyond the scope of this paper. However, let me point to the following fact. Let us take two histograms $g_1, \ldots, g_N$ with a union interval of the length Δ and $\bar{g}_1, \ldots, \bar{g}_{2N}$ with a union interval of the length $\bar{\Delta} = \frac{1}{2}\Delta$. Then, taking into account that $g_j = \bar{g}_{2j-1} + \bar{g}_{2j}$, we obtain $\bar{F}_\gamma(2i, 2n) = (2n\bar{\Delta})^{-\gamma}\sum_{j=1}^{2n}\bar{g}_{j+2i-1} = \left(2n\cdot\frac{1}{2}\Delta\right)^{-\gamma}\sum_{j=1}^{n}\bar{g}_{2j+2i-2} + \bar{g}_{2j+2i-1} = (n\Delta)^{-\gamma}\sum_{j=1}^{n} g_{j+i-1} = F_\gamma(i, n)$. In other words, the table of possible values of $\bar{F}_\gamma(i, n)$ will consist largely of the same values as the table $F_\gamma(i, n)$. Therefore, in not very bad cases, we may suppose that, at a relatively small Δ, the optimal union intervals for these histograms will be close to each other. In this case, the reduction of the unit interval of the histogram will result in a certain refinement of the set of baseline points at an extension of the time of calculations. One of the criteria for selecting Δ (in addition to the number of digits of the analog-to-digital converter, the required precision, and so on) may also be the number of unit intervals in the optimum union interval: if this number is greater than ten, the further reduction of Δ will not improve the accuracy significantly.

RESULTS AND DISCUSSION

Let me consider the functioning of the algorithm in the case when the baseline is distorted with a saw-tooth signal with vertical boundaries c and d (see Fig. 3a). It is clear that the mathematical expectation of the sum $\sum_{j=1}^{n} g_{i+j-1}$ this case will be $M\int_a^b p(c, d, x)dx$, where M is a normalizing factor; a and b are the beginning and end of the union interval segment, respectively ($b - a = n\Delta$); and $p(c, d, x)$ is vertical probability density function, which is equal to $1/(d - c)$ inside the interval $[c, d]$ and zero outside this interval (Fig. 3b). It can easily be

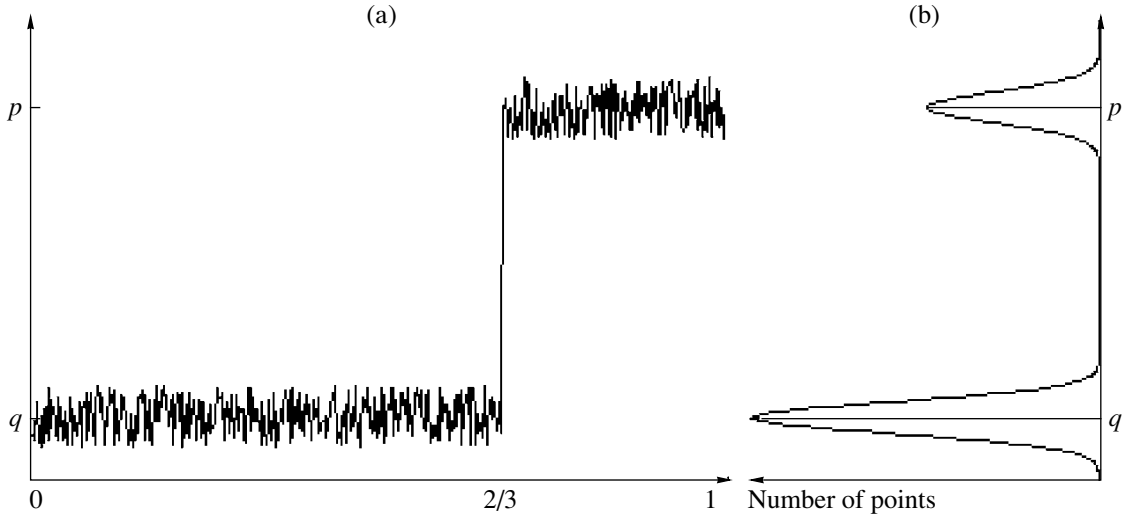

Fig. 5. (a) A stair-like curve and (b) its vertical probability density function.

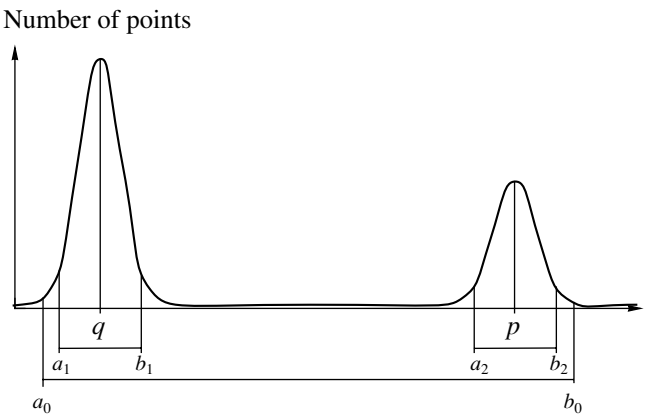

Fig. 6. Local minima of the optimization function of the vertical histogram for the experimental stairlike curve.

found that the function $F_\gamma(a, b) = (b-a)^{-\gamma}\int_a^b p(c, d, x)dx$ at $0 < \gamma < 1$ attains a maximum at $a = c$ and $b = d$. The value $\gamma = 0.377392$ found for the normal distribution of the distorting signal also gives a good result for the saw-tooth distorting signal.

Consider the functioning of the algorithm for a sinusoidal distorting signal. The shape of the histogram in this case can be assessed using only one sinusoid half-period (Fig. 4a). In this case, the vertical probability density function will be as follows: $\Phi(y) = P(\sin(x) < y) = P(x < \arcsin(y)) = \arcsin(y)/\pi + \frac{1}{2}$. Correspondingly, the vertical probability density function (and histogram estimate) will be equal to $p(y) = \frac{1}{\pi\sqrt{1-y^2}}$ (Fig. 4b). Function $(b-a)^{-\gamma}\int_a^b \frac{1}{\sqrt{1-y^2}}dy$ passes through the maximum $a = -1$, $b = 1$ at $\gamma < \frac{1}{2}$ and approaches $+\infty$ at $a = -1$, $b \longrightarrow -1$ and $b = 1$, $a \longrightarrow 1$ when $\gamma > \frac{1}{2}$. Thus, the optimum sum of histogram intervals will cover all baseline points at $\gamma < \frac{1}{2}$ and be restricted to the upper or bottom boundary of the baseline at $\gamma > \frac{1}{2}$. The value $\gamma = 0.377392$ found for the normal distribution of the distorting signal will give a good result for the sinusoidal distorting signal.

Consider the functioning of the algorithm in the case when the experimental curve is a step distorted by a normally distributed noise with SD = σ (Fig. 5a). The vertical probability density function in this case will be a sum of two Gaussian functions (Fig. 5b). Let me suppose for definiteness that the background portion is two-thirds of the whole experimental curves. Then, the "smaller" Gaussian of the vertical distribution will be one-half of the "larger" Gaussian. In this case, function $F_\gamma(a, b)$ will have three local maxima (Fig. 6). The global maximum may be either the maximum (a_1, b_1) including only points of the larger Gaussian, or the maximum (a_0, b_0) including all points of the experimental curve. In the former case, the algorithm will find the foot of the step; in the latter case, the whole experimental curve will be considered the baseline. The selection between the two cases depends on the parameter γ and the ratio between the step height and noise SD, $\frac{p-q}{\sigma}$. At

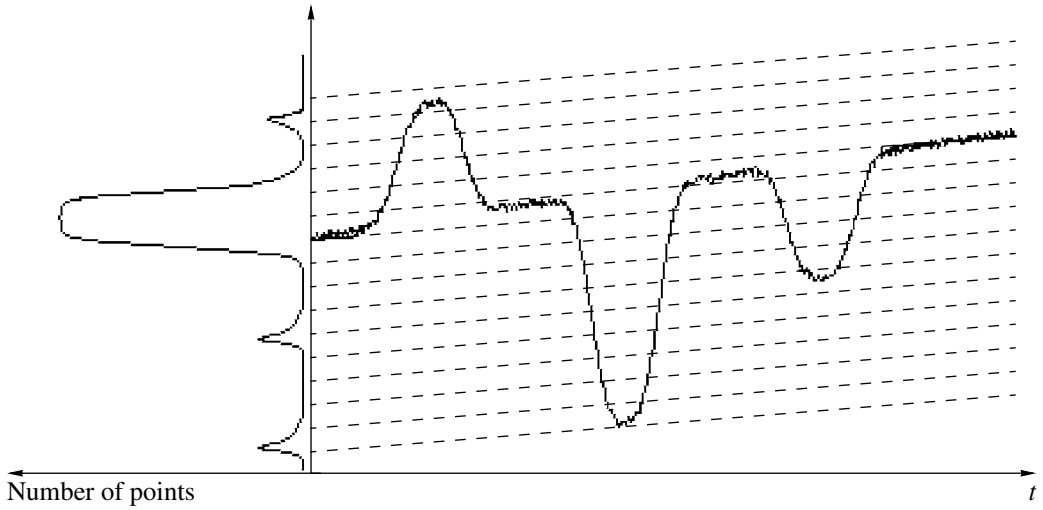

Fig. 7. An experimental curve with an incline baseline and its vertical histogram. The global maximum in the histogram is diffuse because of a difference between the slope of the baseline and the projection angle.

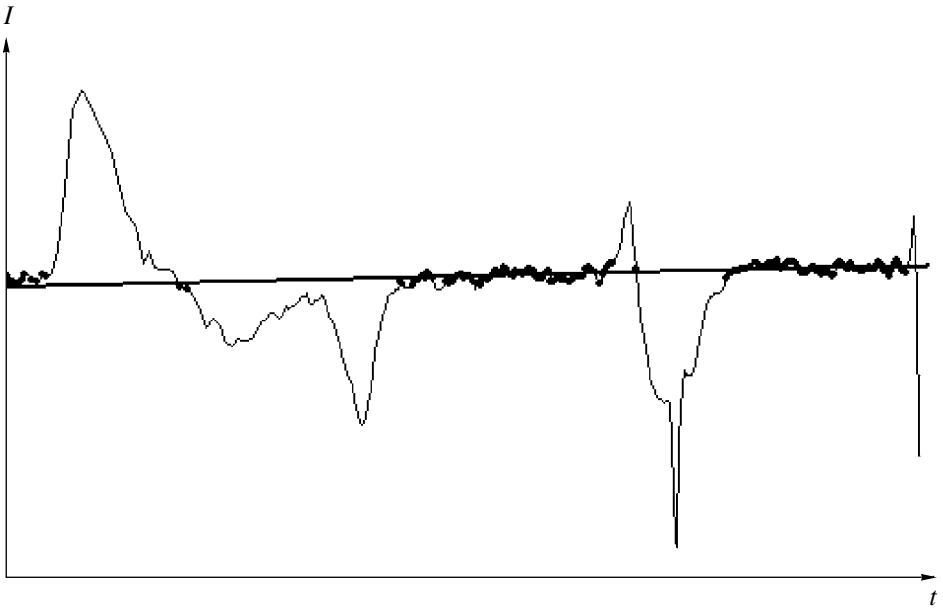

Fig. 8. An experimental curve recorded with a solid-electrolyte analyzer and a baseline constructed using the proposed algorithm.

$\gamma = 0.377392$, the foot of the step will be selected at $p - q > 8.05\sigma$; at $\gamma = 0.5$ the sufficient criterion is $p - q > 4.9\sigma$.

Let me consider the case when the baseline is an inclined line given by the equation $b(t) = At + B$. In this case, setting A equal to a constant, we can project points of the experimental curve onto the axis of ordinates at an angle of $\arctan(A)$ rather than at the right angle. Then, for each A we will obtain its specific baseline, and the question will be what baseline should be considered optimal. If the true slope of the baseline differs from A, the main maximum in the histogram constructed at an angle of $\arctan(A)$ will be diffuse (Fig. 7). Using this property, we can take the ratio between the number of points attributed to the baseline and its variance as the objective function $h(A)$. The problem can be, therefore, reduced to the determination of the maximum of $h(A)$. Note that the function $h(A)$ will not be continuous and will, most probably, have several local maxima, so that the use of fast optimization methods (such as the Newton method) will be excluded and optimization will be difficult. One should most likely use the enumeration of all reasonable possibilities. Therefore, attempts at constructing a baseline of a more intricate shape using this method will be time-consuming. An example of constructing an inclined baseline for an experimental curve recorded with a solid-electrolyte analyzer [1] is shown in Fig. 8.

Note in conclusion that the algorithm is stable to distorting signals of different shapes. This makes it useful for the preliminary processing of experimental data (for example, differentiation, smoothing, Fourier transformations, and so on).

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research, project no. 03-03-32876.

REFERENCES

1. Zuev, B.K., Korotkov, A.A., Filonenko, V.G., Mashkovtsev, A.N., and Zvolinskii, V.P., *Zh. Anal. Khim.*, 2004, vol. 59, no. 2, p. 185 [*J. Anal. Chem.* (Engl. Transl.), vol. 59, no. 2, p. 163].
2. Vapnik, V.N., *Vosstanovlenie zavisimostei po empiricheskim dannym* (Restoration of Relationships from Empirical Data), Moscow: Nauka, 1979.
3. Malinovskii, L.G., *Analiz statisticheskikh svyazei: model'no-konstruktivnyi podkhod* (Analysis of Statistical Relations: A Model-Construction Approach), Moscow: Nauka, 2002.
4. Anderson, T.W., *An Introduction to Multivariate Statistical Analysis*, New York: Chapman, 1957. Translated under the title *Vvedenie v mnogomernyi statisticheskii analiz*, Moscow: Fizmat, 1963.