Probabilistic Evaluation of Growth Models Based on Time Dependent Auger Signals

W. von der Linden, V. Dose, N. Memmel, and R. Fischer

Max-Planck-Institut für Plasmaphysik, EURATOM Association
D-85740 Garching b. München, Germany
e-mail: wvl@ipp-garching.mpg.de

The time-dependent Auger intensity measured during epitaxial growth of ultra-thin films is routinely used to study the growth process. The Auger signal yields integrated information about the layer-thickness distribution which can only be inferred upon solving an inverse problem. Common praxis is to minimize the misfit between experimental data and theoretical prediction made by a growth model under consideration. The purpose of this paper is to demonstrate the consistent application of probability theory to assess possible growth models and determine the respective parameters without resorting to ad-hoc methods.

PACS: 68.55.-a , 02.50.-r , 07.05.Kf , 29.85.+c

1 Introduction

Auger-electron spectroscopy (AES) is a routinely used technique to investigate thin film growth modes [1]. While keeping the evaporation rate fixed, the Auger signal versus time (AS-t) plot is measured continuously during film growth. The Auger intensity depends on the proportion of stacks of surface atoms of height i above the substrate. The stack-height distribution dependence on time and the resulting time-dependence of the Auger intensity allows therefore to differentiate between different growth models. In this paper we use the rules of probability theory to assess possible growth models and to determine the respective parameters. One essential ingredient of probability theory is the marginalization rule. It allows to integrate out unknown parameters which are essential for the specification of the model. There are two competing factors in the theory. The likelihood factor would favor the model with the least deviation from the data. This extreme case is of little use as it would generally vote for the most complex model resulting in zero misfit. The introduction of the prior probability automatically enforces a tradeoff between the fit of the data and the complexity of the model, yielding an intrinsic formalism for Ockham's Razor: *Essentia non sunt multiplicanda praeter necessitatem*, in other words, the result should not be more complicated than necessary.

The intention of this paper is to present a consistent probabilistic theory to analyze experimental AS-t data. In order to illustrate the theory we apply it to published Auger data [2] for Fe epitaxially grown on Cu. We will, however, not give a complete discussion of the physical problem but we will rather demonstrate the features of the probabilistic approach guided by two possible growth models.

2 Growth models

A single layer of surface material s emits Auger-electrons with intensity $I_s^{(1)}$ and energy E_s. If the Auger electrons traverse n layers of material 'm' the signal is weakened by a factor $(\alpha_{m,E_s})^n$. The transmission coefficient α is related to the inelastic mean free path (IMFP) λ via

$$\alpha = e^{-\frac{d}{\lambda \cos(\theta)}} \qquad , \tag{1}$$

where d stands for the layer thickness and θ is the angle under which the electron travels through the surface. If there are n surface layers, each layer emits Auger electrons which are then diminished by the outer layers and the total intensity reads

$$I_s^{(n)} = I_s^{(1)} \sum_{i=0}^{n-1} (\alpha_{s,E_s})^i = I_s^{(1)} \frac{1-(\alpha_{s,E_s})^n}{1-\alpha_{s,E_s}} = I_{\infty,s}(1-(\alpha_{s,E_s})^n) \qquad , \tag{2}$$

where $I_{\infty,s}$ stands for the bulk intensity of the surface material. In general, the substrate will not be covered homogeneously by surface atoms. The surface is characterized by a height distribution p_n with $\sum_n p_n = 1$ and the total Auger intensity becomes

$$I_s = I_{\infty,s} \sum_{n=0}^{\infty} p_n (1-(\alpha_{s,E_s})^n) \qquad , \tag{3}$$

provided inter-diffusion of bulk and surface material does not occur. In order to eliminate experimental variations, e.g. in the incoming beam intensity, the normalized intensity $r = I_s/(I_s + I_b)$ is measured, where I_b represents the

Auger signal of the substrate. The bulk intensity of the substrate $I_{\infty,b}$ is attenuated by the surface layers of material 's'

$$I_b = I_{\infty,b} \sum_{n=0}^{\infty} p_n (\alpha_{s,E_b})^n \qquad . \tag{4}$$

In general the transmission coefficients α_{s,E_s} and α_{s,E_b} as well as the Auger intensities $I_{\infty,s}$ and $I_{\infty,b}$ differ and

$$I_b + I_s = I_{\infty,s} \sum_n p_n \left\{1 - (\alpha_{s,E_s})^n + \rho\, (\alpha_{s,E_b})^n)\right\} \tag{5}$$

with $\rho = I_{\infty,b}/I_{\infty,s}$. The intensity ratio is therefore

$$r = \frac{I_s}{I_s + I_b} = r_\infty \frac{\sum_n p_n \left\{1 - (\alpha_{s,E_s})^n\right\}}{\sum_n p_n \left\{1 - (\alpha_{s,E_s})^n + \rho(\alpha_{s,E_b})^n)\right\}} \qquad . \tag{6}$$

According to the derivation, we expect $r_\infty = 1$ if the assumptions we made are valid. A value for r_∞ which significantly differs from unity will most likely be an indication for inter-diffusion of bulk and surface materials.

The above equations are only correct under certain geometrical constraints which we will digress into now. Auger electrons are commonly detected by means of retarding field or cylindrical mirror analyzers with a relatively large solid angle [3]. This means there is not a fixed angle θ in eq. (1). The detector rather averages over a range $\theta \in [-\theta_{\max}, \theta_{\max}]$. Consequently the damping factors α^n due to n layers of thickness d are replaced by

$$\overline{\alpha^n} = \frac{\displaystyle\int_{|\theta|<\theta_{\max}} e^{-\frac{nd}{\lambda\cos(\theta)}}\, d\Omega}{\displaystyle\int_{|\theta|<\theta_{\max}} d\Omega} = \frac{\left. e^{-\frac{nd}{\lambda\xi}} + c\mathrm{Ei}(-\frac{nd}{\lambda\xi})\right|_{\xi=\cos(\theta_{\max})}^{1}}{1-\cos(\theta_{\max})} \qquad , \tag{7}$$

where Ei stands for the exponential integral. It is common practice to replace $\overline{\alpha^n}$ by $\exp(-nd/\lambda\cos(\theta_{\text{eff}})) = \alpha^n_{\text{eff}}$ upon introducing an effective angle θ_{eff} which can be determined via $\cos(\theta_{\text{eff}}) = \overline{\cos(\theta)} = \cos^2(\theta_{\max}/2)$. Typically $\cos(\theta_{\text{eff}}) = 0.74$ [2] which corresponds to $\theta_{\max} = 61°$. For this value the relative error $\varepsilon = \overline{\alpha^n}/\alpha^n_{\text{eff}} - 1$ of the standard approximation versus the ratio nd/λ is plotted in fig.1. If the thickness of the surface in units of the IMFP (nd/λ) is less than 2 the ubiquitous approximation α^n_{eff} is valid ($\varepsilon < 4\%$) while it rapidly becomes wrong for larger values of nd/λ. The condition is fulfilled for the data set analyzed in section 4.2.

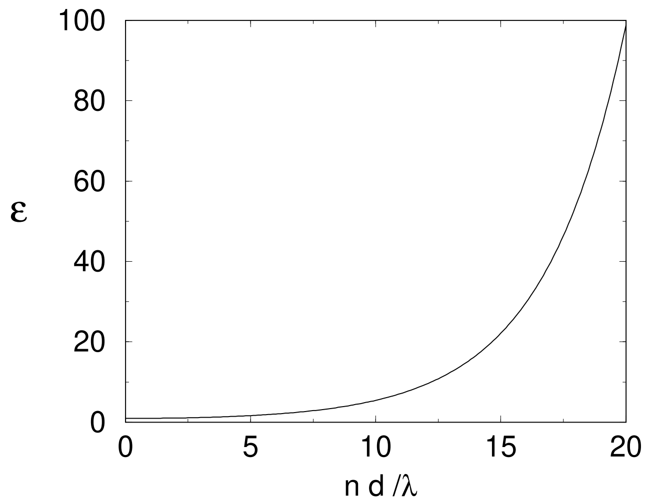

Fig. 1. *Relative error* $\varepsilon = \overline{\alpha^n}/\alpha_{\text{eff}}^n - 1$ *of the standard approximation as explained in the text.*

2.1 Layer-by-layer growth

In the layer-by-layer [Frank van der Merwe (FM)] growth model the layers are gradually filled one after the other. The experimental growth rate is $1/T$, where T stands for the time needed to complete one layer. The average thickness D at time t is $D = t/T$. We define the times t_n at which n layers are exactly filled. In addition we need $t_0 = 0$. For fractional filling of a new layer i.e. for time $t \in (t_n, t_{n+1})$, a fraction $q = (t - t_n)/T$ of the surface is covered by $n+1$ layers, while the remaining part of the surface contains n layers. Therefore, the measured intensity is

$$
\begin{aligned}
r &= r_\infty \left((1-q)\{1-(\alpha_{s,E_s})^n\} + q\{1-(\alpha_{s,E_s})^{n+1}\} \right) /N \qquad (8)\\
N &= (1-q)\{1-(\alpha_{s,E_s})^n + \rho(\alpha_{s,E_b})^n\} + q\{1-(\alpha_{s,E_s})^{n+1} + \rho(\alpha_{s,E_b})^{n+1}\} \quad .
\end{aligned}
$$

In order to keep the formalism transparent and to be able to perform most of the mathematical steps analytically we consider the special case of identical transmission coefficients and intensities for iron and copper, i.e. $N = 1$. It will be shown that this is a valid approximation for the example we will study [2]. In the following discussion we use $\alpha = \alpha_{s,E_s} = \alpha_{s,E_b}$. For $N = 1$ the time dependence of the intensity interpolates linearly between the values of completely filled layers as depicted in fig.2. Due to experimental variations, particularly in the evaporation rate, T may vary slightly from layer to layer. The time to grow layer n be $T_n = t_{n+1} - t_n$. The time dependence of the Auger intensity is still determined by eq. (8), merely q is modified to $q = (t - t_n)/(t_{n+1} - t_n)$. The time-dependence of the normalized Auger intensity becomes

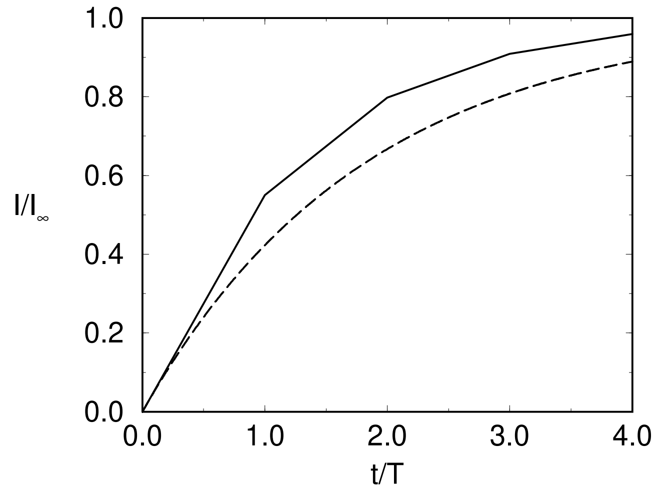

Fig. 2. *Auger intensity as function of time for the two growth models: layer-by-layer (FM) growth (solid line) and simultaneous multilayer (SM) growth (dashed line). Both model curves are computed for transmission coefficient* $\alpha = 0.45$.

$$\begin{aligned}\tilde{r}(t) &= r_\infty \left((1-q(t))(1-\alpha^{n(t)}) + q(t)(1-\alpha^{n(t)+1}\right) \;=\; r_\infty \psi(t) \qquad (9)\\ n(t) &= \text{largest integer } n \text{ with } t_n < t\\ q(t) &= (t - t_{n(t)})/(t_{n(t)+1} - t_{n(t)})\end{aligned}$$

In this model the signal depends on the transmission coefficient α, the normalized bulk intensity r_∞ and the times T_n to grow the various layers.

On passing we want to note that the above model equations are not applicable if the evaporation process is not well controlled, i.e. if the deposition rate is not constant during the growth of one layer but rather has some unknown systematic trend. We will see that this is not the case in the data set we are analyzing.

2.2 Simultaneous multilayer growth

The layer-by-layer (FM) growth model assumes that the atoms which randomly hit the surface during the evaporation process move freely on the surface and try to get as close as possible to the substrate, thereby filling the layers.

Another scenario would be that the atoms may move on the surface but they are inclined not to change layers, which would be the case if there are repulsive forces towards the boundary of a finite layer. The Auger signal, as we derived it in the previous section, depends only on the probability p_i to find atomic stacks of height i on the substrate. The position of the stacks is unimportant, i.e. the Auger spectrum is the same whether the stacks form bizzare structures or compact islands. The diffusion of atoms within the same layer has therefore no effect on the Auger signal and we can ignore the dynamics of the atoms entirely. The growth obeys therefore a Poisson process and

$$p_i = e^{-D}\frac{D^i}{i!} \quad , \tag{10}$$

where $D = D(t) = t/T$ is the mean number of layers at time t. Stacks of height i yield an Auger intensity $I_\infty(1-\alpha^i)$ and the contribution of all stacks is

$$\tilde{r}(t) = r_\infty \sum_{i=0}^{\infty} p_i(1-\alpha^i) \; = \; r_\infty(1-e^{-(1-\alpha)t/T}) = r_\infty\psi(t) \quad . \tag{11}$$

It has the same form as eq. (9). Although the functions ψ differ in the two models we will subsequently use the same symbol to keep the notation simple. The meaning will always be clear from the context. Here the signal depends on the transmission coefficient α, the normalized bulk intensity ratio r_∞ and a mean time T to deposit a layer. We will use the notation $\mathcal{T}$ to specify the set $\{T_n\}$ in the case of the layer growth model and the mean growth time T in the case of simultaneous multilayer (SM) growth.

Like for the FM model it is also necessary for the SM model that the deposition rate is constant in time apart from statistical fluctuations which are covered by the error statistics of the Auger signal.

2.3 Goodness of fit

The experimental data comprise N pairs (r_i, τ_i) of intensity and time. The goodness of fit is defined by

$$\chi^2 = \sum_{i=1}^{N}(r_i - r_\infty\psi(\tau_i))^2/\sigma^2 \quad , \tag{12}$$

where σ is the noise level. The model specifies the theoretical prediction $\tilde{r}(t)$, given the model parameters $r_\infty, \alpha, \mathcal{T}$. The following discussion simplifies considerably if we use vector notation:

$$\begin{aligned}
\vec{\hat{r}} &= r_\infty(\psi(\tau_1),\psi(\tau_2),\ldots,\psi(\tau_N))^T = r_\infty\vec{\psi} &&= r_\infty\sqrt{\psi^2}\,\vec{\hat{\psi}} \\
\vec{r} &= (r_1,r_2,\ldots,r_N)^T &&= \sqrt{r^2}\,\vec{\hat{r}} \\
r^2 &= \langle\vec{r},\vec{r}\rangle \\
\psi^2 &= \langle\vec{\psi},\vec{\psi}\rangle
\end{aligned} \quad . \tag{13}$$

The scalar product is represented by $\langle\vec{x},\vec{y}\rangle$. The maximum likelihood (ML) estimate r_∞^{ML} for the bulk intensity r_∞ is obtained upon minimizing χ^2 defined in eq. (12)

$$r_\infty^{\mathrm{ML}} = \sqrt{\frac{r^2}{\psi^2}}\,\langle\vec{\hat{r}},\vec{\hat{\psi}}\rangle \qquad . \tag{14}$$

For reasons of convenience in subsequent sections we rewrite the misfit introducing r_∞^{ML}

$$\sigma^2\chi^2 = \psi^2\,(r_\infty - r_\infty^{\mathrm{ML}})^2 + r^2\,(1-\langle\vec{\hat{r}},\vec{\hat{\psi}}\rangle^2) \qquad . \tag{15}$$

The ML estimate for the transmission coefficient α follows upon minimizing the remaining misfit once we replaced r_∞ by r_∞^{ML} or alternatively upon maximizing the overlap between unit data and model vector $\langle\vec{\hat{r}},\vec{\hat{\psi}}\rangle$ with respect to α.

3 Bayesian Probability Theory

In this section we outline the basics of Bayesian probability theory (BPT) necessary for the understanding of this paper.

We consider propositions (statements which are either true or false) A, B and C. In this paper we will encounter two types of propositions:
1) A: a given model is the correct model to describe the physics of the problem under consideration. In this case $P(A|B)$ denotes the probability for A being true given that another proposition B is true.
2) A: the value of a parameter lies in the interval $(x, x+dx)$. In this case we will introduce the probability density $p(x|B)$. The probability for A being true, given B is true, is then $p(x|B)dx$.

BPT rests upon two fundamental rules which can be derived from basic axioms [4]. The first rule is the sum rule:

$$p(A \vee B|C) = P(A|C) + P(B|C) \tag{16}$$

The symbols $\vee$ stand for logical "or". Eq. (16) is only correct for exclusive propositions A, B, that is if A is true B must be false and vice versa. In this paper we will only consider exclusive propositions. The sum rule can immediately be generalized to a set of exclusive propositions A_i with $i = 1, \ldots, N$

$$P(\bigvee_{i=1}^{N} A_i|C) = \sum_{i=1}^{N} P(A_i|C) \qquad . \tag{17}$$

If in addition the set is exhaustive, i.e. one of the propositions A_i is certainly true, in mathematical terms ($\bigvee\limits_{i=1}^{N} A_i$ = true), the left hand side of eq. (17) is one and we have the marginalization rule

$$P(B|C) = P(\{\bigvee_{i=1}^{N} A_i\}\, B|C) = \sum_{i=1}^{N} P(A_i\, B|C) \qquad . \tag{18}$$

The logical 'and' of two propositions A and B is denoted by $A\,B$. The second rule of BPT is the product rule

$$P(A\,B|C) = P(A|B\,C)P(B|C) = P(B|A\,C)P(A|C) \tag{19}$$

which allows to express the joint probability for proposition A and B in terms of conditional probabilities for A and B separately. The product rule leads immediately to Bayes' theorem

$$P(A|B\,C) = \frac{P(B|A\,C)P(A|C)}{P(B|C)} \qquad . \tag{20}$$

It allows to determine the posterior probability $P(A|B\,C)$ from the likelihood $P(B|A\,C)$ and the prior $P(A|C)$. The denominator serves in this respect as normalizing factor. Eq. (20) also tells us how probabilities are updated from prior to posterior when new evidence B is given. We will make intensive use of a combination of both rules for the case that the probability $P(A|C)$ can only be determined if we know in addition that one of the propositions of an exhaustive set B_i is true. In this case we use the marginalization rule to introduce the set B_i and the product rule to shift the conditional bar

$$P(A|C) = \sum_{i=1}^{N} P(A\,B_i|C) = \sum_{i=1}^{N} P(A|B_i\,C)\,P(B_i|C) \qquad . \tag{21}$$

In the case of continuous propositions this version of the marginalization rule reads

$$P(A|C) = \int dx\, P(A|x\,C)\, p(x|C) \qquad . \tag{22}$$

The rules of probability theory tell us how to manipulate conditional probabilities. Eventually we also need to assign prior probabilities. The general theory to assign probabilities has been elaborated upon by Jaynes [5]. For the following problem we need two prior probabilities. Firstly, the prior probability density $p(x|I)$ for a parameter x which according to some background information I is restricted to the interval (x_0, x_1). The correct prior for such a location parameters is constant $p(x|I) = 1/(x_1 - x_0)\theta(x_0 < x < x_1)$. The generalization of the step function θ is obvious. Secondly, we need the prior for a scale parameter, like the noise level σ. Due to Jeffreys the correct prior [5] is $p(\sigma|\mathcal{I}) = c_J/\sigma$, with a normalization constant c_J. Jeffreys' prior is improper (not normalizable) and it has to be understood as the limit of a sequence of proper priors. Details about the correct limiting process for improper priors are discussed by Bretthorst [6]. In this paper we do not have to care about these subtleties, as all posterior probabilities which we will encounter are proper and the prior volume factor c_J drops out in model comparison.

Now we are prepared to compute whatever probabilities we are interested in. These probabilities will, however, be generally too complex to be plotted and we need other means to summarize the result. It is common praxis to compute posterior expectation values. Let's assume we have determined the posterior probability $p(\vec{y}|D\,\mathcal{I})$ for a vector of parameters $\vec{y}$ in the light of some data D and further background information $\mathcal{I}$. We could summarize the result by the expectation value $\mathcal{E}[y_i|D\,I]$ of the components and by the covariance matrix $\mathcal{E}[\Delta y_i \Delta y_j|D\,I]$, with $\Delta y_i = y_i - \mathcal{E}[y_i|D\,I]$. In general we might be interested in a functional $f(\vec{y})$. Since the posterior probability is conditional so are posterior expectation values. This point is allowed for in the notation

$$\mathcal{E}[f(\vec{y})|D\,\mathcal{I}] = \int dV_y\, f(\vec{y})\, p(\vec{y}|D\,\mathcal{I}) \qquad . \tag{23}$$

The marginalization rule eq. (22) leads to the intuitively obvious expression

$$\mathcal{E}[f(\vec{y})|D\,\mathcal{I}] = \int dx\, \mathcal{E}[f(\vec{y})|x\,D\,\mathcal{I}]\, p(x|D\,\mathcal{I}) \qquad . \tag{24}$$

The posterior expectation value is the average posterior expectation in the light of an additional parameter x. An example would be that the data D do not contain the noise level. To evaluate the likelihood, the noise level σ is however, necessary. According to eq. (24) we sum over the posterior expectation values for all σ with weight $p(\sigma|D\,\mathcal{I})$. As mentioned earlier, we assume that the noise level is not known and hence $p(\sigma|D\,\mathcal{I}) = p(\sigma|\mathcal{I})$ which for the scale parameter σ is Jeffreys' prior.

4 Application to growth models

We will use Bayesian probability theory to assess various growth models. For each model, in particular for the most probable one, we will be interested in the model parameters: noise level, growth time $\mathcal{T}$, bulk intensity ratio r_∞ and transmission coefficient α.

We start out with the probability for the noise level σ of the data, which is hardly accessible from the experimental side. The analysis for the other parameters is analogous. The key quantity is $P(\sigma|\vec{r}\,\mathcal{I})$, the probability for the noise level in the light of the data. The orthodox approach would be to do a least squares fit for a particular model and use the so-called unbiased estimator

$$\sigma_{\text{est}}^2 = \frac{1}{N - N_p} \sum_{i=1}^{N} (r_i - r_\infty \psi(\tau_i))^2 \tag{25}$$

for the noise level. N_p is the number of parameters in the model. It is obvious that the result will depend upon the model: A more flexible model with the same number of parameters will yield a smaller σ_{est} than a rigid model as it fits the data better. With increasing complexity of the model, the reconstruction will start fitting the noise and eventually lead to a pointwise reconstruction of the data. The rules of probability theory take this point properly into account, as one has to sum over all models eq. (21)

$$p(\sigma|\vec{r}\,\mathcal{I}) = \sum_M p(\sigma|\vec{r}\,M\,\mathcal{I}) P(M|\vec{r}\,\mathcal{I}) \qquad . \tag{26}$$

The term $P(M|\vec{r}\,\mathcal{I})$ contains two competing factors. One stemming from the likelihood which favors the model with the least deviation from the data. The second is introduced by the prior probability and penalizes complexity. The trade-off will be a solution that fits the data but not the noise. To proceed we need to evaluate $p(\sigma|\vec{r}\,M\,\mathcal{I})$, the probability for the parameter in the light of the data, given a particular model. On top of this, we need the probability for the model in the light of the data $P(M|\vec{r}\,\mathcal{I})$. To evaluate $p(\sigma|\vec{r}\,M\,\mathcal{I})$ we use the marginalization rule eq. (22) to introduce further model parameters

$$p(\sigma|\vec{r}\,M\,\mathcal{I}) = \int d\mathcal{T}\, d\alpha\, p(\sigma|\vec{r}\,\mathcal{T}\,\alpha\,M\,\mathcal{I})\, p(\mathcal{T}\,\alpha|\vec{r}\,M\,\mathcal{I}) \qquad . \tag{27}$$

For the posterior expectation of a function $f(\sigma)$ of σ we have according to eq. (24)

$$\mathcal{E}[f(\sigma)|\vec{r}\, M\, \mathcal{I}] = \int d\mathcal{T}\, d\alpha\, \mathcal{E}[f(\sigma)|\vec{r}\, \mathcal{T}\, \alpha\, M\, \mathcal{I}]\, P(\mathcal{T}\, \alpha|\vec{r}\, M\, \mathcal{I}) \quad . \tag{28}$$

The probability for σ reads due to Bayes' theorem

$$p(\sigma|\vec{r}\, \mathcal{T}\, \alpha\, M\, \mathcal{I}) = \frac{p(\vec{r}|\sigma\, \mathcal{T}\, \alpha\, M\, \mathcal{I})\, p(\sigma|\mathcal{I})}{p(\vec{r}|\mathcal{T}\, \alpha\, M\, \mathcal{I})} \quad , \tag{29}$$

where we used the fact that the noise level is logically independent of the model parameters and $p(\sigma|\mathcal{I})$ is according to section 3 Jeffreys prior. The marginal likelihood $P(\vec{r}|\sigma\, \mathcal{T}\, \alpha\, M\, \mathcal{I})$ is related to the likelihood $P(\vec{r}|\sigma\, \mathcal{T}\, \alpha\, r_\infty\, M\, \mathcal{I})$, i.e. the error statistics of the experiment

$$P(\vec{r}|\sigma\, \mathcal{T}\, \alpha\, r_\infty\, M\, \mathcal{I}) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2}\chi^2\right\} \tag{30}$$

This is the entity most statistical evaluations are based upon. The (ML) solution, for example, is obtained upon maximizing the likelihood term with respect to the unknown parameters. So far we have left out the bulk intensity ratio r_∞ which will be introduced now to derive the ultimate expression for the marginal likelihood entering eq. (29).

$$p(\vec{r}|\sigma\, \mathcal{T}\, \alpha\, M\, \mathcal{I}) = \int dr_\infty\, p(\vec{r}|\sigma\, \mathcal{T}\, \alpha\, r_\infty\, M\, \mathcal{I})\, p(r_\infty|\mathcal{I}) \quad . \tag{31}$$

Again the bulk intensity is logically independent of the other parameters and the second factor simplified to the bare prior for r_∞. Our prior knowledge about the bulk intensity is that it is positive and we know that there exists an upper limiting $r^{\max}$ which should not exceed unity by much, since we are analyzing normalized intensities. We do not assume more knowledge about r_∞ and the prior describing this state of knowledge consistently is equal to $1/r^{\max}$ between zero and the upper limit and it is zero otherwise. Using eq. (15) and eq. (30) we have

$$\begin{aligned} p(\vec{r}|\sigma\, \mathcal{T}\, \alpha\, M\, \mathcal{I}) = & \frac{(2\pi\sigma^2)^{-\frac{N}{2}}}{r^{\max}} \exp\left\{-\frac{r^2}{2\sigma^2}(1 - \langle\hat{\vec{r}},\hat{\vec{\psi}}\rangle^2)\right\} \\ & \times \int_0^{r^{\max}} dr_\infty\, \exp\left\{-\frac{\psi^2}{2\sigma^2}\,(r_\infty - r_\infty^{\mathrm{ML}})^2\right\} \quad . \end{aligned} \tag{32}$$

The precise value for $r^{\max}$, however, is not needed since the prefactor will drop out in model comparison and the upper integration limit has no influence on the result. This is because the likelihood as function of r_∞ is a Gaussian centered at $r_\infty^{\rm ML}$ with width $\sigma_L = \sqrt{\sigma^2/\psi^2}$. According to its definition $\psi^2 = NO(1)$. The relative width of the likelihood is therefore

$$\frac{\sigma_L}{r_\infty^{\rm ML}} = \frac{\sigma}{\sqrt{\psi^2} r_\infty^{\rm ML}} \approx \frac{1}{\sqrt{\#\text{ data}} \times \text{signal to noise ratio}} \qquad . \tag{33}$$

For any reasonable data set the signal to noise ratio as well as the number of data points is much greater than one and therefore $\sigma_L/r_\infty^{\rm ML} \ll 1$. In the data set which we will discuss in this paper $\sigma_L/r_\infty^{\rm ML} \approx 2\cdot 10^{-3}$. I.e. the integrand is restricted to a very narrow region around $r_\infty^{\rm ML}$ and the integration limits can therefore be extended to $-\infty$ and ∞ leading to

$$p(\vec{r}|\sigma\, \mathcal{T}\, \alpha\, M\, \mathcal{I}) = \frac{(2\pi\sigma^2)^{-\frac{N-1}{2}}}{r^{\rm max}} \exp\left\{-\frac{I^2}{2\sigma^2}(1-\langle\vec{\hat{r}},\vec{\hat{\psi}}\rangle^2)\right\} / \sqrt{\psi^2} \qquad . \tag{34}$$

We know now the numerator in eq. (29). The denominator is simply the normalization factor, i.e. the integral of the enumerator over σ. The integral can also be determined analytically

$$\begin{aligned} p(\vec{r}|\mathcal{T}\,\alpha\,M\,\mathcal{I}) &= \int_0^\infty d\sigma\, p(\vec{r}|\sigma\,\mathcal{T}\,\alpha\,M\,\mathcal{I})\, p(\sigma|\mathcal{I}) \\ &= \frac{c_J (2\pi)^{-\frac{N-1}{2}}}{r^{\rm max}\sqrt{\psi^2}} \int_0^\infty \frac{d\sigma}{\sigma} (\sigma^2)^{-\frac{N-1}{2}} \exp\left\{-\frac{r^2}{2\sigma^2}(1-\langle\vec{\hat{r}},\vec{\hat{\psi}}\rangle^2)\right\} \\ &= \frac{c_J (r^2\pi)^{-\frac{N-1}{2}}\Gamma(\frac{N-1}{2})}{r^{\rm max}\sqrt{\psi^2}} (1-\langle\vec{\hat{r}},\vec{\hat{\psi}}\rangle^2)^{-\frac{N-1}{2}} \qquad . \end{aligned} \tag{35}$$

Similarly we obtain the expectation value for the noise level

$$\mathcal{E}[\sigma^2|\vec{r}\,\mathcal{T}\,\alpha\,M\,\mathcal{I}] = \frac{r^2(1-\langle\vec{\hat{r}},\vec{\hat{\psi}}\rangle^2)}{N-3} \qquad , \tag{36}$$

which closely resembles the orthodox result eq. (25). But in contrast to the orthodox result we have to integrate over $\mathcal{T}$, and α (eq. (27)) and not just take the ML-values for all model parameters.

For eq. (27) we still need $p(\mathcal{T}\,\alpha|\vec{r}\,M\,\mathcal{I})$, which can be derived upon resorting to Bayes' theorem once more

$$p(\mathcal{T}\,\alpha|\vec{r}\,M\,\mathcal{I}) = \frac{p(\vec{r}|\mathcal{T}\,\alpha\,M\,\mathcal{I})}{p(\vec{r}|M\,\mathcal{I})}\, p(\mathcal{T}\,\alpha|M\,\mathcal{I})$$

$$= \frac{p(\vec{r}|\mathcal{T}\,\alpha\,M\,\mathcal{I})}{p(\vec{r}|M\,\mathcal{I})}\,p(\mathcal{T}|M\,\mathcal{I})\,\theta(0 \le \alpha \le 1) \tag{37}$$

In the last part we have made use of the logical independence of the model parameters $\mathcal{T}$ and α, i.e. $p(\mathcal{T}\,\alpha|M\,\mathcal{I}) = p(\mathcal{T}|M\,\mathcal{I})p(\alpha|\mathcal{I})$. The meaning of $\mathcal{T}$ and hence the prior depends upon the model while the transmission coefficient α is independent. On top of that, since α is a location parameter confined to the interval $[0, 1]$, the prior (section 3) is simply one in this interval and zero outside.

In summary we have for the posterior expectation of the noise variance σ^2 according to eq. (28) and eq. (36)

$$\mathcal{E}[\sigma^2|\vec{r}\,M\,\mathcal{I}] = \int d\mathcal{T} \int_0^1 d\alpha\, \frac{r^2(1-\langle\vec{\hat{r}},\vec{\hat{\psi}}\rangle^2)}{N-3}\,\rho(\mathcal{T},\alpha) \tag{38}$$

$$\rho(\mathcal{T},\alpha) = \frac{p(\vec{r}|\mathcal{T}\,\alpha\,M\,\mathcal{I})\,p(\mathcal{T}|M\,\mathcal{I})}{p(\vec{r}|M\,\mathcal{I})} \tag{39}$$

$$p(\vec{r}|M\,\mathcal{I}) = \int d\mathcal{T} \int_0^1 d\alpha\, p(\vec{r}|\mathcal{T}\,\alpha\,M\,\mathcal{I})\,p(\mathcal{T}|M\,\mathcal{I}) \qquad . \tag{40}$$

Analogously, the expectation value for an arbitrary function $g(\mathcal{T},\alpha)$ of the model parameters $\mathcal{T},\alpha$ (eq. (23)) in the light of the data and for a given model M is

$$\mathcal{E}[g(\mathcal{T},\alpha)|\vec{r}\,M\,\mathcal{I}] = \int d\mathcal{T} \int_0^1 d\alpha\, g(\mathcal{T},\alpha)\,\rho(\mathcal{T},\alpha) \tag{41}$$

The marginal likelihood $p(\vec{r}|\mathcal{T}\,\alpha\,M\,\mathcal{I})$ is given in eq. (35). A final probability assignment is required for $p(\mathcal{T}|M\,\mathcal{I})$. For the SM growth model the situation is simple. Here $\mathcal{T}$ stands for the mean time T needed to grow one layer. This quantity is again confined to an interval $T \in (0, T_{\max})$ and the prior is

$$p(\mathcal{T}|M_{\mathrm{SM}}\,\mathcal{I}) = \frac{1}{T_{\max}}\theta(0 < T < T_{\max}) \qquad . \tag{42}$$

The stopping criterion of a sensible experiment is such that the measured intensity curve is in the asymptotic regime. Therefore, at least one complete layer has been grown during the complete experiment and accordingly a reasonable choice for $T_{\max}$ is t_N. Again the result is independent of the precise value for $T_{\max}$, as long as it lies in a reasonable regime: the prefactor drops out, both in parameter estimation and model comparison and the integration limit is already so large that it has practically no influence. For the layer growth the prior is somewhat more complicated. Here, we need to consider the fact that the time T_i to grow a layer may deviate from the mean time T by a unknown amount ΔT_i.

$$
\begin{aligned}
p(\mathcal{T}|M_{\mathrm{FM}}\,\mathcal{I}) &= p(T\,\Delta T_1\,\ldots\,\Delta T_L|\mathcal{I}) \\
&= p(T|\mathcal{I}) \cdot \prod_{i=1}^{L} \frac{1}{\sqrt{(2\pi\delta^2)}} \exp\left\{-\frac{1}{2\delta^2}(T_i - T)^2\right\} \qquad . \qquad (43)
\end{aligned}
$$

The prior $p(T|\mathcal{I})$ is the same as in the SM growth model. The deviations from the mean time T are assumed independent and Gaussian with width δ. Experimental experience tells us that the growth times may vary by about 10%, i.e. $\delta \approx .1T$. We will show that the results depend only weakly on the values of δ in a reasonable range.

The integrals over $\mathcal{T}$ and α in eq. (38) and eq. (41) can be done easily by Monte Carlo Means [7] using $\rho(\mathcal{T}, \alpha)$ as probability density to guide the random walk.

4.1 Model Selection

To assess different models we need $P(M|\vec{r}\,\mathcal{I})$, the probability for the model in the light of the data and all available background information. Bayes' theorem links this unknown probability to the known marginal likelihood and the prior for model M

$$
P(M|\vec{r}\,\mathcal{I}) = \frac{p(\vec{r}|M\,\mathcal{I})P(M|\mathcal{I})}{p(\vec{r}|\mathcal{I})} \qquad . \qquad (44)
$$

To eliminate nuisance prefactors one computes the odds-ratio for models M_1 and M_2

$$
\frac{P(M_1|\vec{r}\,\mathcal{I})}{P(M_2|\vec{r}\,\mathcal{I})} = \frac{p(\vec{r}|M_1\,\mathcal{I})}{p(\vec{r}|M_2\,\mathcal{I})} \cdot \frac{P(M_1|\mathcal{I})}{P(M_2|\mathcal{I})} \qquad , \qquad (45)
$$

which consists of the likelihood ratio and prior odds. In the present problem our background information does not differentiate between the different models and the prior odds is therefore one. The marginal likelihood is already given in eq. (40).

4.2 Results

To illustrate the Bayesian analysis for AS-t curves we use experimental data for epitaxial growth of Fe on Cu(100) which have been obtained by Glatzel et. al [2]. The experimental data are depicted in the upper panel of fig. 4. The results for the layer-by-layer (FM) growth model are collected in table 1. We begin the discussion of the results with model comparison. In general the transmission coefficient of the 920eV Cu Auger peak in Fe could differ from that of the 705eV Fe peak. Fortunately, the inelastic mean free paths and hence the α-s appear to be almost identical for the two energies[8]. In addition, we can also use $I_{\infty,\mathrm{Fe}} = I_{\infty,\mathrm{Cu}}$ [9]. As discussed in the previous sections, all prior probabilities have been assigned without unidentified parameters. The only prior parameter that has observable influence on the quantitative results is the experimental variation δ in the evaporation time T for a layer; but the data constraints turn out to be so dominant that the precise value for δ is unimportant. The logarithm of the odds ratio $\mathrm{Log}_{10}\left(\frac{P(M_{\mathrm{FM}}|\vec{r}\,\mathcal{I})}{P(M_{\mathrm{SM}}|\vec{r}\,\mathcal{I})}\right)$ changes only slightly when the uncertainty in the evaporation time T varies between $\delta = 1\%$ and $\delta = 40\%$. The true value for δ lies certainly in this regime and according to the first row of table 1 the layer growth model is more probable than the SM growth model by roughly 9 orders of magnitude.

The higher probability for the FM model is corroborated by fig.3, where the experimental data are compared with the theoretical curves for the two models SM (left panel) and FM (right panel). The values for the model parameters are obtained upon maximizing the posterior probabilities (MAP solution). Along with the experimental data we include the estimated error bars within the two models as obtained from eq. (36). Note that the higher probability for the FM model is reflected in considerably smaller error bars. Table 1 also illustrates nicely the balance of likelihood term and Ockham's factor. For $\delta = 1\%$ the prior region is slightly too narrow to describe the experimental variation in the growth times T_i and the likelihood factor, therefore, increases upon increasing δ from 1% to 10%. This is reflected in the decrease of the estimated noise level $\sqrt{\mathcal{E}[\sigma^2|\vec{r}\,M_{\mathrm{FM}}\,\mathcal{I}]}$. The gain in likelihood is, however, not large enough to overcome the loss in Ockham's factor and the net probability decreases slightly. A further increase of δ to 40% does not improve the agreement between fit and data any further (same σ_{est}) and Ockham's razor penalizes the more flexible model and the net probability decreases significantly. This result also tells us that the completion times for the four layers under consideration agree very well and there is therefore no sign of a systematic deviation in the evaporation rates. In hindsight this observation justifies the model equation used for the FM and SM model.

The noise level obtained from $\mathcal{E}[\sigma^2|\vec{r}\,M\,\mathcal{I}]$ for the FM model is $\sigma_{\mathrm{FM}} = 0.006$ while the result for the SM growth model is $\sigma_{\mathrm{SM}} = 0.016$. The model indepen-

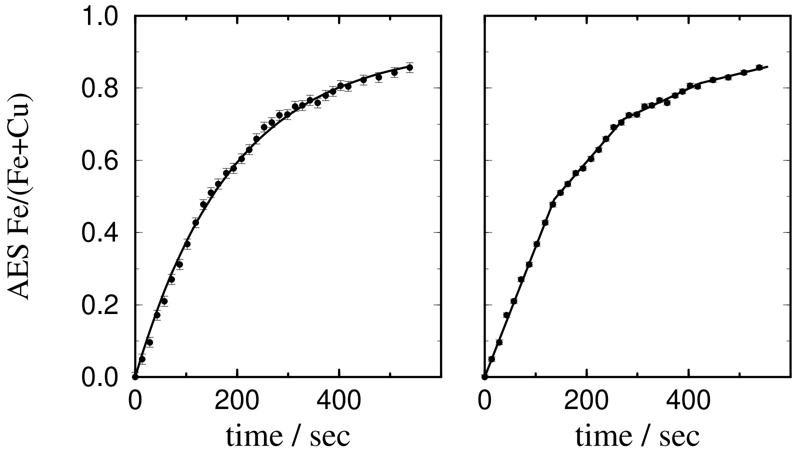

Fig. 3. *Comparison of measured Auger intensities with MAP solution for the two models: simultaneous multilayer growth (left panel) and layer-by-layer growth (right panel)*

dent estimate for σ follows from eq. (26). Since the probability for the layer growth model exceeds overwhelmingly the one for the other model we get $\sigma = \sigma_{\mathrm{FM}} = 0.006$. The lower panel in fig.4 depicts the marginal probabilities $p(t_n = t|rM_{\mathrm{FM}}\mathcal{I})$ for the completion times t_n. Here four breakpoints can be detected for which the numbers are listed in table 1. It is noteworthy that the results for the first two layers (even their variance) depend insignificantly on the prior width δ. Mainly the last breakpoints point t_4, which is obviously only weekly supported by the data, follows the changes in δ. The mean growth time per layer $\overline{T}$ obtained from the first three break points is $137(2)sec$ for $\delta = 10\%$.

Table 1
Results for the layer-by-layer (FM) growth model as function of the prior width δ in the evaporation time. Details are explained in the text.

δ	1%	10%	40%
$\mathrm{Log}_{10}(\frac{P_{\mathrm{FM}}}{P_{\mathrm{SM}}})$	9.2	8.8	7.7
$\sqrt{\mathcal{E}[\sigma^2\|\vec{r}\,M_{\mathrm{FM}}\,\mathcal{I}]}$	0.0064	0.0063	0.0063
t_1/sec	137(2)	135(3)	135(3)
t_2/sec	274(4)	272(6)	273(7)
t_3/sec	411(6)	429(14)	448(23)
t_4/sec	548(8)	575(24)	619(52)
$\overline{T}$/sec	137(1)	137(2)	137(2)
$\mathcal{E}[r_\infty\|\vec{r}\,M_{\mathrm{FM}}\,\mathcal{I}]$	0.89(1)	0.90(1)	0.90(1)
$\mathcal{E}[\alpha\|\vec{r}\,M_{\mathrm{FM}}\,\mathcal{I}]$	0.445(9)	0.462(15)	0.474(19)

The fact that $\mathcal{E}[r_\infty|\vec{r}\, M_{\rm FM}\, \mathcal{I}] = 0.9$ deviates substantially from 1 points out that something is not correct with the assumptions on which the analysis is based. One might argue that this is a consequence of the fact that the model parameters (α, I_∞) for iron and copper are not really identical. This is, however, not true as we will show next. Since the FM model is by far the most probable one we restrict the discussion to this model. Of course, the entire Bayesian analysis could be repeated for the FM model with the increased set of parameters $\{r_\infty, \sigma, \alpha_{s,E_s}, \alpha_{s,E_b}, \rho, \mathcal{T}\}$ but this is not necessary to answer the question under consideration. We can shorten the procedure by just comparing the maximum likelihood parameters. It is noteworthy that the value for the estimated noise level even increases for the more flexible FM model from 0.0063 to 0.0064, which means that the gain in complexity to fit the data is overcome by the loss in degrees of freedom (see eq. (25)). As to the main purpose of this paragraph, the ML estimates for r_∞ are 0.89 in the FM model for both sets of parameters. This value is in perfect agreement with the Bayesian result given in table 1. The difference in transmission coefficients and intensities does therefore not explain the wrong asymptotic value.

Obviously, the discrepancy can be explained if the surface contains not purely iron atoms but copper atoms as well. Several scenarios causing this effect are conceivable. A simple scenario would be a homogeneous admixture of $x = 1 - r_\infty \approx 9\%$ of Cu in all layers. Of course, an incompletely covered Cu substrate (e.g. due to holes in the iron film) would also explain the reduced saturation value. A third alternative would be a fraction $x = (1 - r_\infty)/(1 - \alpha) = 19\%$ of Cu atoms floating up continuously in the top most layer. In summary the present analysis strongly rejects a simultaneous multilayer growth model in favor of a layer-by-layer growth model with some admixture of Cu atoms in the top most layers. In [2] from which the AES data are taken the amount of Fe deposited on the Cu substrate was quantitatively determined by Rutherford-Backscattering. It was found that the first two breaks in the slope in the AS-t curves correspond to 2 and 4 monolayers of Fe, respectively. Thus a layer-by-layer growth mode has been claimed. While – from a compositional point of view – such a "bilayer growth model" is well established by other techniques [10–15] in the initial stages of growth up to 2 monolayers (i.e. up to the first breakpoint in the AS-t curve in fig.3), there is no support for such a growth mode at coverages beyond 2 monolayers. Thus it has been suggested [14] that the first three breakpoints in slope of the AS-t curves may have a structural rather than morphological origin, since these breakpoints coincide with the occurrence of (4×1), (5×1) and (2×1) overlayer structures at D =2,4 and 6 monolayers [15]. However, if this is true it is rather surprising that the AS-t data of fig.3 follow almost perfectly the functional form (eq. (8),eq. (9)) expected for a layer-by-layer (or bilayer-by-bilayer) growth mode. We do not have a definite explanation for this behavior. However, we would like to point out that – in contrast to usual film deposition in a single run – AS-t curves (as the one in fig.3) are often obtained by several subsequent

deposition events, interrupted by the AES-analysis. In consideration of the extreme sensitivity of the Fe/Cu(001) system to deposition parameters this unusual deposition mode, which lowers the effective deposition rate by some two orders of magnitude, could well produce films with different morphology and enhance diffusion of substrate atoms to the very surface. Furthermore, the prolonged exposure to the electron beam for AES-analysis may also alter the film growth. At this stage we want to point out again that the purpose of the present paper is to demonstrate how, by strictly applying the rules of probability theory, different models (even with different complexity and degrees of freedom) can be compared with each other and how model parameters can be consistently obtained. However, the choice of models to be compared as well as their physical meaning is beyond the scope of probability theory and belongs to the responsibility of the scientists. So, even if a model perfectly fits the experimental data, as in the present case of the FM model for the AS-t curves, the model needs not to be the true one. It is conceivable that the mathematical model can also have different physical origin or that other physical models with different model equation even have a higher probability.
Finally, we will briefly discuss the values for the transmission coefficient α as obtained from the FM and SM models. In the FM model $\alpha = 0.462(15)$ which is in good agreement to the values derived from other AS-t experiments[16,17]. One should be able to determine the ratio λ/d, that is the inelastic mean free path (IMFP) λ in units of film thickness d, from the transmission coefficient. The relation between transmission coefficient and IMFP is given in eq. (1). With the values from table 1 we conclude $\lambda/d = 1.8(1)$. Assuming bi-layer growth[2] (d = 2ML) we find a value for the IMFP of $\lambda = 3.5ML$ which deviates strongly from the universal curve reported by Seah and Dench [8], which yields 4.6ML for 705eV and 5.2ML for 920eV electrons. The α value and the reported IMFP data are only in conformity if one uses erroneously the linear approximation of eq. (1) as was done in Ref.[2]. The discrepancy from the universal IMFP curve is however not that worrying since there is a considerable scatter of data around the universal curve. In addition, more results for the IMFP of Auger electrons show that even the data for iron alone scatter over a wide range[18]. So we are inclined to believe that the deviation from the universal curve is genuine.
In the case of simultaneous multilayer growth transmission coefficient α and completion time T cannot be determined separately. We obtain $(1-\alpha)T = 0.00524(12)$. For a qualitative discussion we use $\overline{T} = 137(2)$ of the FM model and obtain $\alpha = 0.28(1)$ which is considerably smaller than in the FM model and would even predict $\lambda/d = 1.06(4)$.

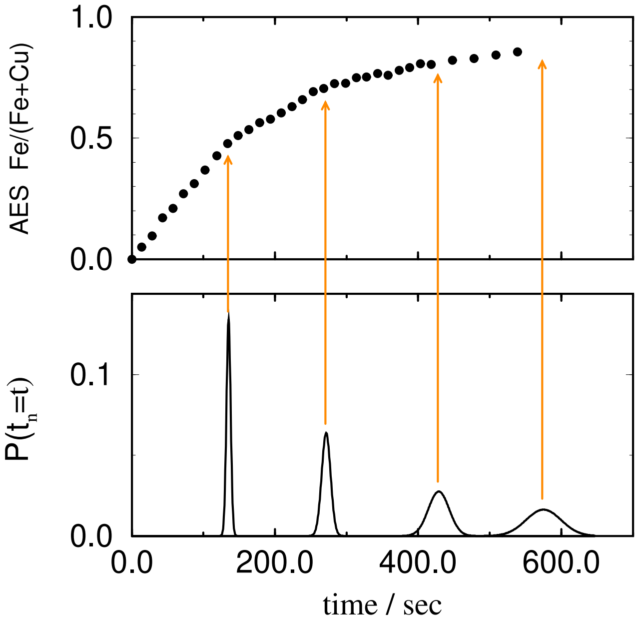

Fig. 4. *Experimental data for the normalized Auger iron signal, Fe(705 eV)/(Fe(705eV)+Cu(920eV)) versus evaporation time (upper panel). Marginal probability for the completion times t_n in the FM model versus time (solid lines) for* $\delta = 10\%$ *(lower panel).*

4.3 Summary

We have illustrated how the rules of probability theory can be used to assess growth models, which may even have different complexities with different numbers of model parameters, and how the model parameters can be determined without resorting to any ad-hoc methods. We have seen that among the two models we have chosen to demonstrate the theory, the layer-by-layer growth process is by far more likely to describe the AS-t curves than the simultaneous multilayer growth model. The consistent Bayesian analysis also revealed a contradiction to the assumption made in [2] that the surface layers are pure iron. The Bayesian analysis rather prompts us to assume that all layers are homogeneously interspersed with copper. It is needless to say that it is straight forward to extend the analysis to other growth models and to include different types of experimental data besides AS-t data, like LEED[19] where models of different complexity have to be compared.

References

[1] C. Argile and G. Rhead, Surf. Sci. Rep. 10 (1989) 277.

[2] H. Glatzel, Th. Fauster, B. Scherzer and V. Dose, Surf. Sci. 254 (1991) 58.

[3] Low Energy Electrons and Surface Chemistry, Eds. G. Ertl and J. Küppers (Verlag Chemie, Weinheim, 1974).

[4] M. Tribus, Rational Descriptions, Decisions and Design (Pergamon, New York, 1969).

[5] E. Jaynes, in: E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics, Eds. R. Rosenkrantz (Reidel, Dordrecht, 1983), p. 114.

[6] Bayesian Spectrum Analysis and Parameter Estimation, Eds. G. Bretthorst (Springer Press, Berlin, Heidelberg, 1988).

[7] J. Hammersley and D. Handscomb, Monte Carlo Methods (Methuen, London, 1964).

[8] M. Seah and W. Dench, Journal Surf. Interface Anal. 1 (1979) 2.

[9] L. D. et al., Handbook of Auger Electron Spectroscopy, Physical Electronics (Eden Praire, (2nd edition 1976), ADDRESS, 1972).

[10] D. Steigerwald, I. Jacob and J. W.F. Egelhoff, Surf. Sci. 202 (1988) 472.

[11] T. Detzel, N. Memmel and T. Fauster, Surf. Sci. 293 (1993) 227.

[12] K. Johnson, D. Chambliss, R. Wilson and S. Chiang, J. Vac. Sci. A11 (1993) 1654.

[13] A. Brodde and H. Neddermeyer, Surf. Sci. 287/288 (1993) 988.

[14] J. Thomassen, B. Feldmann and M. Wüttig, Surf. Sci. 264 (1992) 406.

[15] K. Heinz, P. Bayer and S. Müller, Surface Reviews and Letters 2 (1995) 89.

[16] W. Daum, C. Stuhlmann and H. Ibach, Phys.Rev.Lett. 60 (1988) 2741.

[17] M. Stampanoni, A. Vaterlaus, M. Aeschlimann, F. Meier and D. Pescia, J.Appl.Phys. 64 (1988) 5321.

[18] M. Onellion, M. Thompson, J. Erskine, C. Duke and A. Paton, Surf.Sci. 179 (1987) 219.

[19] K. Heinz, S. Müller and P. Bayer, Surface Science 337 (1995) 215.