Acta Materialia 125 (2017) 145–155

Contents lists available at ScienceDirect

Acta Materialia

journal homepage: www.elsevier.com/locate/actamat

Full length article

Efficient sampling in materials simulation - Exploring the parameter space of grain boundaries

Holger Dette [a], Josua Gösmann [a], Christian Greiff [a], Rebecca Janisch [b, *]

[a] *Department of Mathematics, Institute of Statistics, Ruhr-Universität Bochum, 44780 Bochum, Germany*
[b] *Interdisciplinary Centre for Advanced Materials Simulation (ICAMS), Ruhr-Universität Bochum, 44780 Bochum, Germany*

ARTICLE INFO

Article history:
Received 28 July 2016
Received in revised form
19 October 2016
Accepted 17 November 2016
Available online 8 December 2016

Keywords:
Sequential sampling
Design of experiment
Grain boundary energy
Atomistic simulations
High-throughput methods

ABSTRACT

In the framework of materials design there is the demand for extensive databases of specific materials properties. In this work we suggest an improved strategy for creating future databases, especially for extrinsic properties that depend on several material parameters. As an example we choose the energy of grain boundaries as a function of their geometric degrees of freedom. The construction of many existing databases of grain boundary energies in face-centred and body centred cubic metals relied on the a-priori knowledge of the location of important cusps and maxima in the five-dimensional energy landscape, and on an as-densely-as-possible sampling strategy. We introduce two methods to improve the current state of the art. Based on an existing energy model the location and number of the energy minima along which the hierarchical sampling takes place is predicted from existing data points without any a-priori knowledge, using a predictor function. Furthermore we show that in many cases it is more efficient to use a sequential sampling in a “design of experiment” scheme, rather than sampling all observations homogeneously in one batch. This sequential design exhibits a smaller error than the simultaneous one, and thus can provide the same accuracy with fewer data points. The new strategy should be particularly beneficial in the exploration of grain boundary energies in new alloys and/or non-cubic structures.

1. Introduction

Increasing computational power contributes to the creation and extension of huge materials databases containing a variety of material properties. Such data collections can form the basis for material design and discovery, and thus promise great savings in development time and costs. The materials project [1] or the NOMAD repository [2] which together feature roughly 3.5$\cdot^{6}$ entries are impressive examples of such property collections. They can be used to search for correlations in property maps, or as input for material models.

High-throughput numerical simulations as well as experiments play a key role in determining materials properties, because they allow a systematic variation of individual parameters, such as e.g. composition, and the coverage of a broad range of these parameters. Nevertheless, the majority of properties that are available today are so-called intrinsic properties like phase stability, stiffness, or band gaps. Extrinsic properties which depend on additional structural parameters, like interface distribution functions, energies or mobilities, are available, too (see e.g. Ref. [3]), but much less frequently. As a matter of fact, because they depend on several variables at the same time, such properties form multidimensional databases of their own for one particular material.

For such cases the paper at hand shall promote an efficient sampling of the variable space that is based on design of experiment principles. As a suitable example we pick grain boundary energies, which depend on five geometric degrees of freedom.

The evolution of a microstructure of a metallic or multiphase material during solidification or deformation depends to a large extend on the energy of the interfaces in the microstructure. So do the resulting macroscopic properties, the deformability, thermal stability, and strength. Substantial scientific effort is put into describing the interactions of individual defects that form the microstructure within and across different material length scales to model their influence on the macroscopic properties. Again, numerical simulations are a valuable tool in this respect, because they allow the isolation of specific effects on different length or time

[*] Corresponding author.
E-mail address: rebecca.janisch@rub.de (R. Janisch).

http://dx.doi.org/10.1016/j.actamat.2016.11.044
1359-6454/

scales, as well as the abovementioned systematic variation of parameters that enter any material model.

In this paper we focus on grain boundaries, i.e. homophase interfaces, which separate two regions of undistorted crystal of the same crystal structure and chemistry that are misoriented with respect to each other. The space of macroscopic geometric parameters that define a grain boundary structure is five-dimensional - three degrees of freedom for the misorientation of the grains, two for the interface inclination. It is not per se clear which sampling strategy is the best for this five dimensional space. An intuitive approach might be to sample the parameter space homogeneously as finely as possible, so that afterwards an interpolation between the points is possible without knowing the underlying function that connects them. Alternatively, one can rely on a so-called structure model, which helps to identify the slowly and quickly varying regions in the energy function. Ideally, it already provides a closed form expression for the energy as a function of the structural parameters. This latter approach has been recently used by Bulatov et al. [4] to find a universal function for the crystallography class of face-centred cubic (fcc) metals. It reduces the amount of required data points, but uses a-priori knowledge of the topography of the multi-dimensional space of macroscopic parameters defining the grain boundary geometry. In particular it requires the specification of the number and the location of the cusps and maxima in the energy landscape. The 3D rotational subspaces that represent energy minima can indeed be predicted based on crystal symmetry. To a large extend, especially for fcc metals, the positions of the cusps in the one-dimensional subspaces of energy as a function of misorientation for twist and tilt grain boundaries can be predicted fairly well based on a random grain-boundary model [5], or a lattice-matching approach [6]. However, they also depend on the details of the atomistic structure and hence on the nature of the interatomic bonding, which makes a prediction more difficult already for bcc metals [6], and especially for systems with directional bonds, such as diamond or silicon, no straight forward relationship exists [7].

In this paper we show that the method proposed by Bulatov et al. [4] can be significantly improved in at least two directions. First we develop an estimate, which does not require any knowledge of the number and the location of cusps and maxima of the energy function, but determines these parameters from the data. Secondly, we address the problem of efficient sampling in such experiments. In contrast to previous work, which considers curve fits from all available data (within a subspace) we propose a sequential approach. A first part of the sample is used to obtain a prediction for the energy function. Subsequent sampling is performed sequentially, such that the new experiment is conducted at sampling points at which the uncertainty of the prediction is maximal. Additionally, the prediction is continuously updated using the new information of the previous experiments. In many cases we can thus minimize the number of points which are necessary to give a reliable description of the parameter space, especially cusps and maxima in the energy landscape. Thus, the important parameters of the energy function do not have to be known a priori, and still the computational effort remains manageable.

The remaining part of this paper is organized as follows. In Section 2 we give a brief description of the geometry and structure of the grain boundary and describe the existing data bases of grain boundary energies. We also describe the structural model introduced in Ref. [4] which will be further developed in this paper. Section 3 introduces the new statistical methodology, and the advantages of the sequential sampling approach are illustrated in Section 4.

2. Grain boundary structure and energy

The geometry and structure of grain boundaries are determined by the five macroscopic and three microscopic degrees of freedom (DOF), i.e. by.

- The misorientation of the two grains, e.g. defined by rotation axis $\hat{\boldsymbol{\rho}}$ and angle x, or a rotation matrix $\mathbf{A}$, which rotates grain 1 into grain 2. This rotation determines three macroscopic DOF.
- The inclination of the grain boundary plane, defined by its normal vector $\hat{\mathbf{n}}$, that determines the remaining two macroscopic DOF.
- A translation $\mathbf{t}$ of the grains parallel and perpendicular to the interface plane, which fixes the three microscopic DOF.

Traditionally, grain boundaries for which the plane normal vector $\hat{\mathbf{n}}$ is perpendicular to the misorientation axis $\hat{\boldsymbol{\rho}}$ are called tilt grain boundaries, those with $\hat{\mathbf{n}}$ parallel to $\hat{\boldsymbol{\rho}}$ twist grain boundaries.

Several other representations of the grain boundary geometry are possible, e.g. a representation by two rotation matrices that describe the orientation of each grain with respect to a reference frame, in which one axis is perpendicular to the interface [8]. It depends on the application which representation is the most convenient one. For instance, a representation by two rotation matrices is useful to construct grain boundary structures as input for atomistic simulations, as done e.g. in Ref. [9]. However, to evaluate afterwards the energies of grain boundaries as a function of their degrees of freedom, as done e.g. by Wolf [10] for face centred cubic metals, a representation in terms of rotation axis and angle is more descriptive.

Some grain misorientations lead to a periodic superstructure across the grain boundary, a so-called “coincidence site lattice” (CSL). CSL-based grain boundaries represent low-energy structures that are especially important when trying to capture the material properties. Hence, besides being characterised by their misorientation they are usually described by an additional parameter Σ, which is the ratio of the CSL unit cell volume to that of the original crystal lattice unit cell. In other words it is a measure for the periodicity of the CSL lattice.

In principle the energies of all possible grain boundary structures in a microstructure could be determined from atomistic simulations. The only limiting factors are the computational demand and the availability of a reliable description of the interatomic interactions. Many atomistic studies of lower-dimensional subspaces can be found in the literature, which focus e.g. on symmetrical tilt grain boundaries (e.g. Refs. [10–13]), on the two-dimensional subspace of grain boundary plane inclinations [14], or on examples of general [15], and heterophase [16] boundaries. In addition, few but expansive databases which represent the full five-dimensional space have been created that include several thousand structures [9,17,18].

Up to date there exist two major collections of grain boundary energies which were obtained from atomistic simulations, the database of Kim et al. [17] for body-centred cubic (bcc) iron, and the one of Olmsted and co-workers [9] for face-centred cubic (fcc) aluminium and nickel, and a more recent version also for bcc iron [18]. The structure of the databases reflects two different approaches: Kim et al. sample the 5D parameter space as densely as possible in a homogeneous fashion, i.e. for arbitrary misorientations, to provide sufficient data for a numerical interpolation scheme. Details are given in Section 2.1. In the approach chosen in Ref. [9] the emphasis is put on the CSL-based grain boundaries by sampling the subspace of special misorientations around high-symmetry axes in the cubic lattice. Thus, in the resulting database the minimum energy regions are represented well, but to interpolate safely in the “non-special regions” a model for the energy as a function of the geometric degrees of freedom is needed. This has been provided by Ref. [4], as described below in Section 2.2.

2.1. Homogeneous sampling and numerical interpolation

In the work of Kim et al. the misorientation between two bcc grains is represented by the three Euler angles between a reference coordinate system (grain 1) and a sphere inside (grain 2). The misorientation range is then sampled by dividing each Euler angle into nine equal intervals of 10° ($0 \le \phi_1 \le 90^\circ, 0 \le \Phi \le 90^\circ, 0 \le \phi_2 \le 90^\circ$), leading to 729 misorientations. For a smooth energy function of the misorientation, this might have been a sufficiently dense sampling. However, this homogeneous scheme misses the majority of the special misorientations, which leads to CSL boundaries and to pronounced cusps in the energy function. This is a striking example of the fact that even a strictly systematic parametric study can miss the most important features of a target function, and it motivates again our improved scheme which is able to identify such points. Kim et al. were aware of the missed features in the topology of the parameter space and thus additional 26 CSL boundaries with $\Sigma \le 33$ were constructed separately. For each misorientation, CSL-type or general, 91 different inclinations of the grain boundary plane were selected, representing equally spaced points on the sphere. Fig. 1 shows the anisotropy of the energy as a function of plane inclination for a Σ 9 misorientation extracted from the database of Kim et al. [17].

This anisotropy is an important factor for the evolution of the microstructure e.g. during grain growth. With their dense sampling Kim et al. [17] were able to provide enough data points for interpolation within a phase-field model in the simulation of grain growth in polycrystalline iron of [19]. This demonstrates on the one hand the usefulness of such databases, but on the other hand also their limitations. The next step towards a more realistic material system would be to include a second chemical (alloying) element in the grain boundary structure. In principle, grain boundary energies in a multi-component system could be obtained by a similar procedure as used for the database for pure Fe grain boundaries. However, a good description of the five-dimensional space of geometric degrees of freedom already required including 66.339 structures, each obtained as the result of a multi-step optimization procedure. Every alloying element would add another compositional degree of freedom to the five geometric ones, and thus increase the number of structures by orders of magnitude.

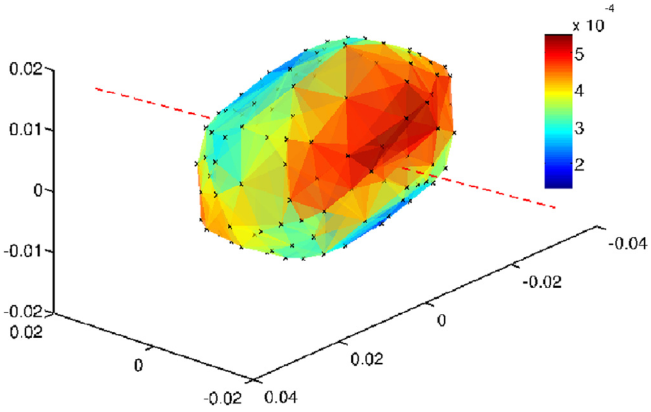


Fig. 1. Energy as a function of plane inclination for a Σ 9 misorientation (data from Ref. [17]).

2.2. Fitting in subspaces and hierarchical interpolation

An alternative strategy was chosen by Bulatov et al. [4], who fitted the parameters of a five-dimensional energy function to the database of Olmsted. Starting point for the construction of this function is the dislocation model for grain boundaries [20]. It describes a grain boundary as a two-dimensional array of lattice dislocations. The energy of this array is the sum of the energies of the individual dislocations, which can be calculated e.g. according to elasticity theory (with the core energy being a free parameter). Strictly speaking this picture brakes down if the density of the dislocations in the grain boundary becomes so large that the dislocation cores start to overlap, i.e. for misorientation angles around 10°. Wolf showed [21], however, that the expression of Read and Shockley [20] can be extended piecewise to the complete range of misorientation angles between 0° and 180°, as shown schematically in Fig. 2. The resulting model, though fully empirical, can be justified by the existence of grain boundary dislocations with Burgers vectors that can assume (continuous and) much smaller values than the one of lattice dislocations and thus exist in array much more dense.

Wolf defined such functions for the one-dimensional subspaces of symmetrical tilt and twist grain boundaries, i.e. $E(x)$ for fixed rotation axes, of the type

$$E(x) = \sin(\pi/2\,x)(E_c - E_s \log(\sin(\pi/2\,x))), \tag{1}$$

in which x varies between 0 and 1. E_s is the strain energy of the dislocation, which can be approximated by $E_s = Gb/2\pi$, with the material parameters shear modulus G and Burgers vector b [10]. E_c is the dislocation core energy, which was used as adjustable parameter to fit the model to grain boundary energies determined via atomistic simulations.

In Ref. [4] Equation (1) is re-phrased by splitting off a dimensionless part

$$f_{\mathrm{RSW}}(x) = \sin(\pi/2\,x)(1 - a \log(\sin(\pi/2\,x))) \tag{2}$$

for each segment of the one dimensional functions, which is then

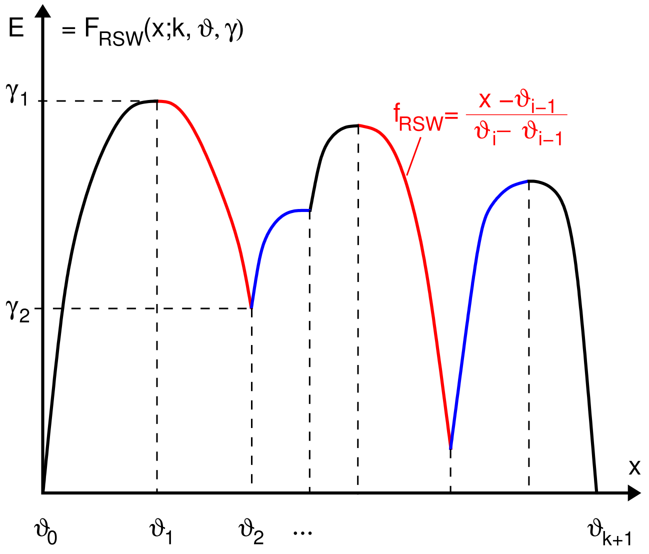


Fig. 2. Schematic of the 1D subspace of energy as a function of misorientation (of twist or symmetrical tilt grain boundaries). The function is expressed piecewise between two change points ϑ_{i-1} and ϑ_i by the Read-Shockley model (1). In this work we explore ways to determine the number of necessary segments f_{RSW} efficiently.

multiplied with the energy of the maximum in the respective segment and (scaled and) shifted to the corresponding interval. For a precise definition see Formula (4) in the following section. With respect to the formulation of Wolf, a could be regarded as the ratio of strain to core energy, but in fact it is a fit parameter.

Reference energies for several hundred grain boundaries in different fcc metals are e.g. available from the work of Olmsted et al. [9], who carried out atomistic simulations using EAM type potentials. As mentioned, the sampling of the 5D space is not homogeneous in this case, but is restricted to misorientations around high-symmetry axes of cubic crystals that lead to coincidence site lattices. This way 388 grain boundaries with 72 different misorientations around the $\langle 111\rangle, \langle 110\rangle$, and $\langle 100\rangle$ axes of the fcc lattice were constructed. As pointed out in Ref. [4], this approach is reasonable if one assumes that the topography of the GB energy function is defined by cusps or "grofs" of order k [4], special subsets of the 5-dimensional space of macroscopics DOF, for which the energy is minimal with respect to variations locally orthogonal to the set. In other words, the energy is smooth along the valley of the $(N-k)$ dimensional subset, but forms a cusp in the complementary k-subspace. The connectivity of grofs is defined by the crystal symmetries.

Accordingly, Bulatov et al. [4] started by fitting the one dimensional scaffolding sets of symmetrical tilt and twist boundaries with misorientation axes $\langle 100\rangle, \langle 111\rangle, \langle 110\rangle$. In the next step the 2D subset including asymmetric tilt grain boundaries was obtained by interpolating between these six 1D sets, and the 3D scaffolding sets of mixed grain boundaries (neither pure twist nor tilt) in turn by interpolating between the 2D subsets. In a last step the interpolation between the 3D scaffolding sets in 5D space was done by approximating the energy of an arbitrary grain boundary as a weighted average over the energies of all best-matching boundaries found within the three 3D scaffolding subsets. This is equal to assuming a rather flat 2D subspace of energy vs. plane inclination. This assumption of a slowly varying function for the energy as a function of plane inclination is a rather strong one, as can be seen from the results of [17], see e.g. Fig. 1. Apart from this, a comparison of four different fcc metals shows that the fitting function is rather universal, since 27 of the 42 dimensionless parameters turned out to be more or less the same for all investigated materials.

To summarize, the importance of the CSL boundaries motivates the "special" sampling approach [9] followed by a hierarchical interpolation scheme [4], starting with the subspaces of grain boundaries with misorientations around high symmetry axes, rather than using a homogeneous sampling scheme over the whole 5D space. However, since the 2D subspace of plane inclinations is by no means flat (especially not in a multi-component system), a sampling of this subspace, which is more or less neglected in the data of [9], has to be included, as done in Ref. [17] or studied separately as in Ref. [14].

3. Improved statistical inference of the grain boundary energy function

As mentioned, an important feature in the description of the topography of the grain boundary energy subspaces are the energy-cusps. Olmsted et al. [9]. and Bulatov et al. [4] investigate the grain boundary energy function for fcc models with a-priori knowledge of the number and location of the cusps. The Read-Shockley-Wolf (RSW) function proposed in Ref. [21] is used to model the energy values of grain boundaries in these subspaces [4]. In this work we introduce an alternative procedure, which still assumes the validity of the RSW function, but improves the current state of the art in at least two directions.

- First - in contrast to [4] - we determine the location and number of the cusps and maxima for the RSW model directly from the available data. Recall from Section 2 that the energy function consists of piecewise models of the form (2) which are scaled with the maximum grain boundary energy in the respective interval and connected at change points as displayed in Fig. 2. We propose to determine the number and the location (more precisely the left and right end-point) of these intervals from the available data. This gives a fit of RSW model without using any a-priori information.
- Secondly, we propose an efficient way of locating profitable measurement spots. This is achieved by a sequential sampling scheme. A first part of the sample is used for a prediction of the energy function. Subsequent sampling is then performed sequentially, such that the new experiment is conducted at sampling points, where the uncertainty of the prediction for the energy function is maximal. Additionally, the prediction is updated after each experiment. Thus we address the problem of the high computational costs in these numerical experiments, and minimize the number of points which are necessary to give a reliable description of the topography of these subspaces.

3.1. Estimating a RSW-Function

In this section we describe a general approach to estimate all parameters in a piecewise RSW model. As introduced in Section 2.2, the RSW-segment function is given by

$$f_{\mathrm{RSW}}(x) = \sin(\pi/2\,x)(1 - a\log(\sin(\pi/2\,x))), \tag{3}$$

where x varies between 0 and 1. Now let k denote the (unknown) number and $0 = \vartheta_0 < \vartheta_1 < \ldots < \vartheta_k < \vartheta_{k+1} = \vartheta_{\max}$ the (unknown) boundary points of the intervals in which different RSW segments are used. $\vartheta_{\max}$ depends on the rotational symmetry of the misorientation axis in question. For instance, the $<110>$ directions in a cubic lattice have two-fold rotational symmetry, thus $\vartheta_{\max} = 180^\circ$, while the $<100>$ axes have four-fold rotational symmetry and $\vartheta_{\max} = 90^\circ$. The boundary points may correspond to cusps, local maxima, or shoulders of the energy function, as schematically shown in Fig. 2. Throughout this paper these points will be called *change points*. Moreover, we denote by $\gamma_0 = 0$ the energy at point $\vartheta_0 = 0$. We define a piecewise RSW function F_{RSW} of the order k with energy levels $\gamma_1, \ldots, \gamma_k$ by

$$\begin{aligned} F_{\mathrm{RSW}}(x,\vartheta,\gamma,k) = {} & \sum_{i=1}^{k+1} I_{[\vartheta_{i-1},\vartheta_i]}(x) \\ & \times \Bigg[I_{[-\infty,0)}(\gamma_i - \gamma_{i-1}) \cdot \left(\gamma_i + (\gamma_{i-1} - \gamma_i) \cdot f_{\mathrm{RSW}}\left(\frac{\vartheta_i - x}{\vartheta_i - \vartheta_{i-1}} \right) \right) \\ & + I_{[0,\infty)}(\gamma_i - \gamma_{i-1}) \cdot \left(\gamma_{i-1} + (\gamma_i - \gamma_{i-1}) \cdot f_{\mathrm{RSW}}\left(\frac{x - \vartheta_{i-1}}{\vartheta_i - \vartheta_{i-1}} \right) \right) \Bigg] \end{aligned} \tag{4}$$

where

$$I_{[a,b]}(\kappa) = \begin{cases} 1 & \text{if } \kappa \in [a,b] \\ 0 & \text{otherwise.} \end{cases}$$

$I_{[a,b]}(x)$ denotes the indicator function that assigns the different RSW-segment functions f_{RSW} to their intervals $[\vartheta_{i-1}, \vartheta_i]$ $(\kappa = x)$, respectively determines the sign of the segment $(\kappa = \gamma_i - \gamma_{i-1})$.

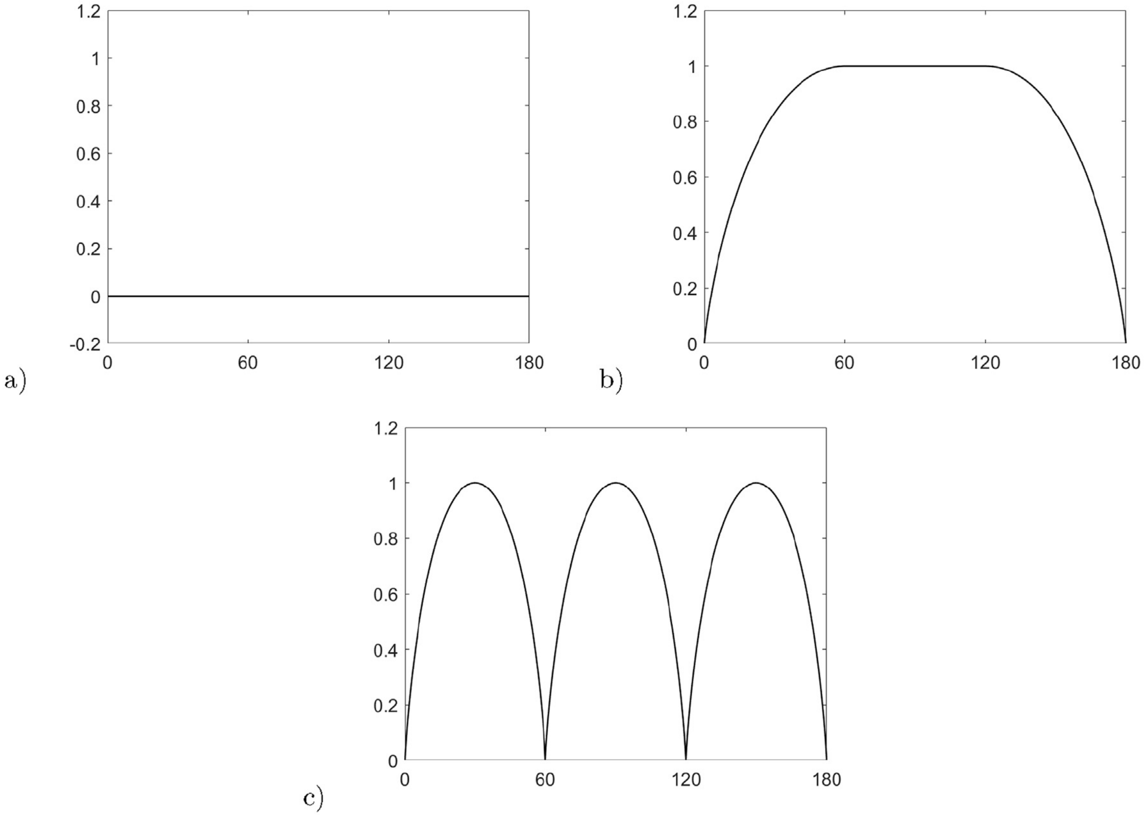


Fig. 3. Examples of RSW energy functions $F_{\mathrm{RSW}}(x)$ for different parameters k, ϑ and γ. A choice of such functions is also used as starting points in the minimization of (5). a) $k = 0$ and $\gamma = 0$, b) $k = 2$, $\vartheta_1 = 60^\circ$, $\vartheta_2 = 120^\circ$ and $\gamma_1 = \gamma_2 = 1$ c) $k = 5$, $\vartheta_1 = 30^\circ$, $\vartheta_2 = 60^\circ$, $\vartheta_3 = 90^\circ$, $\vartheta_4 = 120^\circ$, $\vartheta_5 = 150^\circ$, $\gamma_1 = \gamma_3 = \gamma_5 = 1$ and $\gamma_2 = \gamma_4 = \gamma_6 = 0$.

Note that in Equation (4) ϑ and γ are short notations for the complete sets of angles and energies, $\vartheta = (\vartheta_1, \ldots, \vartheta_k)$, and $\gamma = (\gamma_1, \ldots, \gamma_{k+1})$. Each specification (k,ϑ,γ) corresponds to a different energy function. For example, if $k = 0$ and $\gamma_0 = 0$, the energy as a function of misorientation is constant, as shown in Fig. 3a). The combination $k = 2$, $\vartheta_1 = 60^\circ$, $\vartheta_2 = 120^\circ$ and $\gamma_1 = \gamma_2 = 1$ gives the function shown in Fig. 3b), and $k = 5$, $\vartheta_1 = 30^\circ$, $\vartheta_2 = 60^\circ$, $\vartheta_3 = 90^\circ$, $\vartheta_4 = 120^\circ$, $\vartheta_5 = 150^\circ$, $\gamma_1 = \gamma_3 = \gamma_5 = 1$ and $\gamma_2 = \gamma_4 = \gamma_6 = 0$ would yield the function displayed in Fig. 3c). In the following sections we discuss the problem of fitting a piecewise RSW function to data $\{(x_i, E_i)|\ i = 1, \ldots, n\}$, where the pair (x_i, E_i) denotes the ith observation of a grain boundary with misorientation angle x_i and corresponding simulated energy E_i. This means that we have to determine the number of change points k, locations $\vartheta = (\vartheta_1, \ldots, \vartheta_k)$, and heights $\gamma = (\gamma_1, \ldots, \gamma_{k+1})$ of the (unknown) RSW-function F_{RSW} from the data.

3.1.1. *Estimating a RSW-Function with fixed number of change points*

In Ref. [4] it is assumed that the number k and the location of (some of) the change points $\vartheta_1, \ldots, \vartheta_k$ are known, and the values $\gamma_1, \ldots, \gamma_k$ are determined from the grain boundary energies in the database of [9]. In this section we take the first step to a more flexible approach by assuming that only the number of change points is known (for example from symmetry properties of the crystal) and the other parameters have to be determined from the available data. The problem of additionally determining the number k from the data will then be addressed in the next section.

The estimation of the parameters $\vartheta_1, \ldots, \vartheta_k$ and $\gamma_1, \ldots, \gamma_{k+1}$ can then be done by a constrained non-linear least squares approach minimizing the sum of squared errors,

$$\mathrm{SSQ}(\vartheta, \gamma) = \frac{1}{n} \sum_{j=1}^{n} \left[E_j - F_{\mathrm{RSW}}(x_j, \vartheta, \gamma, k) \right]^2 , \tag{5}$$

with respect to (ϑ, γ) under the constraints $\vartheta_{i-1} \le \vartheta_i$ $(i = 1, \ldots, k + 1)$. The minimization is realized by the non-linear optimization algorithm SQP (see e.g. Ref. [22]), which is implemented e.g. in the optimization routine ”fmincon()” of ©MATLAB. Note that the corresponding optimization problem is not convex. Consequently, the quality of these algorithms depends sensitively on the set of starting points. To ensure reliable results every minimization was performed with three different starting sets. Possible examples of such functions are displayed in Fig. 3) and we used the best of the three obtained minimizers. These values are denoted by $\widehat{\vartheta}$ and $\widehat{\gamma}$. In the statistical literature $\widehat{\vartheta}$ and $\widehat{\gamma}$ are usually called *(constrained) least squares estimates* as the sum of squares is minimized under the constraints $\vartheta_{i-1} \le \vartheta_i$ $(i = 1, \ldots, k + 1)$.

If one evaluates the RSW function F_{RSW} for a given angle x and the estimates $\widehat{\vartheta}$, $\widehat{\gamma}$ one obtains a prediction $\mathrm{pred}(x) := F_{\mathrm{RSW}}(x, \widehat{\vartheta}, \widehat{\gamma}, k)$ of the unknown RSW function $F_{\mathrm{RSW}}(x, \vartheta, \gamma, k)$ at the point x. As $\widehat{\vartheta}$, $\widehat{\gamma}$ contain a statistical estimation error it is of particular interest to specify the uncertainty of this prediction. For this purpose we require some mathematical assumptions and assume that the statistical model is given by

$$E_i = F_{\mathrm{RSW}}(x_i, \vartheta, \gamma, k) + \varepsilon_i \ , i = 1, \ldots, n \ ; \tag{6}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are independent, normal distributed random variables with mean 0 and the variance σ^2. In (6) the variable ε_i denotes a random error which is introduced to model deviations from the functional relation $y = F_{\mathrm{RSW}}(x, \vartheta, \gamma, k)$, which ideally describes the relation between energy y and angle x through the RSW function. It then follows from standard results of mathematical statistics (see for example [23]) that the variance of the prediction $\text{pred}(x)$, for the energy at point x is approximately given by

$$\text{Var}(\text{pred}(x)) \approx \widehat{V}^2(x) := SSQ\left(\widehat{\vartheta},\widehat{\gamma}\right)\cdot m\left(x,\widehat{\vartheta},\widehat{\gamma}\right)\left(\sum\left(m\left(x_i,\widehat{\vartheta},\widehat{\gamma}\right)^T m\left(x_i,\widehat{\vartheta},\widehat{\gamma}\right)\right)\right)^{-1} m\left(x,\widehat{\vartheta},\widehat{\gamma}\right)^T, \tag{7}$$

where

$$m\left(x,\widehat{\vartheta},\widehat{\gamma}\right)=\left(\frac{\partial}{\partial\vartheta}F_{\text{RSW}}(x,\vartheta,\gamma,k)\Big|_{\vartheta=\widehat{\vartheta},\gamma=\widehat{\gamma}},\frac{\partial}{\partial\gamma}F_{\text{RSW}}(x,\vartheta,\gamma,k)\Big|_{\vartheta=\widehat{\vartheta},\gamma=\widehat{\gamma}}\right)$$

denotes the gradient of the piecewise RSW function evaluated at the estimates $\widehat{\vartheta}$ and $\widehat{\gamma}$. This means that the standard error of the prediction is approximately given by $\widehat{V}(x)$. Obviously we are interested in predictions with a *small* standard error for all angles of interest.

3.1.2. Determining the number $\widehat{k}$ of change points

The estimation method discussed in Section 3.1.1 requires the specification of the number of change points k, which is often not available in simulations of grain boundary energies. In this section we use an *information criterion* from mathematical statistics to determine the unknown order k of the function in Equation (4) from the available data, i.e. we determine the number of segments f_{RSW} from which F_{RSW} is constructed. Roughly speaking, such a criterion is used to balance the complexity and the quality of a parametric fit of the energy function. For example, in an extreme case, one would just interpolate the data using a model with a large number of parameters. Obviously, such a model would be useless, because the number of parameters coincides with the number of data points. On the other hand a model with too few parameters might not give a reasonable fit to the observed data. For a detailed discussion of statistical model selection the reader is referred to [24].

Consider a piecewise RSW function $F_{\text{RSW}}(x,\vartheta,\gamma,k)$ of a fixed order k and let $(\widehat{\vartheta}^{(k)},\widehat{\gamma}^{(k)})$ denote the (constrained) least squares estimates determined by the procedure described in Section 3.1.1. The *Bayesian Information Criterion* (BIC) is given by

$$\text{BIC}(k) := n\log\left(\text{SSQ}\left(\widehat{\vartheta}^{(k)},\widehat{\gamma}^{(k)}\right)\right)+k\log(n), \tag{8}$$

where SSQ is the sum of squares defined in Equation (5) and n is the number of data points. The order of the RSW function is now estimated by the value $\widehat{k}$ for which $\text{BIC}(k)$ is minimal, that is

$$\widehat{k} = \text{argmin}_k\text{BIC}(k).$$

If the minimum is attained at two (or more) values, say $\widehat{k}_1, \widehat{k}_2$ we choose the smaller value as estimate of the order. The estimates of the parameters are the values $(\widehat{\vartheta}^{(\widehat{k})},\widehat{\gamma}^{(\widehat{k})})$ corresponding to the optimal value $\widehat{k}$. Note that by increasing the order k we increase the number of parameters in the RSW function and the sum of squares in Equation (5) decreases. Therefore, the first term $n\log(\text{SSQ}(\widehat{\vartheta}^{(k)},\widehat{\gamma}^{(k)}))$ in Equation (8) becomes smaller, if the number of parameters in the model is increased. However this decrease is penalized by the second term $k\log(n)$ which increases with k. Thus, by minimizing the Bayesian Information Criterion one tries to obtain a reasonably good fit (that is a small value for the sum of squared deviations between the observed data and the predicted values) with a reasonably small number of parameters. General mathematical arguments as given in Ref. [24] show that the BIC criterion estimates the unknown order of the model correctly if the sample size is reasonably large and the “true” model is in fact a piecewise RSW function.

3.2. Sequential design of experiments

In this section we describe the estimation procedure which simultaneously estimates the model (including the number and location of the change points) and additionally determines the location of the new data points in an efficient way. We suggest this procedure to replace any sampling strategy in which the number and spacing of data points to obtain (e.g. from atomistic simulations) is fixed prior to the screening run. Note that in the following the term “experiment” is used in the sense of “numerical simulation”. The steps are the following:

(1) (Initialization) Start with an initial sample of size n_0, say $\{(x_i,E_i)|i=1,\ldots,n_0\}$, of experiments taken (for example from a homogeneous design) on the given interval, which is $[0,45]$, $[0,60]$, $[0,90]$ or $[0,180]$.

(2) (n_s further sequential experiments) For $j=n_0,\ldots,n=n_0+n_s$ do.

(2.1) (Estimation of parameters) Determine a fit $F_{\text{RSW}}(x,\widehat{\vartheta}_j,\widehat{\gamma}_j,\widehat{k}_j)$ of the RSW function from the data points $\{(x_i,E_i)|i=1,\ldots,j\}$. For this purpose we determine the estimates $\widehat{k}_j$, $\widehat{\vartheta}_j$, $\widehat{\gamma}_j$ using the least squares method and the BIC criterion described in Section 3.1.1 and 3.1.2.

(2.2) (Find the next point) Calculate a prediction $\text{pred}_j(x)$ for any point x of interest, where the uncertainty of the prediction is specified by the function $\widehat{V}(x)$ defined in (7). The next point x_{j+1} is then determined as the maximizer of the function $\widehat{V}(x)$. In other words we run the next simulation under an experimental condition x_{j+1}, where we expect the largest uncertainty. Interestingly, in the simulated examples the maximum variance is not necessarily obtained between the points with a maximum variation of the misorientation angle. The outcome of this experiment is denoted by E_{j+1}.

(2.3) (Increase the sample size) Put $j \leftarrow j+1$.

Note that the current goal of our project is not to find the best function to fit an existing database, but to find the best strategy to create a new database. Therefore the proposed algorithm is based on a specific model assumption for the grain boundary energy function. For illustration we considered the RSW function, but our approach is generally applicable for any parametric model which is used to describe the grain boundary energy function. If the model is misspecified, least squares estimation produces a best

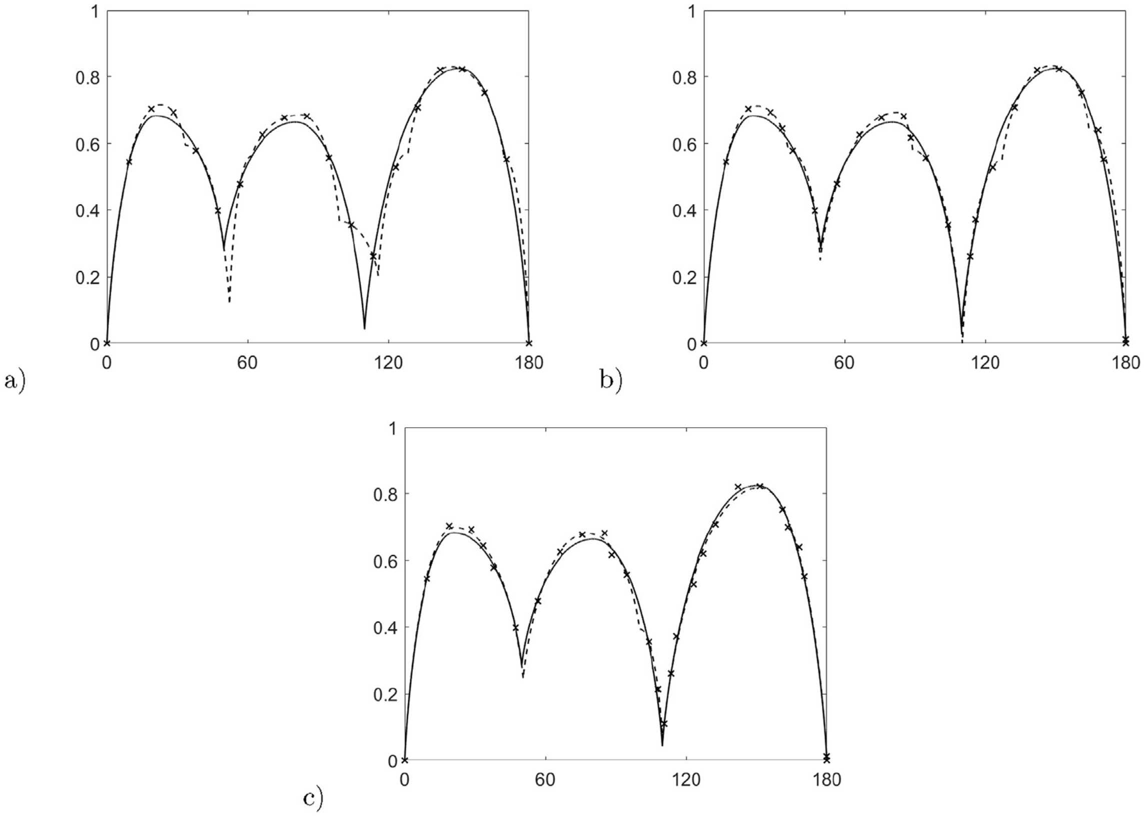


Fig. 4. Estimates (dashed curves) of function F_2 (solid curves) for sample sizes 20, 25 and 30. After 20 initial points (homogeneously sampled), all further points have been obtained by the sequential procedure described in Section 3.2.

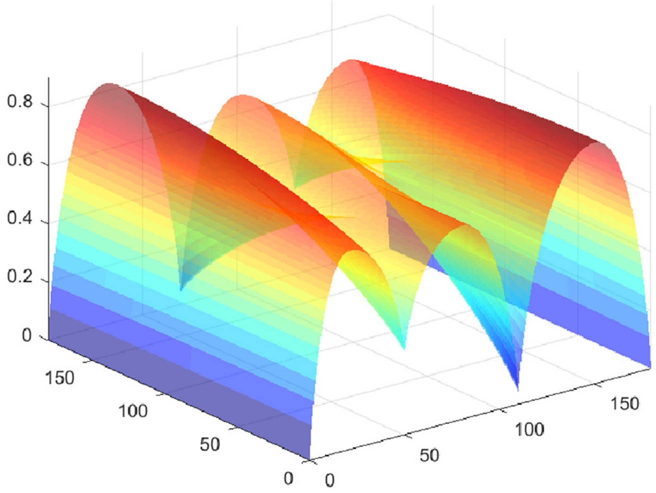


Fig. 5. Energy function of the two-dimensional subspace of [110]-tilt grain boundaries as it was proposed by Bulatov et al. [4], generated with ©MATLAB.

approximation of the unknown model by the parametric class (see Ref. [24]) and sequential sampling will improve the estimation of this approximation. On the other hand the quality of the approximation depends on the chosen model class and the true model. Nevertheless, if RSW functions are considered, our approach exhibits some more flexibility as it does not assume knowledge of the change points (in contrast to [4]).

3.3. Example

In this section we give a brief illustration of the estimation and sequential sampling procedure proposed in the previous paragraphs. Note that instead of carrying out an atomistic simulation to obtain the grain boundary energy for a given misorientation angle, we have simply used the model (6) to calculate it. Thus, the simulated data points follow piecewise a RSW-curve (subject to a random error), but we pretend this is not known and estimate it instead. To mimick the deviations which usually occur, between the results of real atomistic simulations and an idealized model of grain boundary energies, we have added a normal distributed measurement error.

Suppose that the true model is given by the function F_2, which is shown in Fig. 8. In order to estimate this function we run 20 experiments homogeneously sampled over the interval $[0, 180]$, which corresponds to the initial step in Section 3.2. The estimate of the function F_2 based on this starting set is displayed by a dashed curve (in Fig. 4a) (the solid curve represents the “true” function F_2). The next points are then sampled sequentially as described in the previous section. In Fig. 4b) we display the estimate (dashed curve) on the basis of 5 additionally sampled points (this means that here the total sample size is 25). We observe that the quality of the estimate has already improved. Performing the sequential procedure five further times, we can estimate the function on the basis of 30 observations and the result is shown in Fig. 4c). The quality of the estimate has further improved and the function F_2 is nearly reconstructed.

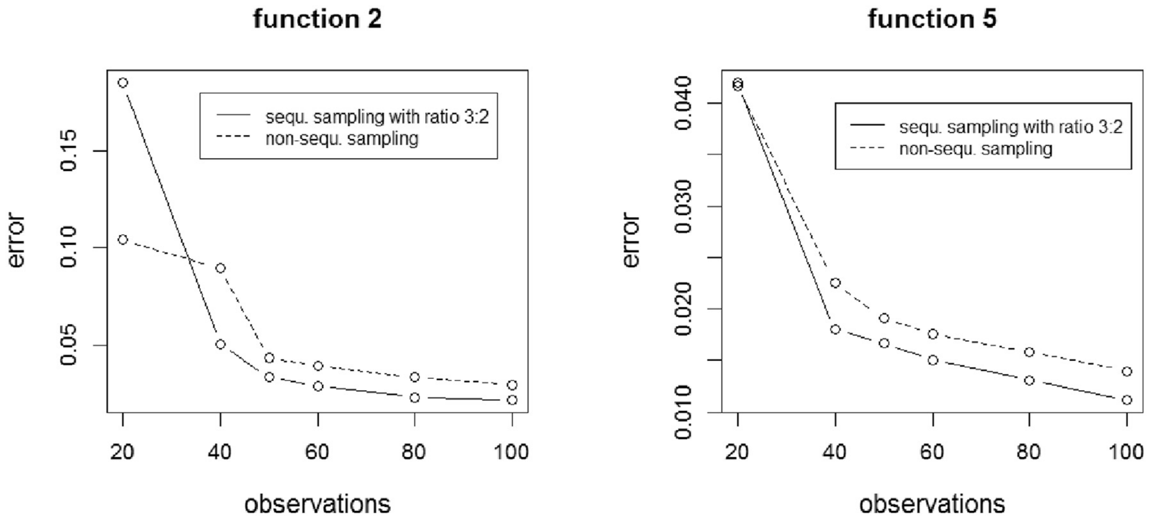


Fig. 6. Estimation error of fits from a sequential (solid line) and non-sequential design (dashed line) for various sample sizes for RSW functions F_2 and F_5.

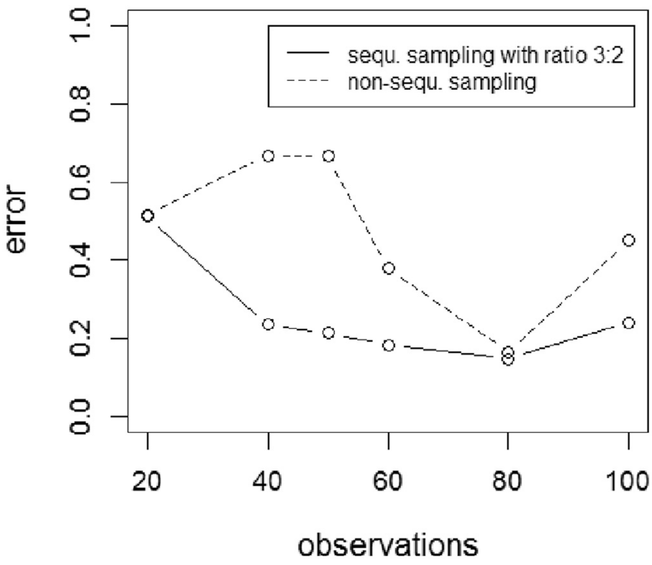


Fig. 7. Estimation error of fits from a sequential (solid line) and non-sequential design (dashed line) for various sample sizes for the two-dimensional RSW functions F_{2D}.

3.4. Interpolation to higher dimensions

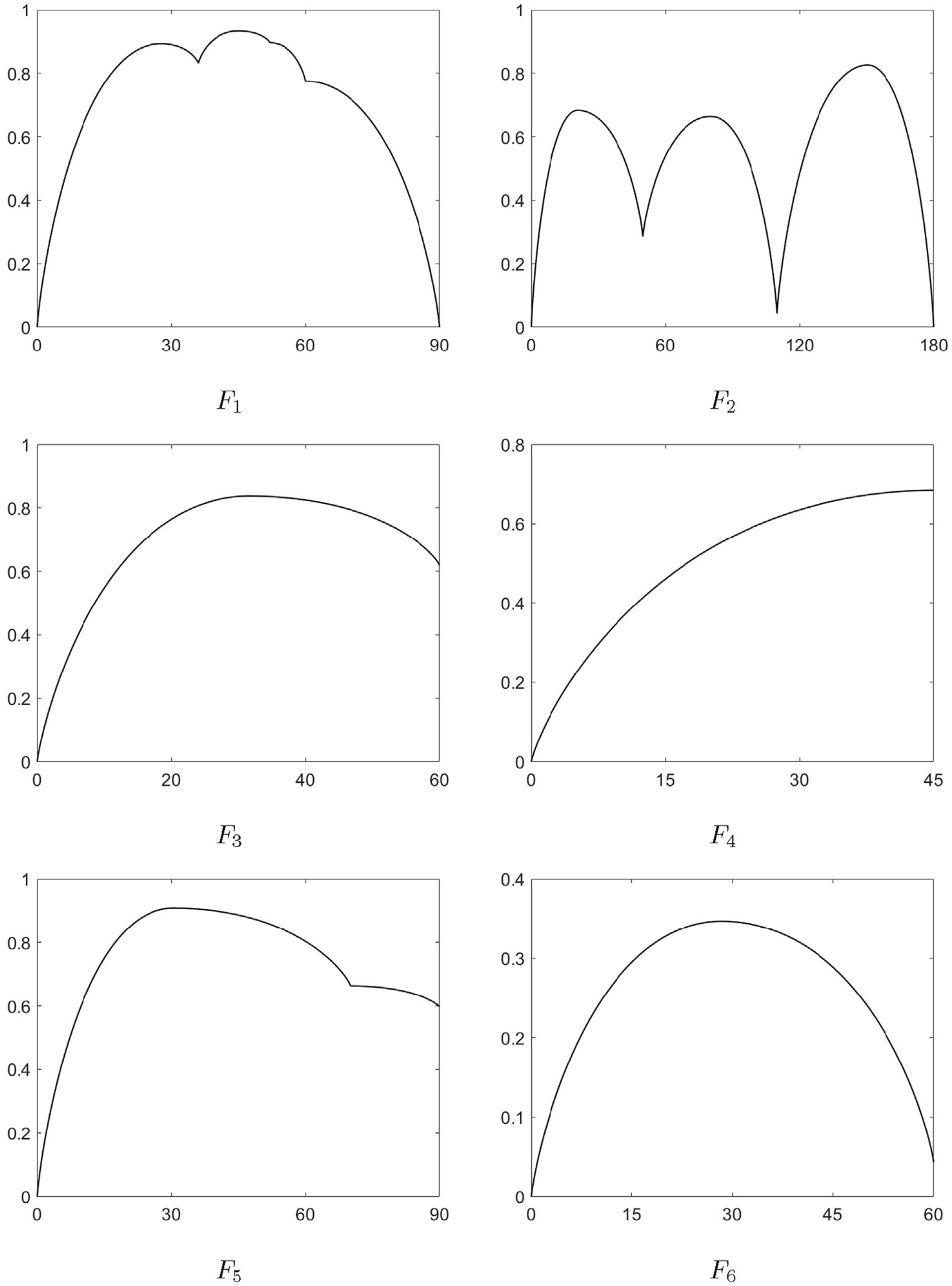
As mentioned in Section 2.2 the approach of Bulatov et al. [4] has its focus on an accurate modeling of the 1D subspaces with the characteristic change points. This part is based on a physical model, even if the extension to the complete misorientation range is empirical. The extension to higher dimensions, however, is then simply done by linear interpolation between the scaffolding sets. The first step is to extend the 1D subset of symmetric tilt grain boundaries to the 2D subset of asymmetric and symmetric tilt grain boundaries, this means including the effect of the interface plane inclination. As argued above, a linear approximation is probably not the best one, since the space of plane inclinations is not flat. Nevertheless, we will use this two-dimensional model to demonstrate the general applicability of the sequential sampling method introduced in this paper. Thus, we follow the approach of [4] also for a higher-dimensional case study. An improved modeling of the grain boundary energy function is left for future work. As an example we pick the subspace of [110]-tilt grain boundaries. For this purpose, denote with F_{110} the one-dimensional function of symmetric [110]-tilt boundaries. In our model, this is a RSW-function, so we have $F_{110}(\,\cdot\,) = F_{\mathrm{RSW}}(\,\cdot\,, \vartheta, \gamma, k)$ for an unknown parameter triple (ϑ, γ, k). The two-dimensional function F_{2D} for the regarded subspace is then given by the formula

$$F_{2D}(x_1,x_2) = \begin{cases} F_{110}(180-x_1) \\ +(F_{110}(x_1) - F_{110}(180-x_1))f_{\mathrm{RSW}}\left(1-\frac{x_2}{180}, a\right) \\ \text{if } F_{110}(x_1) > F_{110}(180-x_1), \\ F_{110}(x_1) \\ +(F_{110}(180-x_1) - F_{110}(x_1))f_{\mathrm{RSW}}\left(\frac{x_2}{180}, a\right) \\ \text{otherwise}. \end{cases} \tag{9}$$

It is displayed in Fig. 5 for the actual choice $F_{110} = F_2$, where the latter function was proposed in Ref. [4] for the symmetric [110]-tilt boundaries and is shown in Fig. 8. The function $f_{\mathrm{RSW}}(\,\cdot\,, a)$ denotes an RSW-segment on the interval $[0,1]$ as defined in Formula (2), only that now a is also considered as an additional shape parameter, which controls the linear interpolation between $F_{110}(x_1)$ and $F_{110}(180-x_1)$. The sequential sampling method described in Section 3.2 can be straighforwardly adapted to work in this two-dimensional setting. An estimator $\widehat{a}$ for the shape parameter a is defined together with $\widehat{\vartheta}$ and $\widehat{\gamma}$ by extending Formula (5) accordingly. Including $\widehat{a}$ in Formula (7) leads to an adjusted point selection criterion and thus we are directly in position to provide a sampling scheme for this two-dimensional subspace. Other models can be treated in the same way.

4. Empirical results and discussion

In this section we present a simulation study investigating the differences between sequential and homogeneous sampling for the one-dimensional space of energy as a function of misorientation angle for symmetric tilt grain boundaries, as well as the two-dimensional space of symmetric and non-symmetric tilt grain



F_1 F_2

F_3 F_4

F_5 F_6

Fig. 8. Six RSW energy functions considered in Ref. [4]. The RSW functions F_1, F_2 and F_3 belong to the sets symmetric of tilt grain boundaries with fixed misorientation axis $\langle 100\rangle$, $\langle 110\rangle$ and $\langle 111\rangle$, respectively. The other triple F_4, F_5 and F_6 belongs to the sets of twist grain boundaries with fixed misorientation axis $\langle 100\rangle$, $\langle 110\rangle$ and $\langle 111\rangle$, respectively. Note that, the shape parameter a given in Ref. [4] was set to $a = 0.5$, which only affected functions F_4 and F_6.

boundaries. The experimental conditions $x_1, \ldots, x_n$, in this case the angles for which we require the grain boundary energies, were obtained in two ways. First the misorientation space was sampled homogeneously with n equidistant points $x_1, \ldots, x_n$. This method will be denoted by (M1) in the following discussion. Secondly, we applied the sequential procedure proposed in Section 3.2 to determine the n experimental conditions, where we used homogeneous designs of different sizes as initial sample. This method will be denoted by (M2).

We have simulated data according to model (6) with a normal distributed measurement error with a variance of 0.02. Six different RSW functions are investigated which are taken from Ref. [4] and are displayed in Fig. 8. The RSW functions F_1, F_2 and F_3 belong to the sets of symmetric tilt grain boundaries with fixed misorientation axis $\langle 100\rangle$, $\langle 110\rangle$ and $\langle 111\rangle$, respectively. The other triple, F_4, F_5 and F_6, belongs to the sets of twist grain boundaries with fixed misorientation axis $\langle 100\rangle$, $\langle 110\rangle$ and $\langle 111\rangle$, respectively. For simplicity, the shape parameter a given in Ref. [4] was set to $a = 0.5$, which only affected functions F_4 and F_6. In Tables 1 and 2 we display the simulated error

$$\max_x \left| F_{\mathrm{RSW}}\left(x, \widehat{\vartheta}, \widehat{\gamma}, \widehat{k}\right) - F_{\mathrm{RSW}}(x, \vartheta, \gamma, k) \right|$$

between the estimated and “true” piecewise RSW function

Table 1
Simulated estimation error for different designs and the RSW functions F_1, F_2, F_3 defined in Ref. [4] and shown in Fig. 8. The total number of measurement points (misorientation angles), n, is always 50, of which n_0 were chosen in a homogeneous fashion in the initial step, and n_s sequentially at positions of maximum uncertainty. Thus, the first row in the table corresponds to a completely homogeneous sampling scheme, the last row to a mostly sequential one. The numbers in bold face mark the smallest estimation error for the different designs. For more details see text.

n_0	n_s	F_1	F_2	F_3
50	0	**0.059938**	0.043230	0.009729
40	10	0.067764	0.056167	0.007985
30	20	0.070953	**0.033227**	0.007900
20	30	0.073819	0.049871	**0.007786**
10	40	0.079652	0.080023	0.007877

Table 2
Simulated estimation error for different designs and the RSW functions F_4, F_5, F_6 defined in Ref. [4] and shown in Fig. 8. The total number of measurement points (misorientation angles) is n is always 50, of which n_0 were chosen in a homogeneous fashion in the initial step, and n_s sequentially at positions of maximum uncertainty. Thus, the first row corresponds to a completely homogeneous sampling scheme, the last row to a mostly sequential one. The numbers in bold face mark the smallest estimation error for the different designs. For more details see text.

n_0	n_s	F_4	F_5	F_6
50	0	0.003103	0.019071	0.010073
40	10	0.003044	**0.016468**	0.008094
30	20	0.003112	0.016708	0.008085
20	30	0.003243	0.017488	0.008056
10	40	**0.002992**	0.017668	**0.007677**

$F_{\text{RSW}}(x, \vartheta, \gamma, k)$ for the function F_1, F_2, F_3 and F_4, F_5, F_6, respectively. All results are based on 1000 simulation runs. For comparison of the different sampling strategies, we limited the number of observations (grain boundary energies) to fit the RSW function by $n = 50$. The first column of each table shows how many points were chosen homogeneously in the initial step, while the second columns give the number of points chosen sequentially. This means that the first row of each table represents the results of sampling method (M1), where no points have been chosen sequentially and misorientation angles are chosen completely by a homogeneous design in the initial step. Similarly, e.g. the third row with the values 30 and 20 in the first and second columns correspond to the sequential method (M2) with an initial design of size 30 and a sequential design of size 20 in the second step. In this case the simulated estimation error for the function F_2 is given by 0.033227. The numbers printed in boldface show the best estimation error for the different designs. For example, for the function F_1 the non-sequential design is the best, while the sequential design with 30 observations in the initial step yields a 18% larger error. Similarly, for the function F_3 the sequential design with 10 observations in the initial step is the best and the non-sequential design has a 25% larger error.

Note that F_1 is the only one for which the homogeneous sampling design shows the best performance. In all other cases a sequential design yields a smaller estimation error, where the improvement varies between 15% and 25%. Thus the sequential approach seems to be an interesting alternative to homogeneous sampling.

A particular important question in applications is how many observations are needed to obtain a fit of the RSW function with an estimation error smaller than a given threshold. To investigate this issue we consider the performance of a non-sequential and a sequential design with 60% of the observations in the first step for different sample sizes. In Fig. 6 we display exemplarily the estimation error in the estimation of the function F_2 and F_5, where the sample size varies between 20 and 100. For example, if the sample size is 40, the sequential design takes 24 observations in the first step and 16 observations are taken sequentially as described in Section 3.2. 1000 simulation runs have been used to calculate this error. For the RSW function F_2 we observe a better performance of the new sequential sampling scheme whenever the sample size is larger than 40, and this design always yields a smaller estimation error for the function F_5.

The results in Fig. 6 can also be interpreted in the following way. Assume that one is interested in the estimation of the function F_5 with an error not larger than 0.020. If one uses a non-sequential design, 60 observations are required to achieve this precision, while the same precision can be obtained with only 40 measurements sampled according to the sequential design. These results demonstrate the benefits of the sequential sampling approach proposed in this paper.

In Fig. 7 we display the error that occurs while estimating a two-dimensional RSW-Function, as described in Section 3.4. The simulated model is given by the function F_2 in Fig. 8 and the ratio between the initial and remaining sample size is again $3:2$. Except for the case of twenty observations, the sequential sampling outperforms the non-sequential one. Furthermore, one can observe that the quality of a sequential sampling scheme is relatively stable. In contrast to this, the error of the non-sequential sampling can rise significantly when the observation size is increased, due to a worse location of the grid with respect to the cusps of the one-dimensional RSW-Function F_2. With the current model, this effect is particularly strong. The cusps in the energy landscape are located on the boundaries of the two-dimensional subspace (i.e. in the one-dimensional subspace of symmetric tilt grain boundaries). Due to the linear interpolation scheme (see Equation (9)) the energy increases or decreases monotonously between these boundaries, see Fig. 5. The sequential scheme adapts to this specific form and is able to detect this strongly inhomogeneous distribution of energy cusps, by selecting new points mainly along the boundary. In contrast, the homogeneous sampling also adds points in those regions where the energy varies monotonously.

We can summarize the empirical study of this section as follows: in many cases the sequential procedure shows a better performance, but there also some models, where the homogeneous sampling is not improved. These are typically situations, where the homogeneous sampling method (sequential or non-sequential) already captures the critical points of the unknown function. Obviously in such a situation a homogeneous sampling method will perform better. At the same time, the sequential sampling method is always superior when the critical points are distributed very inhomogeneously, as shown by the two-dimensional example. Note however, that the sample size in the initial step has to be sufficiently large. If it is too small, one does not obtain precise information about the model in the first step of the algorithm, which is updated in the subsequent experiments. In such cases the error at the beginning of the procedure is too large and transferred to the following predictions.

In the present paper we have considered a sequential sampling scheme where in each step of the algorithm one additional atomistic simulation is performed. This approach has the drawback that such a sampling cannot be parallelized. However, our approach can easily be modified such that it allows the selection of several points, say $x^{(1)}, \ldots, x^{(d)}$, in one step. These points are chosen by minimizing the determinant of the covariance matrix of the predictions $\text{pred}(x^{(1)}), \ldots, \text{pred}(x^{(d)})$ and the next d simulations at these experimental conditions can be run in parallel. Even if it takes fewer points to get the same accuracy with a sequential scheme adding only one point in each step, it may still take less time to get the same accuracy with a sequential scheme adding d points in each step, if the samples can be generated in parallel. The optimum number d in each new step depends sensitively on the type of numerical experiment that is carried out (e.g. whether ab-initio calculations are performed or an empirical potential is used) and the computational power that is available to the user. Since the number of available CPUs is usually limited, it might for instance be necessary to use it completely on one DFT data point at a time, and the sequential sampling with one additional point in each step will be highly beneficial. With a less time-consuming approach, CPU power can be diverted to a parallel approach. For such cases an important question for future research is to investigate the trade-off (in computing time) between these different sequential sampling schemes.

5. Summary and conclusion

In the framework of materials design there exists an increasing demand for databases of specific materials properties. In this work we have suggested an improved strategy for creating future databases.

Among the existing databases, the one of Kim et al. [17] is based on a homogeneous sampling of the misorientation as well as the plane-inclination subspace, extended by the energies of special grain boundaries that were missed in the homogeneous scheme. In contrast, the database of Olmsted and co-workers [9] is restricted to several subsets of special misorientations. In combination with a physics based interpolation function as suggested by Bulatov [4] the latter seems more efficient, but in the current state is lacking sufficient data in the plane-inclination subspace.

In any case both schemes rely on the a-priori knowledge of the location of important cusps and maxima in the five-dimensional energy landscape of grain boundaries. We suggest two ways to improve the current state of the art. Firstly, the location and number of the energy minima along which the hierarchical sampling takes place, can be predicted from existing data points without any a-priori knowledge, using a predictor function. Second, instead of a sampling a fixed number of equally spaced data points in a given subspace, we show that it is more efficient to use a sequential sampling in a “design of experiment” scheme. In most cases this sequential design exhibits a smaller estimation error than the simultaneous one, and thus can provide the same accuracy with fewer data points. This new strategy should be particularly beneficial for the exploration of extrinsic material properties, if the “property landscape” has an unknown topography, for example grain boundary energies in new alloys and/or non-cubic structures.

Acknowledgements

The authors would like to thank the reviewers for their careful reading of the manuscript and very constructive comments. This work has been supported in part by the Collaborative Research Center ”Statistical modeling of nonlinear dynamic processes” (SFB 823, Project C1, C2) of the German Research Foundation (DFG).

References

[1] A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K.A. Persson, The Materials Project: a materials genome approach to accelerating materials innovation, APL Mater. 1 (1) (2013) 011002.

[2] NOMAD, http://nomad-repository.eu/cms/, version of October 18, 2016.

[3] The Grain Boundary Data Archive, http://mimp.materials.cmu.edu/~gr20/Grain_Boundary_Data_Archive/, version of March 10, 2016.

[4] V. Bulatov, B. Reed, M. Kumar, Grain boundary energy function for fcc metals, Acta Mater 65 (2014) 161–175.

[5] D. Wolf, A broken bond model for grain boundaries in face-centered cubic metals, J. Appl. Phys. 68 (1990) 3220–3236.

[6] B. Runnels, I.J. Beyerlein, S. Conti, M. Ortiz, An analytical model of interfacial energy based on a lattice-matching interatomic energy, J. Mech. Phys. Solids 89 (2016) 174–193.

[7] O. Shenderova, D. Brenner, Atomistic simulations of structures and mechanical properties of polycrystalline diamond: symmetrical ⟨001⟩ tilt grain boundaries, Phys. Rev. B 60 (1999) 7043–7052.

[8] D.L. Olmsted, A new class of metrics for the macroscopic crystallographic space of grain boundaries, Acta Mater. 57 (2009) 2793–2799.

[9] D.L. Olmsted, S.M. Foiles, E.A. Holm, Survey of computed grain boundary properties in face-centered cubic metals: I. Grain boundary energy, Acta Mater. 57 (2009) 3694–3703.

[10] D. Wolf, Correlation between structure, energy, and ideal cleavage fracture for symmetrical grain boundaries in fcc metals, J. Mater. Res. 5 (1990) 1708–1730.

[11] D. Wolf, Correlation between the energy and structure of grain-boundaries in bcc metals. 2. Symmetrical tilt boundaries, Philos. Mag. A 62 (4) (1990) 447–464.

[12] C. Schmidt, M.W. Finnis, F. Ernst, V. Vitek, Theoretical and experimental investigation of structures and energies of $\Sigma=3$, [112] tilt grain boundaries in copper, Phil. Mag. A 77 (1998) 1161–1184.

[13] J. Wang, I.J. Beyerlein, Atomic structures of symmetric tilt grain boundaries in hexagonal close packed (hcp) crystals, Model. Simul. Mater. Sci. Eng. 20 (2) (2012) 024002.

[14] A.D. Banadaki, S. Patala, A simple faceting model for the interfacial and cleavage energies of Σ 3 grain boundaries in the complete boundary orientation space, Comp. Mat. Sci. 112 (2016) 147–160.

[15] D. Wolf, Structure and energy of general grain-boundaries in bcc metals, J. Appl. Phys. 69 (1) (1991) 185–196, http://dx.doi.org/10.1063/1.347741.

[16] K. Kang, J. Wang, I.J. Beyerlein, Atomic structure variations of mechanically stable fcc-bcc interfaces, J. Appl. Phys. 111 (2012) 053531.

[17] H.-K. Kim, W.-S. Ko, H.-J. Lee, S.G. Kim, B.-J. Lee, An identification scheme of grain boundaries and construction of a grain boundary energy database, Scr. Mater 64 (2011) 1152–1155.

[18] S. Ratanaphan, D.L. Olmsted, V.V. Bulatov, E.A. Holm, A.D. Rollett, G.S. Rohrer, Grain boundary energies in body-centered cubic metals, Acta Mater 88 (2015) 346–354. ISSN 1359–6454.

[19] H.-K. Kim, S.G. Kim, W. Dong, I. Steinbach, B.-J. Lee, Phase-field modeling for 3D grain growth based on a grain boundary energy database, Model. Simul. Mater. Sci. Eng. 22 (2014) 034004–034019.

[20] W. Read, W. Shockley, Dislocation models of crystal grain boundaries, Phys. Rev. 78 (1950) 275–289.

[21] D. Wolf, A Read-Schockley model for high-angle grain-boundaries, Scr. Metall. 23 (1989) 1713–1718.

[22] J.-F. Bonnans, J.C. Gilbert, C. Lemarechal, C.A. Sagastizábal, Numerical Optimization: Theoretical and Practical Aspects (Universitext), second ed., Springer, 2006.

[23] P.J. Bickel, K.A. Doksum, Mathematical Statistics, Basic Ideas and Selected Topics, second ed., vol. 1, Pearson Prentice Hall, New Jersey, 2006, p. 2.

[24] G. Claeskens, N.L. Hjort, Model Selection and Model Averaging, Cambridge University Press, Cambridge, 2008.