Simulated Annealing for Mixture Distribution Analysis and its Applications to Reliability Testing

Cher Ming Tan and Nagarajan Raghavan
Nanyang Technological University, Singapore
National University of Singapore

1. Introduction

1.1 Objective

Reliability is a very important field of study in today's era of technology. It is essential to quantitatively estimate the reliability of a product or device before it is mass produced and sold in the market through accelerated life tests. In reliability testing and data analysis, global optimization of the log-likelihood function plays a key role. An effective technique for this optimization is *Simulated Annealing (SA)*.

The objective of this chapter is to illustrate the applicability of *SA* to reliability data analysis. In particular, this optimization technique is very useful for *mixture distribution analysis* which will be described in detail later. The flow of the chapter goes as follows. A brief introduction to reliability statistics will be provided, intended to provide a basic outlook into this fascinating field to readers who are new to it. The role of SA in reliability statistics will be made clear through the developed log-likelihood function which needs to be optimized. This is followed by an insight into the need for mixture distribution analysis in reliability testing and assessment. The origin and methodology underlying the SA algorithm is then described in detail. The application of SA to mixture distribution analysis is presented and two practical examples of this application are provided from the microelectronics industry where electronic device reliability for gate oxide breakdown and electromigration phenomenon is assessed. Towards the end, techniques proposed in the literature to improve the efficiency of search for SA is presented and a concluding section directs the reader on the path to pursue further research investigations in simulated annealing.

1.2 Scope

The most fundamental form of the SA algorithm is employed in the reliability analysis presented in this chapter. Although more efficient designs of the SA algorithm have been made, they are not utilized in this work. The application case studies illustrate the application of SA for reliability analysis only in the field of microelectronics. The approach presented in this paper is nevertheless applicable to all practical reliability studies.

2. Introduction to reliability testing

2.1 Need for reliability

In today's competitive world and the age of globalization, it has become very essential for the manufacturing industry to keep up to the rapidly rising demands and expectations of the customers. To stay upbeat in the industry and capture large market shares, companies have been marketing aggressively through various strategies one of which is the value-add to their products. Cost competition will always be a lose-lose strategy. Reliability of a product is a good value-add to a product, and this is even more so for high-end electronic products that have revolutionized the world that we live in today. Good reliability brings good reputation and near zero field return during warranty period of the product.

Prior to implementing techniques to improve reliability of a manufactured product, it is essential to characterize and quantify reliability in a statistically credible manner so that improvement efforts can be evaluated. Reliability in itself is a large and wide field of research that encompasses statistical distributions to model failure characteristics and physics of failure to understand the nature of a failure mechanism and its associated failure mode.

The statistics of reliability modeling has been well investigated in the past few decades and therefore the statistical techniques for reliability quantification are established. However, the successful application of these models to practical usage has not been very fruitful. The problem lies in the inappropriate usage of these theories by practicing reliability engineers in the field as the assumptions behind some of the theories are not well understood by engineers in the industry and the lack of familiarity with the methodology to account for the presence of multiple failure mechanisms in a given reliability test data.

2.2 Accelerated life testing

As product reliability is being enhanced, the time taken to obtain the product failure time gets excessively longer, and a common practice therefore is to evaluate product reliability using **accelerated life testing (ALT)**. It is a technique whereby a product is stressed to failure at a much higher stress condition than the normal field operating condition experienced by it. The high stress condition serves to accelerate the failure mechanism so that failures can be observed sooner and adequate time-to-failure (TTF) data for the product can be collected for reliability analysis. Reliability analysis is most useful when it is obtained at the earliest stages of product development so as to facilitate improvements targeted at prolonging the lifetime of a product before it is mass produced and marketed to the customers. ALT is therefore gaining more relevance today as time-to-market gets shorter and shorter.

It is crucial that the high stress in ALT should only accelerate the failure mechanism observed in the field use condition and that it should not give rise to new unseen failure mechanisms which are not typically found at use stress levels. Otherwise, estimation and extrapolation of the product lifetime to the field conditions will not reflect field failures appropriately and this will defeat the very purpose of performing an ALT. On the other hand, there are many potential pitfalls to ALT as outlined by Meeker *et al.* (1998) and Suhir (2002), one of which is the occurrence of multiple failure mechanisms due to the high stress applied during the ALT.

To uncover the different failure modes and mechanisms underlying a product failure, it is necessary to perform a failure analysis (FA) on the failed products. Such an analysis requires

precise sample preparation and accurate examination using various FA instrumentation tools and these observations do provide useful information on the physical nature of the localized failure site and the necessary corrective action to be taken to prolong or avoid these failures in the field. However, performing failure analysis for all the reliability test failures is practically impossible, and a way to classify and categorize failed products into different failure mechanisms based on statistical means is necessary so that only a few representative failed products in each category will need to be analyzed. In this chapter, we will present this method using the Simulated Annealing.

2.3 Statistical analysis of failure data

Having discussed the relevance of reliability and ALT in quality and reliability improvement strategies, we shall now get acquainted with the fundamental probabilistic and statistical definitions underlying reliability. We shall then show where *Simulated Annealing (SA)* comes into the picture of reliability data analysis. Please note that the terms "*device*", "*system*" and "*product*" all refer to the same entity and they may be used interchangeably in this work.

2.3.1 Reliability Fundamentals

Reliability is defined as the probability that a product will perform its intended function over a time period *t* given that it is being used under stated operating conditions (Ebeling, 2005). It is a continuous function of time *t* and is denoted by *R(t)*. Mathematically, reliability can be expressed as in (1) where T is a continuous random variable representing the time to failure of the product with $T \geq 0$. For a given value of *t*, *R(t)* is the probability that the time to failure is greater than or equal to *t*. Note that $\forall t \geq 0$, $0 \leq R(t) \leq 1$; $R(0) = 1$ and $\lim_{t \to \infty} R(t) = 0$.

$$R(t) = \Pr\{T \geq t\} \tag{1}$$

The complementary function of R(t) is the failure probability or *cumulative density function* (CDF) which is denoted by F(t) and is written as:

$$F(t) = 1 - R(t) = \Pr\{T < t\} \tag{2}$$

The time derivative of the CDF gives the probability density function (PDF) of the distribution, denoted by f(t). The failure rate, $\lambda(t)$, represents the instantaneous rate of failures in a sample. It is an alternative way of describing a failure distribution. It is expressed as in (3). When $\lambda(t)$ is an increasing, decreasing or constant function of time, they are characterized by increasing failure rate (IFR), decreasing failure rate (DFR) or constant failure rate (CFR) respectively (Ebeling, 2005). A typical characteristic of the failure rate of any product is represented by the *bathtub curve* in Fig 1 (US Army, 2005) which shows the life pattern of the failure rate of the product from the instant of manufacturing and initial field use up to the later stages of the aging (wear-out) phenomenon. DFR relates to the infant mortality initial period of a system's operation; CFR is the period of useful system operation with the lowest failure rate and IFR accounts for the aging mechanism in the product due to gradual wear-out which is characteristic of its intrinsic failure. It is always desirable to ship out a product after its DFR regime so as to minimize field returns (this is

done using a screening mechanism known as ***burn-in***) and prolong the CFR regime as long as possible such that the product's useful desired lifetime does not include the wear-out IFR region where increasing number of failures are expected to be seen.

$$\lambda(t) = \frac{dF(t)/dt}{R(t)} = \frac{f(t)}{R(t)} \qquad (3)$$
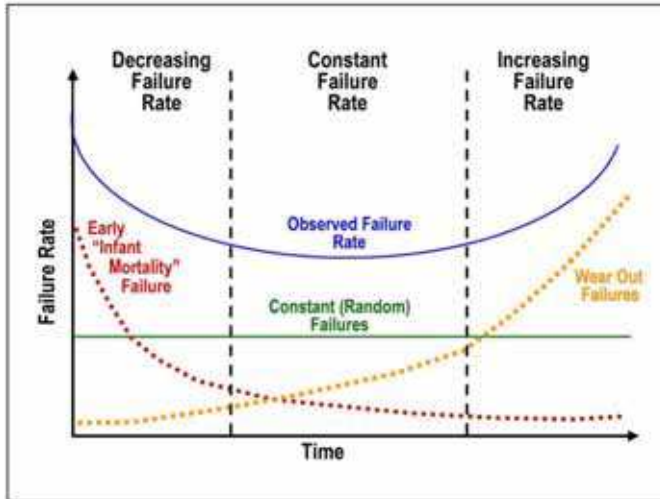

Fig. 1. Bathtub curve representing the failure rate regimes of a typical system (*US Army*, 2005).

The mean time to failure (***MTTF***) of a product is an important reliability metric and is given by (4). It represents the area under the reliability function spanning the whole range of the time continuum starting at t = 0.

$$MTTF = \int_{0}^{\infty} R(t)\,dt \qquad (4)$$

There are various statistical distributions that are used to model the statistics of failure time of products. Examples are the exponential, weibull, normal, lognormal and gamma distributions, to name a few. The most widely applicable distribution amongst these for microelectronic devices are the Weibull and Lognormal distributions. The parameters of these distributions can be tuned so as to fit any set of failure data be it decreasing, increasing or a constant failure rate in the bathtub curve.

These statistical distributions are associated with different degradation behaviors of a product. For example, Weibull distribution is typically used to characterize catastrophic failures while the Lognormal distribution is used to represent gradual rates of degradation (Tobias *et al.*, 1995). We shall be considering the Weibull and Lognormal distributions in our case studies later on in this chapter and it is therefore useful to know the form that their

reliability functions take as shown below in (5) and (6). In (5), β is the shape parameter and η is the scale parameter which is also called the **characteristic life** of the product. The ranges $\beta < 1$, $\beta = 1$ and $\beta > 1$ represent the early failure, constant failure and wear-out failure regions of the bathtub curve respectively. The lognormal reliability function in (6) is also characterized by two parameters viz. the shape parameter (σ) and the median time to failure (t_{50}). The function $\Phi(z)$ is the standardized normal CDF.

$$R_{weibull}(t) = e^{-\left(\frac{t}{\eta}\right)^{\beta}} \tag{5}$$

$$R_{\log normal}(t) = 1 - \Phi\left(\frac{1}{\sigma} \cdot \ln \frac{t}{t_{50}}\right) \tag{6}$$

2.3.2 Failure data analysis

As mentioned earlier, to expedite the process of reliability testing and obtain useful information on the system (device) reliability prior to mass production, it is necessary to perform ALT. Having performed the ALT, there is a standardized set of rules to follow in collecting, transforming, fitting and analyzing the obtained failure data.

Having collected the failure data (t_i), the empirical CDF values, $F(t_i)$ are calculated as in (7) where n is the total number of product sample under test (i.e. sample size of ALT test) and i is the failure order number when the Time to Failure (TTF) values (t_i) and their order numbers are listed in ascending order. This approach to compute the empirical $F(t_i)$ values is known as the **median rank method** (O'Connor, 2002). Note that the expression in (7) is valid only for the case where all the devices are tested until failure is observed. In some cases however, we might terminate an ALT prior to all the devices failing in which case the test is said to be **censored**. The test could be censored after a certain number of failures are observed (**failure terminated test**) or after a pre-determined fixed test time duration (**time-terminated test**). In the case of censored data, the index i is slightly modified to account for the effect of censoring. The details of these changes due to censoring are not critical here and readers interested in gaining more in-depth knowledge on this subject could refer to (O'Connor, 2002) for the complete details.

$$\hat{F}(t_i) = \frac{i - 0.3}{n + 0.4} \tag{7}$$

The empirical CDF values obtained are plotted on a graph paper that is unique to different statistical distributions and it is necessary to find the best fitting distribution parameters to fit this set of data accurately. One of the approaches to fit the data is to maximize the so called **Likelihood function** (**LKL**) which is given by the expression in (8) where t_f refers to the time-to-failure (TTF) data of the failed samples and t_c represents the censored data corresponding to those devices whose failure was not observed (Jiang *et al.*, 1992). Note that there could be various reasons for observing censor data such as unprecedented withdrawal of the test device during the test for other purposes, failure of the device due to some other

failure mechanism which is not of interest in the analysis etc… In (8), *n* is the total number of devices under test while *r* is the number of actual failures observed during the test (r ≤ n); the symbol ψ represents the set of distribution parameters which is {β, η} for a Weibull distribution and {t_{50}, σ} for a Lognormal distribution and *f(t)* and *R(t)* are the PDF and reliability functions respectively.

$$LKL(t;\psi) = \frac{n!}{(n-r)!} \cdot \prod f(t_f|\psi) \cdot \prod R(t_c|\psi) \tag{8}$$

Typically, instead of maximizing the likelihood (LKL) function, it is conventional to perform a monotonic transformation of *LKL* by taking its natural logarithm and maximizing ln(LKL), which is called the **Log-Likelihood function**, denoted here by *LLKL*. The expression for LLKL is given by (9) and the best fitting distribution parameter set, ψ, that fits the median rank CDF data in (7) is determined by optimizing the LLKL function.

$$LLKL(t;\psi) = \ln\left(\frac{n!}{(n-r)!}\right) + \sum \ln(f(t_f|\psi)) + \sum \ln(R(t_c|\psi)) \tag{9}$$

This *LLKL* function is the focus of our attention from now on since it is this function we will be optimizing to obtain the best fit to the ALT test data.

Fig 2 illustrates the median rank data plotted on a Lognormal distribution plot for a set of ALT data. The straight line fit shown is obtained by maximizing the *LLKL* function for the Lognormal distribution using the expression given in (9).
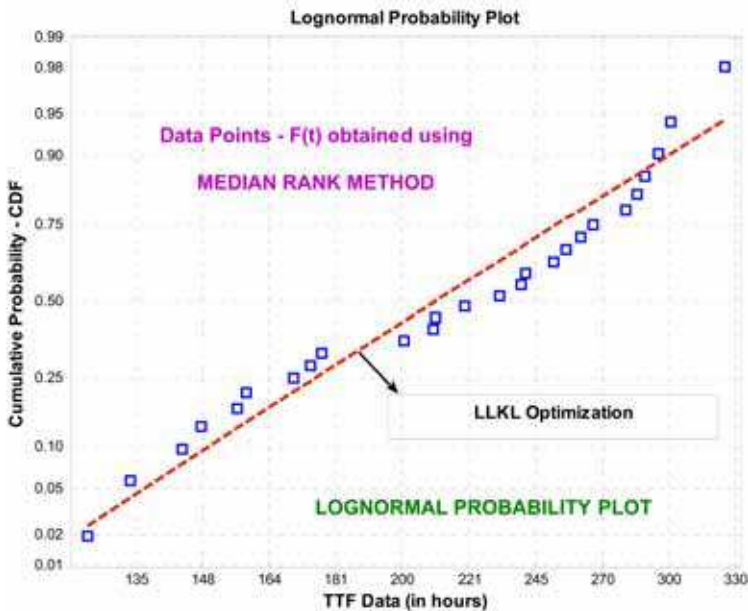

Fig. 2. Illustrating the Median Rank CDF data plotted on a Lognormal probability plot.

2.3.3 Optimization techniques for likelihood function

The *LLKL* function described in (9) is a multi-dimensional function of the statistical distribution parameters (ψ) that *f(t)* and *R(t)* are associated with depending on the distribution used. When the median rank CDF data needs to be fitted by suitable distribution parameters, various techniques could be used to do so. One of the approaches would be to plot the $F(t_i)$ points on the distribution probability plot graph paper and fit the points approximately by a straight line. The other approach would be to perform a least-square line of best fit regression analysis to determine the correlation coefficient and the slope and intercept of the fitting line which would in turn provide the values for ψ.

Other approaches include the Newton-Raphson method, Brent's method, Downhill Simplex method, Conjugate Gradient methods, Quasi-Newton methods etc… that are described in sufficient detail in (Press *et al.*, 2002). Some of these methods might only work under special cases where the gradients of the function to be maximized are defined at all points. In other words, they could be *gradient sensitive*. A few other methods above might work but might not be easy to code and implement. The most important fact regarding the above methods are that they are all *local optimization* algorithms and therefore, if we make use of these methods to find the global optimum, we would most likely end up getting a local optimum depending on the initial guess for the distribution parameters at the beginning of the algorithm execution. Since these methods are highly sensitive to the initial guess and are capable only of local optimization, it is necessary to look out for other techniques which are capable of finding the global optimum and are relatively insensitive to the user's initial guess.

One of the most useful, easy to implement and robust techniques for global optimization of the *LLKL* function is **Simulated Annealing** (SA) (Brooks *et al.*, 1995). It is useful for *LLKL* functions here because SA is capable of efficiently finding the *global optimum* of any *n-dimensional* function. As mentioned earlier, since *LLKL* functions are typically multi-dimensional, SA would be an easy approach to use to optimize them.

In optimization literatures, the function to be optimized is usually referred to as the **objective function** (Press *et al.*, 2002). We will follow the same convention here and call our *LLKL* function as the *objective function*. The equation to be solved for obtaining the best fit distribution parameters to the ALT test data is now expressed as in (10) where φ₁, φ₂, φ₃, …, φ$_k$ etc… are the statistical distribution parameter elements belonging to the set ψ. In short, ψ = {φ₁, φ₂, φ₃, …, φ$_k$}. The log-likelihood function is being maximized with respect to each of the distribution parameters to obtain the optimal solution.

$$\frac{\partial LLKL(\psi)}{\partial \varphi_1} = \frac{\partial LLKL(\psi)}{\partial \varphi_2} = \frac{\partial LLKL(\psi)}{\partial \varphi_3} = ... = \frac{\partial LLKL(\psi)}{\partial \varphi_k} = 0 \qquad (10)$$

Remember we mentioned earlier that a simple straight line fit to the median rank data on the probability plot paper would be sufficient to approximately determine the value of the distribution parameters. This method would however work only in the case where the failure data collected consists of a single distribution. In many cases, if there are more than one failure mechanisms in a device, then each failure mechanism would have its own statistical distribution with a uniquely defined set of parameters and hence, the overall failure data plotted would contain more than one distribution which cannot be represented by a single straight line. Such distributions which comprise of a mixture of more than one distribution are called **mixture distributions** (Titterington *et al.,* 1985). In the case of parameter estimation for a mixture distribution, simulated annealing is all the more useful

since it helps determine the optimal set of parameters for any mixture distribution given its ability to perform global optimization for any n-dimensional system.

In summary, this section has provided a brief overview on the fundamentals of reliability statistics. The statistical theory on reliability data analysis was introduced and the relevance of simulated annealing in this context was highlighted. It should now be clear that our purpose is to make use of SA to globally maximize the n-dimensional log-likelihood function which will in turn help determine the optimal values of the distribution parameters that fit a given set of failure test data. In the next section, we shall treat mixture distribution in greater detail and show the stochastic process associated with it and the final form of the log-likelihood (*LLKL*) function for a mixture distribution that we will be maximizing.

3. Mixture distribution analysis

3.1 What are mixture distributions?

Any product or device in the field could fail due to various reasons. It is rare to find a device which fails due to only a single cause. In some cases, the presence of a failure mechanism triggers other failure mechanisms to evolve and as a result, the final device failure could be caused by the interaction and combined effects of these multiple failure mechanisms.

Every failure mechanism has its own statistical distribution that is dictated by the inherent physics of failure and the rate of degradation of the device is governed by the kinetic processes embedded in the failure physics. When there are more than one failure mechanisms, there are more than one statistical distributions present in the failure data. Therefore, the overall distribution describing the data is a ***mixture distribution***. Fig 3 shows the mixture distribution PDF plot consisting of two component distributions corresponding to two different failure mechanisms.
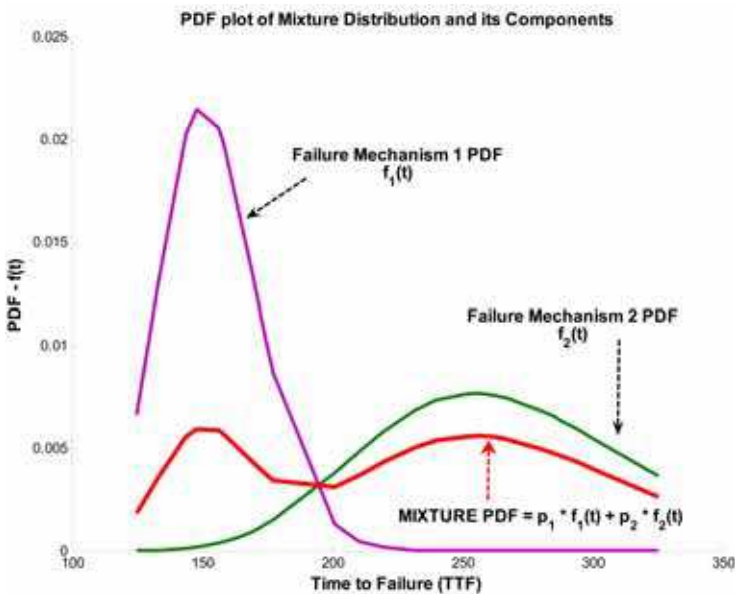

Fig. 3. PDF plot illustrating the effect of multiple failure mechanisms on the mixture distribution PDF.

Such mixture distributions are very commonly observed in ALT test data due to the high stress employed in the test. In the field of microelectronics, for example, package failures could occur due to solder electromigration at the bond pads or due to corrosion effects as a result of exposure to moisture. Similarly, the gate oxide layer in a MOS transistor could fail due to manufacturing induced voids and defects causing extrinsic failures, or percolation / leakage path evolution connecting the metal gate to the Si substrate due to intrinsic failures that evolve as a consequence of aging and wear-out phenomenon. Another popular observation includes the void nucleation and growth in the aluminium or copper interconnect metal lines that connect the transistor devices in an integrated circuit. While some voids evolve in the narrow high current density vertical vias connecting different levels of metallization, others evolve in the main interconnect line itself. One of the later sections clearly describes case studies that show the presence of mixture distributions in electronic device reliability.

3.2 Assumptions

In the mixture distribution analysis technique to be presented in this section, there are a few critical assumptions to be taken note of. One of the key assumptions is that the different failure mechanisms in the tested device are *independent* of each other and hence they do not influence the degradation or failure rate of each other. This assumption helps us in making use of the *principle of superposition* to model the overall mixture PDF. The other assumption we make is that the components of the mixture distribution belong to the same type of statistical distribution (e.g. Weibull, Lognormal etc…) and they are different only in the values of the distribution parameters that they take. Although this assumption is not necessary, it helps simplify the theory presented. Also, we assume that the number of distribution components in a given set of data is known apriori. There are various methods in the statistical literature that help estimate the number of distribution components in a given set of failure test data (Akaike, 1974; Bucar *et al.,* 2004). However, they are beyond the scope of our study here.

3.3 Mixture distribution theory

Let us now take a closer look at the statistics underlying the mixture distribution theory (Titterington *et al.,* 1985). For a device / system with *n* failure mechanisms each with its own failure distribution, the probability density function of the mixture distribution, $f_{MIX}(t)$, is given by (11) where $\{p_1, p_2, p_3, …, p_n\}$ refer to the mixing weight or proportion of each component distribution in the overall mixture and $f_k(t)$ refers to the PDF of the kth component failure distribution; $k \in [1, n]$.

$$f_{MIX}(t) = p_1 \cdot f_1(t) + p_2 \cdot f_2(t) + ... + p_k \cdot f_k(t) + ... + p_n \cdot f_n(t) = \sum_{k=1}^{n} p_k \cdot f_k(t) \quad (11)$$

Based on this expression, the mixture CDF is given by (12) where $F_{MIX}(t)$ refers to the overall mixture distribution CDF and $F_k(t)$ corresponds to the CDF of each individual component distribution. The expression in (12) is obtained by a simple integration of (11) respect to time, *t*.

$$F_{MIX}(t) = \int_0^t f_{MIX}(t)dt = \sum_{k=1}^n \left[p_k \cdot \int_0^t f_k(t)dt \right] = \sum_{k=1}^n p_k \cdot F_k(t) \tag{12}$$

A similar expression to that in (12) may be written for the mixture distribution reliability function, $R_{MIX}(t) = 1 - F_{MIX}(t)$. Based on the above expressions, the log-likelihood function (*LLKL*) for a mixture distribution may be expressed as in (13), where $\psi = \{\varphi_1^{(1)}, \varphi_2^{(1)}, \varphi_1^{(2)}, \varphi_2^{(2)}, \ldots, \varphi_1^{(n)}, \varphi_2^{(n)}\}$ represents all the parameters of the n-component mixture distribution. The superscript (n) in the above notation refers to the index of each component distribution and **n'** and **r'** are the number of failure and censor data.

$$\begin{aligned}
LLKL(t;\psi) &= \ln\left(\frac{n'!}{(n'-r')!} \right) + \sum\nolimits_f \ln\left[f_{MIX}\left(t_f | \psi \right) \right] + \sum\nolimits_c \ln\left[R_{MIX}\left(t_c | \psi \right) \right] \\
&= \ln\left(\frac{n'!}{(n'-r')!} \right) + \sum\nolimits_f \ln\left[\sum_{k=1}^n p_k \cdot f_k\left(t_f | \psi \right) \right] + \sum\nolimits_c \ln\left[\sum_{k=1}^n p_k \cdot R_k\left(t_c | \psi \right) \right]
\end{aligned} \tag{13}$$

For example, in the case of a *Weibull n-component* mixture distribution, $\psi = \{\beta_1, \eta_1, p_1, \beta_2, \eta_2, p_2, \ldots, \beta_n, \eta_n, p_n\}$ where β, η and p refer to the shape parameter, scale parameter and mixing weight of each component of the mixture distribution. Since every failure results from one of the n-component distributions, the sum of all the mixing weights must add up to 1 as shown in (14). A higher mixing weight for a particular component implies that the failure mechanism corresponding to it is more dominating than other existing secondary failure mechanisms.

$$\sum_{k=1}^n p_k = 1 \tag{14}$$

Having developed the expression for *LLKL*, the optimal set of parameters in the set, ψ, are obtained by global maximization of the multi-dimensional *LLKL* function for which, as mentioned earlier, SA is one of the best techniques to use. The number of dimensions or independent variables in the *LLKL* function may be determined using (15) where n is the number of component distributions, r is the number of distribution parameters for each component distribution and m is the overall dimension of the objective function (*LLKL*) to be optimized. In (15), (r + 1) represents the number of distribution parameters accounting for the mixing weight p in addition to the standard parameters of the statistical distribution function.

$$m = n \cdot (r+1) - 1 \tag{15}$$

In this section, a concise description of the mixture distribution theory has been provided. The usefulness and relevance of mixture distributions in device reliability analysis is highlighted. The log-likelihood function for the case of mixture distributions is developed and this well-defined objective function will next need to be optimized using the SA method.

The next section describes the SA methodology and its algorithm in sufficient detail along with an insight into the interesting origin of this method. This will then be followed by two case studies on applying SA to reliability analysis of two critical microelectronic device failure mechanisms viz. gate oxide failures and electromigration phenomenon.

4. Application of simulated annealing

The methodology of Simulated Annealing can be traced back to the thermodynamic and kinetic processes of cooling, annealing and crystallization of materials such as metals and some liquids (Brooks *et al.*, 1995). When the initial temperature of a material is high, the atoms in it have a high diffusivity such that they can hop around to various lattice positions with different energies. When the temperature is reduced slowly or under conditions of slow cooling or gradual annealing, the atoms have sufficient time to diffuse and find lattice points of subsequently lower energies. Even if an atom happened to settle down at an intermediate local energy metastable equilibrium, the slow cooling and high initial temperature of the material system ensure that the atom overcomes the kinetic activation barrier to jump from the local energy equilibrium well and enter into the well that could possibly contain the global energy minimum. Note that any material would always want to globally minimize its Gibbs free energy and reach the most stable equilibrium state if sufficient time is provided for such a phenomenon to occur. Eventually, as this process of slow cooling is continued and as we let the material system equilibrate at each of the stepwise reductions in the ambient temperature, the atoms are most probable to have entered the energy well containing the global minimum and finally attain the global minimum energy level at sufficiently low temperatures. These low temperatures at later stages of the cooling (annealing) schedule are required in order to ensure that an atom which has entered the global energy minimum well does not jump out of it again.

In contrast, if the rate of cooling was rapid (also called *quenching*), then the atoms are most likely to settle down at local minimal energy metastable equilibrium states, which could result in the formation of polycrystalline or amorphous structures as opposed to crystalline structures that would result during a slow cooling process. Therefore, in order to reliably attain a global minimum energy state, two things are necessary – (A) High initial temperature (T_0) and (B) Slow rate of cooling.

Fig 4 (a) – (c) clearly illustrates the various transitions of atomic energies that could take place at a high initial temperature of T_0, lower intermediate temperature, T_K and final temperature of T_∞ under slow cooling rate conditions. Fig 5 shows the case of rapid cooling that results in a metastable local equilibrium. These figures supplement the explanation above and hopefully give a clear and simple illustration of the physics of annealing.

Based on the kinetics of annealing described above, the above concept has been adopted in optimization literatures as *simulated annealing*. The energy of the atom is analogous to the value of the objective function to be optimized. The position and movement of the atom is analogous to the parameter settings and shifts in its values as the optimization is carried out. The concept of temperature is adopted as it is using the Boltzmann criterion here and we could call the temperature as the "*mathematical temperature*" of the optimization system.

Having understood the origin of simulated annealing as an optimization procedure and its analogy to the physical cooling phenomenon, we shall now get acquainted with some of the common terminologies that we will be using in this section and then proceed on to explain the annealing algorithm in an easy to understand fashion.
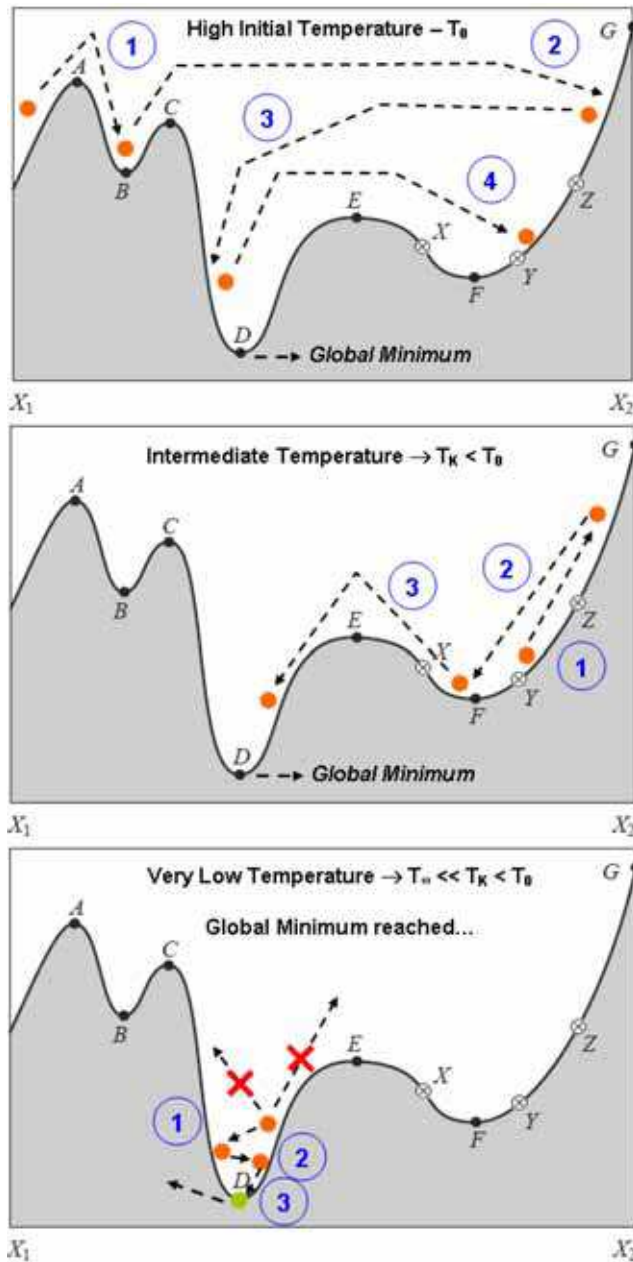
Fig. 4. (a) – (c): Illustrating the energy transitions of an atom in a material which undergoes slow cooling from a high initial temperature of T_0. Note that the basic sketch of the figure has been adopted from (Press *et al.*, 2002). The numbers in circles represent the successful energy transitions from one state to the other.
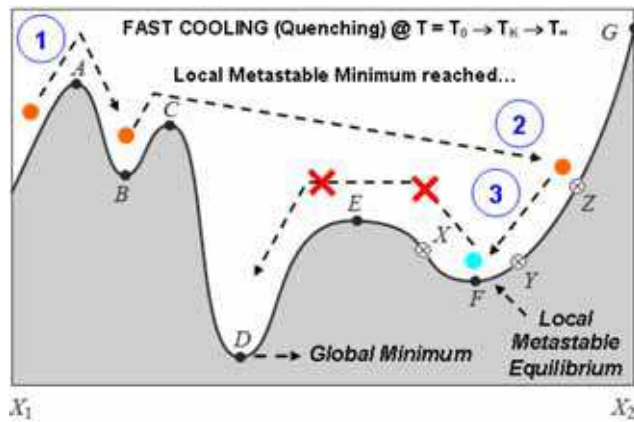
Fig. 5. Illustrating the energy transitions of an atom in a material which is subjected to rapid cooling that prevents it from entering the global energy minimum well. The atom finally attains a local metastable equilibrium state. Note: The basic sketch of the figure has been adopted from (Press *et al.*, 2002). The numbers in circles represent successful energy transitions from one state to the other.

4.2 Terminologies

Different authors use different terminologies to refer to the same parameter in the field of simulated annealing. Therefore, it is necessary to familiarize ourselves with the terms that we will be using in our approach to designing the SA algorithm. Let us now look at some of these terminologies:

- **Temperature Reduction Coefficient (λ)** – Since the accuracy of reaching the global minimum in SA depends on the *rate of cooling*, it is necessary to define rate of cooling as an input parameter of the SA algorithm.

$$T_K = \lambda^K \cdot T_0 \; ; \; K \in Z_0^+ \; ; \; 0 < \lambda < 1 \tag{16}$$

- **Initial Temperature (T$_0$)** – As mentioned earlier, we define the so-called "mathematical temperature" of a system analogous to the physical temperature in a kinetic process. The initial temperature for optimizing an objective function (such as the *LLKL* in our case) has to be kept high and it could be set to the approximate range of variation of the objective function which can be determined by a random space search of the objective function value for different input parameter combinations. Note that a precise value for T$_0$ is not required. We only need to specify a reasonable value for it that would ensure a successful SA algorithm run and a good start for this would be to set T$_0$ to the range of the objective function. We have been advocating that a high value of T$_0$ is essential for the SA to attain the global optimum value. It should be realized however that setting too high a value for T$_0$ makes the SA inefficient as it takes a longer time to reach the global optimum in this case. Very high values of T$_0$ make SA less preferred as it implies slower processing speed, larger memory requirements and increased computational load, all of which are undesirable. There is no hard and fast rule for setting the T$_0$ value. However, good judgment should be exercised when a user sets a value for it.

- **Markov Chain Iteration Number (N)** – This refers to the number of random space searches for the objective function at every temperature value. For every temperature the SA algorithm performs N random space searches in order to approach towards the global minimum. This sequence of N searches in the space spanned by the parameters is considered a *Markov Chain*. The value of N is to be set by the user and it should be set such that a quasi-static equilibrium state is reached at each temperature before transiting to the next lower temperature.
- **Objective Function (E_J)** – As discussed at the end of Section 2, the function to be optimized is called the *objective function* and we denote it by E_J here where E is the equivalent of energy in the physical annealing scenario described earlier and J is the index of the function which means E_J is the value of E after J successful state transitions from the initial value of $E = E_0$ at time t = 0.
- All the objective functions that we would be optimizing are to find the *global minimum* although SA could be easily tuned to find the global maximum too. This is because it is easier to interpret SA for global minimization given its analogy to Gibb's energy minimization of the atoms in a material. Therefore, if any of the functions (such as *LLKL* in our case) need to be maximized, then we can tune the objective function so that it is to be minimized. As an example, instead of maximizing the log-likelihood function (**LLKL**), we could equivalently minimize its negated function (**-LLKL**) since the maximum of a function is the same as the minimum of its negative
- **Boltzmann Theorem of Statistical Physics** – The probability that a system is in some state with energy E is given by (17) where k_b is the Boltzmann's constant and Z(T) is normalization function (Brooks *et al.*, 1995).

$$\Pr(Energy = E) = \frac{1}{Z(T)} \cdot \exp\left(-\frac{k}{k_b T}\right) \tag{17}$$

- **Metropolis Acceptance Criterion (MAC)** – In the physical process of annealing, Boltzmann's theory suggests that an atom could temporarily move from a lower energy state to a higher energy state at times in order to jump out a local minimum energy well in search of the well containing a global minima, although the probability of such jumps to higher energies is quite low. Applying this analogy to our mathematical optimization system, the probability that the system transits from a lower objective function value (E_J) to a higher objective function value (E_{J+1}); $E_{J+1} > E_J$ is given by (18) where $\Delta = (E_{J+1} - E_J)$. Since the accuracy of reaching the global minimum in SA depends on the *rate of cooling*, it is necessary to define rate of cooling as an input parameter of the SA algorithm.

$$\Pr(E_J \rightarrow E_{J+1} | E_{J+1} > E_J) = \exp\left(-\frac{E_{J+1} - E_J}{T}\right)$$
$$\Pr(E_J \rightarrow E_{J+1} | E_{J+1} < E_J) = 1 \tag{18}$$

The Metropolis Acceptance Criterion (MAC) suggests that a state transition from a lower value (EJ) to a higher value (EJ+1) will successfully occur if and only if the probability

of such a transition as given by (18) is more than the random number U generated from a uniform distribution with end limits 0 and 1. In short:

$$\left(E_J \rightarrow E_{J+1} \middle| E_{J+1} > E_J\right) \Leftrightarrow \exp\left(-\frac{E_{J+1} - E_J}{T}\right) \geq U(0,1) \tag{19}$$

With all the terminologies involved in the SA algorithm listed out and interpreted we know that the three main input parameters for executing SA are λ, T_0 and N. Having familiarized with these notations and their meaning, we can now dive straight into the stepwise procedure to be adhered to in implementing the actual SA algorithm.

4.3 Annealing algorithm

The stepwise execution of the annealing algorithm is as follows:

1. For every input parameter, φ_r, of the objective function, $E(\varphi_1, \varphi_2, \varphi_3, \ldots, \varphi_r)$, define the range of values $[\varphi_r^{MIN}, \varphi_r^{MAX}]$ that each of these parameters can take. This range could be guessed by the user's intuition or it could be set as wide as possible based on the realistic limits the parameters could take. For example, the obvious lower and upper limits for the mixing weight in the log-likelihood function has to be 0 and 1 respectively since it is a proportional quantity. If the user has some prior knowledge of the parameters of the objective function, then a narrower range can be defined for the parameters and this would help reduce the size of the parameter space to be spanned for locating the optimum and could help to reach the global minimum faster. Therefore, a narrow range for the parameters would be very useful in improving the efficiency of the SA algorithm.

2. Determine the initial value of temperature, T_0, by performing a random space search over the parameter subspace and finding the range of the objective function values obtained. Set this range to the value of T_0 as indicated in (20).

$$T_0 \equiv \left| E_{MAX} - E_{MIN} \right| \tag{20}$$

3. Set the values for the parameters λ and N. Typical values for λ range from 0.75 to 0.95 while for N, which is the number of markov state transitions for every temperature, values ranging from 1000 to 5000 can be set as a good rule of thumb.

4. At the initial temperature of $T = T_0$, start with an initial guess of the objective function E_0 by using some randomly generated combination of values for the input parameter based on the range defined for them.

5. Compare this value of E_0 with another randomly generated objective function value, E_1 and do the following:

> **IF** $E_1 < E_0$
> $\rightarrow$ Transit from State $E_0 \rightarrow E_1$.
>
> **ELSE-IF** $(E_1 > E_0)$ and ***MAC*** criterion is satisfied
> $\rightarrow$ Transit from State $E_0 \rightarrow E_1$.
>
> **ELSE**
> $\rightarrow$ Remain in State E_0
> $\rightarrow$ Repeat the above steps for a different value of E_1.
>
> **END**

6. The above pseudo code is iteratively followed N times for the initial temperature, T_0. If a transition takes place from $E_J \rightarrow E_{J+1}$ based on the above criteria, we call the transition a *successful transition*. Else, if some iteration does not lead to any transition to a different objective function value, then we refer to it as an *unsuccessful transition*. For the given temperature, T_0, the *fraction of successful state transitions* or what we call **success ratio**, Ω is recorded.

7. Based on the annealing schedule defined by the temperature reduction coefficient in (16), transit to the new lower temperature of $T_1 = \lambda \cdot T_0$ and follow steps (5) and (6) iteratively N times.

8. This temperature reduction takes place sequentially and as steps (5) and (6) are executed for K successive cycles of temperature transitions, the objective function enters the global minima well and slowly approaches the global minimum point.

9. During the initial high temperature conditions, the Boltzmann probability is expected to be high and therefore, the MAC criterion is likely to be accepted most of the time thus resulting in a high value for the success ratio, Ω. However, as temperatures are reduced, MAC criterion is rarely satisfied and further transitions to lower objective function values also becomes less likely thus causing the Ω value to decrease. Eventually, after a large number of temperature cycles, we would reach a stage when the Ω value could be as low as 0.0001 which means that 1 in every 10000 transitions is successful. Such low values of Ω clearly indicate that the objective function has approached very close to the global minimum. In such a case, further algorithm execution is no longer necessary and the SA routine can be stopped. Therefore, the Ω parameter helps us define a stopping criterion that dictates the end of the SA routine code. Typically, we set values of $\Omega = 0.0001$ (1×10^{-4}) or 0.00001 (1×10^{-5}).

```
START

Initial Definition – {N, λ, T₀}

Define Ω = 1, K = 0, J = 0.

Random space search → E_J.

WHILE (Ω > 0.0001)

      T_K = λ^K • T₀;

            FOR (x = 1:N)

                  Random space search → E_{J+1}.

            IF E_{J+1} < E_J
                        → Transit from State E_J → E_{J+1}.

            ELSE-IF (E_{J+1} ≥ E_J) and MAC criterion is satisfied
                        → Transit from State E_J → E_{J+1}.

            ELSE
                        → Remain in State E_J
                        → Repeat the above steps for a different value of E_{J+1}.

                  END

            END

END

K = K + 1;

STOP
```

The SA algorithm execution explained above may be summarized in the form of a simple pseudo code as shown. The readers should by now have realized that the SA algorithm is a very easy to understand and simple to implement technique, yet so powerful in its ability to perform multi-dimensional global optimization. Note that for our case of mixture distribution analysis, the objective function, E, above would be replaced by the log-likelihood function (*LLKL*) defined earlier in (13).

4.4 Benefits and drawbacks of simulated annealing

Simulated annealing is a very powerful technique since it enables optimization of any n-dimensional objective function. It is one of the very few methods that can reliably find the global minimum or maximum value as desired. It is easy to comprehend and can be implemented using a simple C program code without any complexities. Another advantage of this approach is the gradient insensitivity since it does not optimize the function by taking its derivative unlike other methods such as the Newton-Raphson for example. Therefore, SA would be able to work for functions which have singularities and also non-analytical functions which might not have a closed form.

There are however, a few drawbacks in using the SA method. It is highly dependent on the initial temperature, T_0. For very large values of T_0, the algorithm might take too long to converge. Compared to other optimization techniques, it is more computationally intensive and relatively slow. The computational time scales exponentially with the dimensions of the objective function to be solved. We have been mentioning that the SA method converges to the global minimum as the success ratio, Ω, attains a very low value such as 0.0001. It should be noted that $\Omega \neq 0$ implies that there is still a possibility of a lower value existing in the global minima well which could not be found by the random searches performed. Therefore, although the SA method helps approach the global minima, it may not necessarily reach the exact optimum point.

These drawbacks necessitate the use of other local optimization algorithms in conjunction with the SA method so that the exact global minimum may be located (Tan *et al.*, 2007a). Also, in order to get around the problem of long execution times of SA, approaches to localize the search of the parameter subspace during the later executions when the global well might have been found could be investigated. A combined global – local search method would help make SA more efficient than its simplest version implemented above. These techniques to improve the SA algorithm will be briefly touched upon in Section 6. In the next section, we show the application of the SA algorithm to two real case studies on the reliability analysis of electronic device failure mechanisms viz. gate oxide breakdown and electromigration.

5. Case study – microelectronic device reliability

The very first attempts to apply *Simulated Annealing* for reliability analysis in microelectronic devices was taken up by Tan *et al.*, (2007a, 2008). This section brings out the application of SA into microelectronics reliability.

5.1 Gate oxide breakdown in MOSFETs
5.1.1 Physics of gate oxide breakdown
The progress in the electronics industry has been accelerating exponentially after the advancement of semiconductor technology and materials. In order to achieve higher

Thank You for previewing this eBook

You can read the full version of this eBook in different formats:

- ➢ HTML (Free /Available to everyone)

- ➢ PDF / TXT (Available to V.I.P. members. Free Standard members can access up to 5 PDF/TXT eBooks per month each month)

- ➢ Epub & Mobipocket (Exclusive to V.I.P. members)

To download this full book, simply select the format you desire below