The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

Level Set Estimationを用いた太陽電池用シリコンのレッドゾーンの効率的推定

Efficient estimation for red-zone in silicon wafers for solar cells using Level Set Estimation

穂積 祥太 *1 Shota Hozumi　松井 孝太 *2 Kota Matsui　沓掛 健太朗 *2 Kentaro Kutsukake　宇治原徹 *3*4*2 Toru Ujihara　竹内 一郎 *1*2*5 Ichiro Takeuchi

*1名古屋工業大学 情報工学専攻 Nagoya Institute of Technology
*2理化学研究所 革新知能統合研究センター RIKEN Center for Advanced Intelligence Project
*3名古屋大学 未来材料・システム研究所 Institute for Materials and Systems for Sustainability, Nagoya University
*4名古屋大学大学院 工学研究科 Department of Engineering, Graduate School of Nagoya University
*5物質・材料研究機構 情報統合型物質・材料研究拠点 Center for Materials Research by Information Integration, National Institute for Materials Science

For the task of estimating a spacial distribution of a physical quantity, it is common to fix the measurement positions to meshgrid points evenly allocated along the coordinates of the space. However, such fixed measurement positions often contain redundancy in the sense that not all the measurements in the meshgrid points are required for the target task. Especially when a measurement of the physical quantity is costly, it is thus beneficial to allocate the measurement points *adaptively* and reduce the number of measurements. In this study, we applied Level Set Estimation (LSE), which is a method to efficiently estimate the boundary position, to carrier lifetime mapping of silicon for solar cells, and estimated the low quality region. Our approach can reasonably estimate the boundary position by measuring less than 1% position compare to conventional approach.

1. 背景

太陽光電池の基板材料であるシリコンインゴッドの品質を測る尺度の 1 つとして，キャリアライフタイムの値がある．キャリアライフタイムは，光で励起されたキャリアが基底状態に戻るまでの時間であり，太陽電池の変換効率と正の相関がある [Sze & Ng 06]. 実際のインゴット製造では，材料のうちルツボに触れている外周部分は製造過程で不純物が入ってしまい，ライフタイムの値が低くなる．材料内でライフタイムの値が低い範囲は特に “レッドゾーン”と呼ばれ，この範囲が狭いほど，良い材料であるという 1 つの指標となっている [Ferrazza 02]. 通常, レッドゾーンの領域はインゴット表面で 1 点ずつライフタイムを測定し，その結果得られるライフタイムの分布から求められる [Hsieh et al. 14].

従来は，測定領域を格子状に区切り，各格子点上でライフタイムを測定することでレッドゾーンの広さを評価していた．しかし，材料内の 1 点のライフタイム値を測定するためには相応のコストがかかる．また，上述の測定方法ではレッドゾーンと無関係な領域にも大きな測定コストをかけてしまう可能性があり，非常に測定効率が悪い．したがって，少ない測定回数でレッドゾーンの広さを正確に推定する方法を開発することは重要な課題である．

本研究では，能動学習手法の 1 つであるレベルセット推定 (Level Set Estimation, LSE) [Gotovos 13] を用いてレッドゾーンの境界を推定する方法を提案する．提案法では, ライフタイム値の分布に対して統計モデルを仮定することでレッドゾーン境界に関する統計的推測を行い，これに基づいて次にライフタイム値を測定する点を適応的に選択する．太陽電池インゴッドのライフタイム値データを用いた実データ実験を通して，提案法が測定点数を劇的に減らすことができることを示す.

連絡先: 穂積祥太，名古屋工業大学大学院 竹内研究室所属，466-8555 愛知県名古屋市昭和区御器所町，hozumi.s.mllab.nit@gmail.com

2. 問題設定

あるブラックボックス関数 $f : \mathcal{X} \to \mathbb{R}$ と観測対象の候補点 $\{\boldsymbol{x}_i\}_{i=1}^N$ が与えられているとする．ここで, $\mathcal{X}$ は測定点の空間を表し, 測定値は実数値であるとする．また，入力 $\boldsymbol{x} \in \mathcal{X}$ に対する関数値 $f(\boldsymbol{x})$ の観測値 y は，誤差 $s \sim \mathcal{N}(0, \sigma^2)$ を伴って $y = f(\boldsymbol{x}) + s$ と表されるとする．本研究で考察するレベルセット推定 (Level Set Estimation，LSE)[Gotovos 13] は，ある閾値 $h \in \mathbb{R}$ に対して，

$$\boldsymbol{x}_i \in \{\boldsymbol{x} \in \mathcal{X} \mid f(\boldsymbol{x}) > h\} \quad \text{or} \quad \boldsymbol{x}_i \in \{\boldsymbol{x} \in \mathcal{X} \mid f(\boldsymbol{x}) \leq h\} \tag{1}$$

を判定する判別問題として定式化される．

特に，本研究で取り扱うレッドゾーン推定問題では，f として試料表面上の 2 次元座標を入力，その座標におけるライフタイムの値を出力とするような 2 入力 1 出力の関数を考える．通常の回帰分析の文脈では，入力 $\boldsymbol{x}$ と誤差を伴った出力 y の複数のペア $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ を学習データとして f を推定することで，新たな入力に対する判別を行うことができる．しかし，材料科学をはじめとする基礎科学分野では，入力 $\boldsymbol{x}$ に対する出力 y を観測するのに大きなコストがかかる場合が少なくなく，本研究で対象とするライフタイム値の測定も同様である．したがって，できるだけ少ない y の観測回数 (すなわち，少ないデータ数) で (1) の判別問題を解くことがしばしば要求される．

Algorithm 1 Level Set Estimation

Require: sample set D, GP prior$(m(\boldsymbol{x}) = a, k, \sigma_0)$
threshold ratio ω , accuracy parameter $\varepsilon = a/5$, max iteration T
Ensure: predicted sets $\hat{H}$, $\hat{L}$, U_t

$H_0, L_0, U_0, Z_0 \leftarrow \emptyset, C_0(\boldsymbol{x}) \leftarrow \mathbb{R}, t \leftarrow 1$
while $U_{t-1} \neq \phi$ and $t \leq T$ **do**
  $H_t \leftarrow H_{t-1}, L_t \leftarrow L_{t-1}, U_t \leftarrow U_{t-1}, Z_t \leftarrow Z_{t-1}$
  for all $\boldsymbol{x} \in D$ **do**
    $h_t^{opt} \leftarrow \omega \ \max_{\boldsymbol{x} \in Z_{t-1}} \max(Q_t(\boldsymbol{x}))$
    $f_t^{pes} \leftarrow \max_{\boldsymbol{x} \in Z_{t-1}} \min(Q_t(\boldsymbol{x}))$
    $h_t^{pes} \leftarrow \omega f^{pes}$
    if $\min(Q_t(\boldsymbol{x})) + \varepsilon \geq h_t^{opt}$ **then**
      $U_t \leftarrow U_t \backslash \{\boldsymbol{x}\}$
      if $\max(Q_t) < f_t^{pes}$ **then** $H_t \leftarrow H_t \cup \{\boldsymbol{x}\}$
        else $M_t^H \leftarrow M_t^H \cup \{\boldsymbol{x}\}$
    else if $\max(Q_t(\boldsymbol{x})) - \varepsilon \geq h_t^{pes}$ **then**
      $U_t \leftarrow U_t \backslash \{\boldsymbol{x}\}$
      if $\max(Q_t) < f_t^{pes}$ **then** $L_t \leftarrow L_t \cup \{\boldsymbol{x}\}$
        else $M_t^L \leftarrow M_t^L \cup \{\boldsymbol{x}\}$
    end if
  end for
  $Z_t \leftarrow U_t \cup M_t^H \cup M_t^L$
  $\boldsymbol{x} \leftarrow \arg\max_{\boldsymbol{x} \in Z_t}(w_t(\boldsymbol{x}))$
  $y_t \leftarrow f(\boldsymbol{x}) + n_t$
  Compute $\mu_t(\boldsymbol{x})$ and $\sigma_t(\boldsymbol{x})$, for all $\boldsymbol{x} \in D$
  $t \leftarrow t + 1$
end while
$\hat{H} \leftarrow H_{t-1} \cup M_{t-1}^H, \hat{L} \leftarrow L_{t-1} \cup M_{t-1}^L$

本研究では，f を, ブラックボックスかつ少データゆえの不確実性の下で適応的に推定するために統計モデルを導入し，モデルに基づいて得られる予測を用いて (1) を効率的に解くアプローチを提案する．

3. 提案手法

以下では，まずブラックボックス関数としての f の不確実性を考慮に入れたガウス過程によるモデリング [Rasmussen & Williams 06] を説明し，その後 (1) の判別問題のための提案アルゴリズムについて説明する．

3.1 ガウス過程によるモデリング

まず，ブラックボックス関数としての f の不確実性を考慮するため，f の統計モデルとしてガウス過程事前分布

$$f(\boldsymbol{x}) \sim GP(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}'))$$

を仮定する．ここで，$GP(m(\boldsymbol{x}, k(\boldsymbol{x}, \boldsymbol{x}')))$ は平均関数 m とカーネル関数 k で特徴付けられるガウス過程を表す．本研究では，k として Matern3/2 カーネル

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j|\theta) = \beta_f^2(1 + \frac{\sqrt{3}r}{\beta_l}) \exp(-\frac{\sqrt{3}r}{\beta_l})$$

を用いる．ここで，$r = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^T(\boldsymbol{x}_i - \boldsymbol{x}_j)}$ とおいた．Matern カーネルは，ガウスカーネルと比較してより非滑らかな関数を表現しやすいという特徴を持つ．このとき，観測点 $\boldsymbol{x}_{1:n} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$ に対して，関数値ベクトル $\boldsymbol{f} = (f(\boldsymbol{x}_1), ..., f(\boldsymbol{x}_n))$ は，定義から n 変量の正規分布に従う

$$\boldsymbol{f}|x_{1:n} \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{K})$$

ここで，$\boldsymbol{m} = (m(\boldsymbol{x}_1, ..., m(\boldsymbol{x}_n)))$, $\boldsymbol{K}_{i,j} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ である．以上より，データ $\mathcal{D}_n = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ が観測された下での新たな入力 $\boldsymbol{x}$ の観測値 $f(\boldsymbol{x})$ の予測分布は $\mathcal{N}(\mu_n(\boldsymbol{x}), \sigma_n^2(\boldsymbol{x}))$ と書ける．ただし，

$$\mu_n(\boldsymbol{x}) = m(\boldsymbol{x}) + \boldsymbol{k}(\boldsymbol{x})^T(\boldsymbol{K} + \sigma^2\boldsymbol{I})^{-1}(\boldsymbol{y} - \boldsymbol{m})$$
$$\sigma_n(\boldsymbol{x}) = k(\boldsymbol{x}, \boldsymbol{x}) + \boldsymbol{k}(\boldsymbol{x})^T(\boldsymbol{K} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{k}(\boldsymbol{x})$$

はそれぞれ予測平均と予測分散を表す．

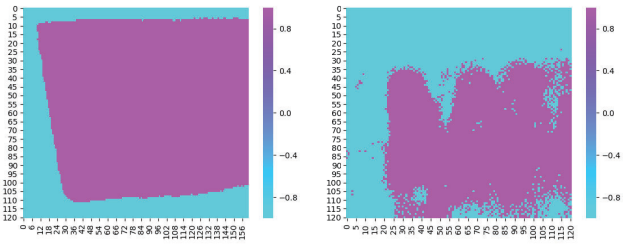

図 1: 左図:データ 1 のライフタイム値のプロット，右図:データ 2 のライフタイム値のプロット．青い領域がライフタイム値が低い低品質領域 (レッドゾーン) を表す．実験では，上図を正解とし，提案法がどの程度復元できたかを評価指標の 1 つとしている．

3.2 レベルセット推定 (LSE)

LSE では，前節のモデリングに基づいて confidence region

$$Q_n(\boldsymbol{x}) := [\ \mu_{n-1}(x) \pm \beta_t^{1/2}\sigma_{n-1}(\boldsymbol{x})\ ]$$

を構成し，閾値 h と $Q_n(\boldsymbol{x})$ の関係から $\boldsymbol{x}$ が h を超えるかどうかの判定を行う．具体的には，閾値に対するマージン ε を導入して, $\min(Q_n) + \varepsilon \geq h$ であれば上側領域, $\max(Q_n) - \varepsilon \geq h$ であれば下側領域であると判別する．ただし，実問題では h が未知であったり，一度どちらかの領域に判別された点でも，不確実性が大きい場合は後の反復で領域が反転したりするといった状況が起こりうるため，より詳細な判別基準を設ける必要がある．以上の考察に基づき，提案する LSE のアルゴリズムを Algorithm 1 に示す．

提案法では，未判定点の集合 U，$\boldsymbol{x}_i \in \{\boldsymbol{x} \in \mathcal{X} \mid f(\boldsymbol{x}) > h\}$ かつ不確実性が十分小さいと判定された点の集合 H，$\boldsymbol{x}_i \in \{\boldsymbol{x} \in \mathcal{X} \mid f(\boldsymbol{x}) > h\}$ かつ不確実性が大きいと判定された点の集合 M^H，$\boldsymbol{x}_i \in \{\boldsymbol{x} \in \mathcal{X} \mid f(\boldsymbol{x}) \leq h\}$ かつ不確実性が十分小さいと判定された点の集合 L，$\boldsymbol{x}_i \in \{\boldsymbol{x} \in \mathcal{X} \mid f(\boldsymbol{x}) \leq h\}$ かつ不確実性が大きいと判定された点の集合 M^L の 5 つの集合を繰り返し更新し，全ての観測候補点が H または L に振り分けられた時点で終了する．

提案法では，最も不確実性の高い点で次の関数値を測定する (uncertainty sampling) アプローチを採用する．これは，獲得関数

$$w_t(\boldsymbol{x}) = \max(Q_t(\boldsymbol{x})) - \min(Q_t(\boldsymbol{x}))$$
$$= 2\beta_n^{1/2}\sigma_{n-1}(\boldsymbol{x})$$

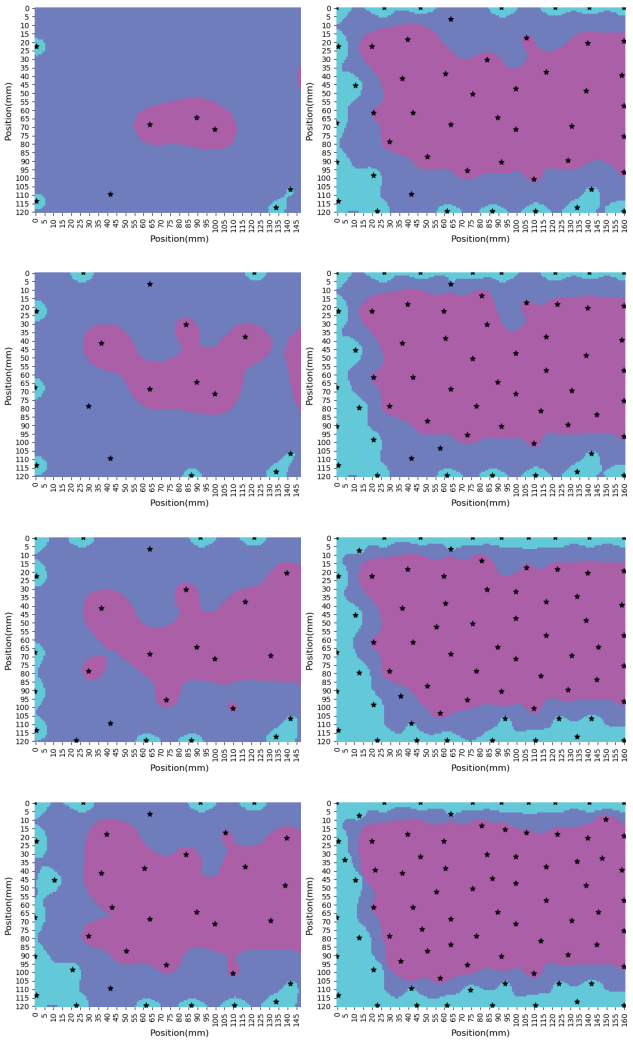

図 2: データ 1 に対する提案法の挙動．左上段から左下段に向けて，獲得点数が 10，20，30，40 時点，右上段から右下段に向けて，獲得点数が 50，60，70，80 時点の推定レッドゾーン (水色領域) を表し．濃青領域は，不確実性が残っていることを表す．また，黒点はライフタイムの観測を行った座標を示す．

の最大値を与える点として特徴付けられる．アルゴリズムの各ステップで，この基準で選ばれた $\boldsymbol{x}$ に対して y の値を観測し，ガウス過程を更新する．

4. 計算機実験

4.1 実験設定

本研究では，太陽電池インゴッドのライフタイム値を測定した 2 種類のデータを解析した．データ 1 については試料表面の 2 次元平面上に格子点を 19481 点，データ 2 については 14641 点とり，その上でライフタイム値の測定を実施したものである．計算機実験では，レッドゾーンの閾値をライフタイム値の最大値の 15 %値と定め，これを下回る範囲をレッドゾーンと定義する．図 1 にそれぞれのデータにおけるライフタイム分布のプロットを示す．図の青い領域が真のレッドゾーンを表している．本実験ではライフタイムを評価する点の数の最大値を 100 に固定し，アルゴリズムが停止した時点で推定されたレッドゾーンを F 値及び可視化によって評価した．

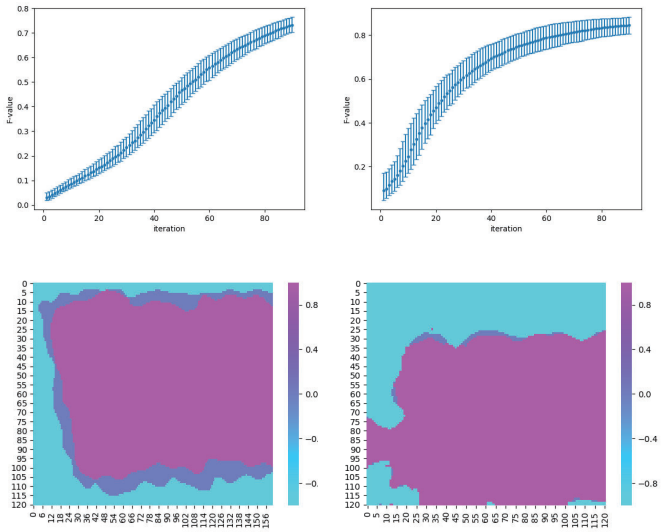

図 3: 上段:F 値のプロット，下段:推定されたレッドゾーンの可視化 (水色領域)．左側がデータ 1，右側がデータ 2 の結果をそれぞれ示す．

ガウス過程に基づく LSE では，観測誤差の分散 σ^2 をデータから推定するとともに，次の 5 種類のハイパーパラメータを設定する必要がある:

1. カーネル関数の分散 β_f
2. カーネル関数の長さスケール β_l
3. ガウス過程の平均関数 $m(\boldsymbol{x})$
4. LSE における分類マージン ε
5. confidence region Q_n で平均と標準偏差のバランスを制御するパラメータ β_t.

文献 [Gotovos 13] では，ライフタイム値の取りうる範囲を被覆するように取ったデータの一部をを用いて，対数尤度最大化に基づいてパラメータの設定を行っている．しかし，実際の実験ではライフタイム値の取りうる範囲が未知であるため，この方法でパラメータを調整しておくことが難しい．本研究では，経験的な知見から $\sigma = 0.001, \beta_l = 25, \beta_t = 1.96$ と設定した．他のパラメータ設定は，まず一様ランダムに 10 点のライフタイム値を測定し，その中央値の半分の値を a とおき，$\beta_f^2 \in [25a, 100a], m(x) = a, \varepsilon = a/5$ とするヒューリスティクスを用いた．特に β_f に関しては，各反復で 1 点のライフタイムを測定する毎に上記の範囲内で対数尤度最大化に基づいて適切な値を設定した．

4.2 実験結果

計算機実験は 1000 回の試行を行い，平均的な振る舞いを観察した．まず，データ 1 に対する提案法の挙動を図 2 に示す．探索初期では大部分が濃青の不確実な領域を占めているが，探索が進むにつれて不確実性が減少していき，徐々にレッドゾーンが推定されていっていることが見てとれる．また，実際にライフタイム値を観測する点も適応的に選択されており，格子点上を全探索する従来のアプローチに比べて探索が効率的になっ

ていることがわかる．次に，データ 1 及びデータ 2 に対する結果をそれぞれ図 4. に示す．図 4. 上段の F 値のグラフは中央値を中心にエラーバーの上限は第一四分位数，下限は第三四分位数として描画した．F 値は，データ 1 で 0.7325，データ 2 で 0.8458 程度を達成した．また，推定されたレッドゾーンを可視化した図 4. の下段と真のレッドゾーンを示した図 1 及び図 4. の下段を比較すると，最大 100 点の測定で比較的外形が良く推定できていることがわかる．ここで，図 4. 下段の濃青の領域は，境界周辺の不確実性がまだ残っていることを表している．真のレッドゾーンを同定するために測定したデータ点数に対しては，データ 1 で 0.5%（100/19481），データ 2 で 0.6%（100/14641）の測定点数で上記の推定結果を得ることができた．探索コストを大幅に削減できていることがわかる．

5. まとめ

本研究では, 太陽電池インゴッドの品質管理において, ライフタイム値の低い低品質領域 (レッドゾーン) を推定するために, レベルセット推定に基づく能動学習を提案した．レベルセット推定は, ライフタイムをブラックボックス関数としてガウス過程でモデリングすることで, レッドゾーンの推定とデータの測定を適応的に行うことができるため, 探索効率の改善が期待できる．計算機実験によって, レッドゾーンを正確に同定できる一方で探索コストの非常に高い格子点上測定に対して, 提案法では非常に少ない探索コストでレッドゾーンの概形を推定することができた．本研究における課題として, レベルセット推定は欠損値を含むデータには適用できない, という点が挙げられる．今回の実験では, データに対して専門家の知識を利用した欠損値処理などの前処理が多数行われているが, 実際の実験でリアルタイムにレベルセット推定を行う場合, これらの前処理を同時に実施することは難しい．今後は, 欠損値を含むデータに対して適用できるようにレベルセット推定アルゴリズムを拡張し, リアルタイムの実験に適用したい．

参考文献

[Ferrazza 02] Ferrazza, F. (2002). Large size multicrystalline silicon ingots. Solar energy materials and solar cells, 72(1-4), 77-81.

[Gotovos 13] Gotovos, A., Casati, N., Hitz, G., & Krause, A. (2013). Active learning for level set estimation. In IJCAI (pp. 1344-1350).

[Hsieh et al. 14] Hsieh, C. C., Lan, A., Hsu, C., & Lan, C. W. (2014). Improvement of multi-crystalline silicon ingot growth by using diffusion barriers. Journal of Crystal Growth, 401, 727-731.

[Rasmussen & Williams 06] Rasmussen, C. E., & Williams, C. K. (2006). Gaussian process for machine learning. MIT press.

[Sze & Ng 06] Sze, S. M., & Ng, K. K. (2006). Physics of semiconductor devices. John wiley & sons.