A Fast Spatial Variation Modeling Algorithm for Efficient Test Cost Reduction of Analog/RF Circuits

Hugo Gonçalves[1,2], Xin Li[1], Miguel Correia[2], Vitor Tavares[2], John Carulli[3] and Kenneth Butler[4]
[1]Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA
[2]Faculdade de Engenharia da Universidade do Porto, Rua Dr. Roberto Frias, s/n 4200-465 Porto PORTUGAL
[3]GLOBALFOUNDRIES, 2600 Great America Way, Santa Clara, CA 95054, USA
[4]Texas Instruments Inc., 12500 TI Boulevard, MS 8741, Dallas, Texas 75243, USA

***Abstract*—In this paper, we adopt a novel numerical algorithm, referred to as dual augmented Lagrangian method (DALM), for efficient test cost reduction based on spatial variation modeling. The key idea of DALM is to derive the dual formulation of the L_1-regularized least-squares problem posed by Virtual Probe (VP), which can be efficiently solved with substantially lower computational cost than its primal formulation. In addition, a number of unique properties associated with discrete cosine transform (DCT) are exploited to further reduce the computational cost of DALM. Our experimental results of an industrial RF transceiver demonstrate that the proposed DALM solver achieves up to 38× runtime speed-up over the conventional interior-point solver without sacrificing any performance on escape rate and yield loss for test applications.**

I. Introduction

As the integrated circuit (IC) technology moves to the nanoscale era, large-scale parametric variations have made analog/RF circuits increasingly difficult to design and test. The performance of an analog/RF circuit can substantially vary from lot to lot, from wafer to wafer, and from die to die. For this reason, analog/RF circuit testing has contributed to a large, or even dominant, portion of the overall test cost for today's complex systems on chip (SOCs) [1]-[2].

To reduce the test cost and, consequently, the manufacturing cost of analog/RF circuits, a large number of algorithms and methodologies have been extensively studied over the past decade [3]-[8]. Among them, spatial variation modeling has emerged as a promising technique in recent years [9]-[18]. The key idea is to accurately capture the spatial variation pattern over all dies on the same wafer by using advanced statistical algorithms such as Virtual Probe (VP) [9]-[14] and Gaussian process (GP) [15]-[18]. As such, we only need to measure a small number of dies of the wafer, while the performance metrics, referred to as test items, of other dies are accurately predicted without physical measurement. It, in turn, substantially reduces the test cost.

In particular, the VP method is derived from the theory of sparse approximation where the spatial variation pattern of a test item is assumed to carry a sparse structure in frequency domain. Namely, the spatial variation pattern can be accurately represented by a small number of spatial frequency components based on discrete cosine transform (DCT) [9]-[14]. These DCT coefficients can be determined by solving a L_1-regularized least-squares problem. Such a L_1-regularized problem can be cast to a convex optimization and robustly solved by the interior-point method with guaranteed global optimum. However, even though efficient interior-point methods have been developed for large-scale sparse approximation in the literature [19], its computational cost remains prohibitively high for the application of test cost reduction where we must run VP in real time during the testing process. For instance, when applying the interior-point method to solve VP, it may take a few minutes to process the measurement data and reconstruct the wafer-level spatial variation pattern, thereby posing an important limitation of the test cost reduction method based on VP.

In this paper we propose an efficient numerical solver that is particularly tuned for VP. Our proposed solver is derived from the dual augmented Lagrangian method (DALM) [20]-[22]. It exploits three unique properties of VP to reduce its runtime. First, VP often measures very few dies for spatial variation modeling and, hence, the number of measurements is substantially less than the number of unknown DCT coefficients (i.e., the optimization variables). In this case, we can mathematically map the original L_1-regularized optimization to its dual formulation where the number of optimization variables of the dual problem is significantly reduced. Second, since VP relies on DCT to model the spatial variations, we can exploit the orthogonal property of DCT basis functions to simplify the computation of DALM. Finally, we apply a fast DCT transform to efficiently calculate matrix-vector multiplications within the iteration loop of DALM to further reduce the computational cost.

To validate the efficacy of our proposed DALM solver for test cost reduction, a set of wafer probe measurement data of an industrial RF transceiver is used. The data set is collected from approximately 1.2M dies distributed over 176 wafers and 9 lots where each wafer contains over 6,000 dies. For each die, more than 50 test items are considered in our experiment. Since VP was previously compared with other algorithms [13], our numerical experiments in this paper mainly focus on the prediction accuracy and computational cost for VP. For testing and comparison purposes, two different solvers are implemented for VP and applied to the same test flow: (i) the conventional interior-point solver for large-scale L_1-regularized least-squares [19], and (ii) the proposed DALM solver. Our preliminary results demonstrate that DALM achieves up to 38× runtime speed-up without sacrificing any

978-3-9815370-4-8/DATE15/©2015 EDAA

performance on escape rate and yield loss.

The remainder of this paper is organized as follows. In Section II, we review the background of VP, and then describe the proposed DALM solver in Section III. Several implementation issues are further discussed in Section IV to make DALM of practical utility. The efficiency of DALM is demonstrated by our experimental results in Section V. Finally, we conclude in Section VI.

II. Background

In this section, we briefly summarize the background of VP [9]-[14]. Without loss of generality, we consider T test items $\{g_t;\ t = 1, 2, \cdots, T\}$ over M wafers. Each test item represents a performance metric (e.g., power consumption, bit error rate, etc.) of a given analog/RF circuit. We use a two-dimensional function $g_{t,m}(x, y)$ to model the spatial variation for the t-th test item of the m-th wafer, where the coordinate (x, y) denotes the spatial location of a die on the wafer. The spatial variation $g_{t,m}(x, y)$ can be further expressed as the linear combination of a set of DCT basis functions:

$$g_{t,m}(x,y) \approx \sum_{k=1}^{K} \alpha_{t,m,k} \cdot b_k(x,y), \tag{1}$$

where $\{b_k(x, y);\ k = 1, 2, \cdots, K\}$ denotes the DCT basis functions, $\{\alpha_{t,m,k};\ k = 1, 2, \cdots, K\}$ stands for the DCT coefficients, and K is the total number of DCT basis functions.

The key idea of VP is to measure a small number of dies $\{(x^{(n)}, y^{(n)}, g_{t,m}{}^{(n)});\ n = 1, \cdots, N\}$ for the t-th test item from the m-th wafer, where $(x^{(n)}, y^{(n)})$ and $g_{t,m}{}^{(n)}$ denote the spatial location and the measured value of the n-th die respectively and N ($N << K$) represents the total number of measured dies. Based on the measurement data, VP formulates the following linear equation:

$$\mathbf{B} \cdot \boldsymbol{\alpha}_{t,m} \approx \mathbf{g}_{t,m}, \tag{2}$$

where

$$\mathbf{B} = \begin{bmatrix} b_1\left(x^{(1)}, y^{(1)}\right) & b_2\left(x^{(1)}, y^{(1)}\right) & \cdots & b_K\left(x^{(1)}, y^{(1)}\right) \\ b_1\left(x^{(2)}, y^{(2)}\right) & b_2\left(x^{(2)}, y^{(2)}\right) & \cdots & b_K\left(x^{(2)}, y^{(2)}\right) \\ \vdots & \vdots & \vdots & \vdots \\ b_1\left(x^{(N)}, y^{(N)}\right) & b_2\left(x^{(N)}, y^{(N)}\right) & \cdots & b_K\left(x^{(N)}, y^{(N)}\right) \end{bmatrix} \tag{3}$$

$$\boldsymbol{\alpha}_{t,m} = \begin{bmatrix} \alpha_{t,m,1} & \alpha_{t,m,2} & \cdots & \alpha_{t,m,K} \end{bmatrix}^T \tag{4}$$

$$\mathbf{g}_{t,m} = \begin{bmatrix} g_{t,m}^{(1)} & g_{t,m}^{(2)} & \cdots & g_{t,m}^{(N)} \end{bmatrix}^T. \tag{5}$$

We need to solve the DCT coefficient vector $\boldsymbol{\alpha}_{t,m}$ from (2). Once $\boldsymbol{\alpha}_{t,m}$ is known, the spatial variation $g_{t,m}(x, y)$ can be recovered by applying the inverse discrete cosine transform (IDCT), which is equivalent to (1).

Note that the linear equation in (2) is underdetermined, since the number of measured dies (i.e., N) is substantially less than the number of unknown DCT coefficients (i.e., K). To find the unique solution of (2), VP further assumes that the solution $\boldsymbol{\alpha}_{t,m}$ is sparse (i.e., containing a small number of non-zeros). Given this assumption on sparsity, the following optimization problem can be formulated to solve (2):

$$\min_{\boldsymbol{\alpha}_{t,m}} \ \frac{1}{2} \cdot \left\| \mathbf{B} \cdot \boldsymbol{\alpha}_{t,m} - \mathbf{g}_{t,m} \right\|_2^2 + \lambda \cdot \left\| \boldsymbol{\alpha}_{t,m} \right\|_0, \tag{6}$$

where $\|\bullet\|_0$ stands for the L_0-norm of a vector (i.e., the number of non-zeros in the vector) and $\|\bullet\|_2$ denotes the L_2-norm of a vector (i.e., the square root of the summation of the squares of all elements in the vector). In (6), λ is a parameter that explores the trade-off between the least-squares error and the sparsity of the solution. The optimal value of λ can be determined by cross-validation.

The optimization in (6) is NP hard and, hence, difficult to solve exactly. To make the problem tractable, VP further relaxes the L_0-norm to L_1-norm, resulting in the following optimization problem:

$$\min_{\boldsymbol{\alpha}_{t,m}} \ \frac{1}{2} \cdot \left\| \mathbf{B} \cdot \boldsymbol{\alpha}_{t,m} - \mathbf{g}_{t,m} \right\|_2^2 + \lambda \cdot \left\| \boldsymbol{\alpha}_{t,m} \right\|_1, \tag{7}$$

where $\|\bullet\|_1$ denotes the L_1-norm of a vector (i.e., the summation of the absolute values of all elements in the vector). The optimization problem in (7) is convex and can be robustly solved by the interior-point method with guaranteed global optimum. However, most conventional interior-point solvers are computationally expensive, especially if the problem size (i.e., the number of unknown DCT coefficients) is large. As will be demonstrated by the experimental results in Section V, the state-of-the-art interior-point solver may take a few minutes to recover the spatial variation $g_{t,m}(x, y)$ for a single test item of a single wafer. It, in turn, substantially slows down the testing process and poses a major limitation of the VP-based method for test cost reduction. To address this issue, we will use an efficient dual augmented Lagrangian method (DALM) to solve (7) so that the computational cost can be significantly reduced. In what follows, the mathematical formulation of DALM will be discussed in detail.

III. Proposed Approach

A. Dual Formulation

Instead of directly solve the L1-norm problem in (7), our proposed DALM algorithm attempts to solve an equivalent dual problem so that the number of unknowns is substantially reduced. To derive the dual formulation, we first add an auxiliary variable $\boldsymbol{\theta}_{t,m} = \mathbf{B} \cdot \boldsymbol{\alpha}_{t,m}$, and re-write (7) as:

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \ & \frac{1}{2} \cdot \left\| \boldsymbol{\theta}_{t,m} - \mathbf{g}_{t,m} \right\|_2^2 + \lambda \cdot \left\| \boldsymbol{\alpha}_{t,m} \right\|_1 \\ \text{S.T.} \ & \boldsymbol{\theta}_{t,m} = \mathbf{B} \cdot \boldsymbol{\alpha}_{t,m} \end{aligned}. \tag{8}$$

The Lagrangian of (8) is expressed as [23]:

$$\begin{aligned} L(\boldsymbol{\alpha}_{t,m}, \boldsymbol{\theta}_{t,m}, \mathbf{d}_{t,m}) = & \frac{1}{2} \cdot \left\| \boldsymbol{\theta}_{t,m} - \mathbf{g}_{t,m} \right\|_2^2 + \lambda \cdot \left\| \boldsymbol{\alpha}_{t,m} \right\|_1 \\ & + \mathbf{d}_{t,m}^T \cdot \left(\boldsymbol{\theta}_{t,m} - \mathbf{B} \cdot \boldsymbol{\alpha}_{t,m} \right) \end{aligned}, \tag{9}$$

where $\mathbf{d}_{t,m} \in \Re^N$ is the dual variable of (8).

Based on (9), we derive the Lagrange dual function by minimizing the Lagrangian over $\boldsymbol{\theta}_{t,m}$ and $\boldsymbol{\alpha}_{t,m}$ [23]:

$$\min_{\boldsymbol{\alpha}_{t,m}, \boldsymbol{\theta}_{t,m}} \ \frac{1}{2} \cdot \left\| \boldsymbol{\theta}_{t,m} - \mathbf{g}_{t,m} \right\|_2^2 + \lambda \cdot \left\| \boldsymbol{\alpha}_{t,m} \right\|_1 + \mathbf{d}_{t,m}^T \cdot \left(\boldsymbol{\theta}_{t,m} - \mathbf{B} \cdot \boldsymbol{\alpha}_{t,m} \right). \tag{10}$$

After a number of mathematical manipulations, we can derive the minimum of (10) as a function of $\mathbf{d}_{t,m}$:

$$-\frac{1}{2}\cdot\left\|\mathbf{d}_{t,m}-\mathbf{g}_{t,m}\right\|_2^2+\frac{1}{2}\cdot\left\|\mathbf{g}_{t,m}\right\|_2^2-\delta_\infty^\lambda\left(\mathbf{B}^T\cdot\mathbf{d}_{t,m}\right), \quad (11)$$

where

$$\delta_\infty^\lambda\left(\mathbf{B}^T\cdot\mathbf{d}_{t,m}\right)=\begin{cases}0 & \left\|\mathbf{B}^T\cdot\mathbf{d}_{t,m}\right\|_\infty\le\lambda\\ +\infty & \left\|\mathbf{B}^T\cdot\mathbf{d}_{t,m}\right\|_\infty>\lambda\end{cases}. \quad (12)$$

In (12), $\|\bullet\|_\infty$ stands for the infinite norm of a vector (i.e., the maximum of the absolute values of all elements in the vector). According to the duality theorem [23], the dual problem is to maximize the Lagrange dual function in (11):

$$\max_{\mathbf{d}_{t,m}} \quad -\frac{1}{2}\cdot\left\|\mathbf{d}_{t,m}-\mathbf{g}_{t,m}\right\|_2^2+\frac{1}{2}\cdot\left\|\mathbf{g}_{t,m}\right\|_2^2-\delta_\infty^\lambda\left(\mathbf{B}^T\cdot\mathbf{d}_{t,m}\right). \quad (13)$$

It can be further proven that the strong duality holds for the primal problem in (7) and the dual problem in (13) [20]-[21]. Namely, the minimum of (7) is exactly equal to the maximum of (13). Hence, we can find the solution of the primal problem in (7) by solving its dual problem in (13).

Note that the solution vector $\boldsymbol{\alpha}_{t,m}$ in (7) contains K unknown DCT coefficients, while the solution vector $\mathbf{d}_{t,m}$ in (13) contains N unknown dual variables. For our application of test cost reduction, the linear equation in (2) is under-determined, and the number of measurements (i.e., N) is substantially less than the number of DCT coefficients. In this case, the dual problem in (13) has significantly less unknowns and, hence, is computationally less expensive to solve than the primal problem in (7). In the next sub-section, we further describe an efficient augmented Lagrangian method to solve the dual problem with low computational cost.

B. Augmented Lagrangian Method

The box constraint defined by $\delta_\infty^\lambda(\mathbf{B}^T\cdot\mathbf{d}_{t,m})$ makes it non-trivial to solve (13), since updating $\mathbf{d}_{t,m}$ with $\mathbf{B}^T\cdot\mathbf{d}_{t,m}$ inside the feasible space is not easy. One way to overcome this issue is to replace $\mathbf{B}^T\cdot\mathbf{d}_{t,m}$ by an auxiliary variable $\mathbf{z}_{t,m}$ and add the corresponding equality constraint to (13):

$$\begin{aligned}\max_{\mathbf{d}_{t,m},\mathbf{z}_{t,m}} \quad & -\frac{1}{2}\cdot\left\|\mathbf{d}_{t,m}-\mathbf{g}_{t,m}\right\|_2^2+\frac{1}{2}\cdot\left\|\mathbf{g}_{t,m}\right\|_2^2-\delta_\infty^\lambda\left(\mathbf{z}_{t,m}\right)\\ \text{S.T.} \quad & \mathbf{z}_{t,m}=\mathbf{B}^T\cdot\mathbf{d}_{t,m}\end{aligned}. \quad (14)$$

This problem is now adequate to be solved by the Augmented Lagrangian Method (ALM) [20]-[21]. The method builds on top of the Augmented Lagrangian formulation of (14):

$$\min_{\boldsymbol{\alpha}_{t,m}}\ \max_{\mathbf{d}_{t,m},\mathbf{z}_{t,m}} \begin{aligned}&-\frac{1}{2}\cdot\left\|\mathbf{d}_{t,m}-\mathbf{g}_{t,m}\right\|_2^2+\frac{1}{2}\cdot\left\|\mathbf{g}_{t,m}\right\|_2^2-\delta_\infty^\lambda\left(\mathbf{z}_{t,m}\right)\\ &+\boldsymbol{\alpha}_{t,m}^T\cdot\left(\mathbf{z}_{t,m}-\mathbf{B}^T\mathbf{d}_{t,m}\right)-\frac{\eta}{2}\cdot\left\|\mathbf{z}_{t,m}-\mathbf{B}^T\mathbf{d}_{t,m}\right\|_2^2\end{aligned}, \quad (15)$$

where η is a penalty parameter and $\boldsymbol{\alpha}_{t,m}$ is the Lagrange multiplier of the equality constraint.

Incidentally, the duality theory is circular, in the sense that the Lagrange multiplier of the dual problem is in fact the primal variable. Therefore, the reason of applying ALM is to keep track of $\boldsymbol{\alpha}_{t,m}$ while solving the dual problem. ALM iteratively maximizes (15) jointly w.r.t. $\mathbf{d}_{t,m}$ and $\mathbf{z}_{t,m}$ and then updates $\boldsymbol{\alpha}_{t,m}$ with a proximal gradient descent. However, the joint maximization is again complicated by the box constraint $\delta_\infty^\lambda(\mathbf{z}_{t,m})$. A simple alternative is to update $\mathbf{d}_{t,m}$ while keeping $\mathbf{z}_{t,m}$ fixed and vice-versa. This approach, known as Alternating Direction Method [22], makes each maximization step easy to solve.

When solving for $\mathbf{z}_{t,m}$, Eq. (15) is reduced to:

$$\max_{\mathbf{z}_{t,m}} \quad -\delta_\infty^\lambda\left(\mathbf{z}_{t,m}\right)-\frac{\eta}{2}\cdot\left\|\mathbf{z}_{t,m}-\mathbf{B}^T\mathbf{d}_{t,m}-\frac{\boldsymbol{\alpha}_{t,m}}{\eta}\right\|_2^2. \quad (16)$$

Its solution can be expressed as:

$$\mathbf{z}_{t,m}^{(k+1)}=P_\infty^\lambda\left(\frac{\boldsymbol{\alpha}_{t,m}^{(k)}}{\eta}+\mathbf{B}^T\cdot\mathbf{d}_{t,m}^{(k)}\right), \quad (17)$$

where

$$P_\infty^\lambda\left(x\right)=\min\left(\left|x\right|,\lambda\right)\cdot\operatorname{sign}\left(x\right) \quad (18)$$

is an element-wise operator and the superscript (k) denotes a variable at the k-th iteration. When solving for $\mathbf{d}_{t,m}$, Eq. (15) is reduced to:

$$\max_{\mathbf{d}_{t,m}} \quad -\frac{1}{2}\cdot\left\|\mathbf{d}_{t,m}-\mathbf{g}_{t,m}\right\|_2^2-\frac{\eta}{2}\cdot\left\|\mathbf{z}_{t,m}-\mathbf{B}^T\mathbf{d}_{t,m}-\frac{\boldsymbol{\alpha}_{t,m}}{\eta}\right\|_2^2. \quad (19)$$

Its solution can be expressed as:

$$\mathbf{d}_{t,m}^{(k+1)}=\left(\mathbf{I}+\eta\cdot\mathbf{B}\mathbf{B}^T\right)^{-1}\cdot\left(\mathbf{g}_{t,m}-\mathbf{B}\cdot\left(\boldsymbol{\alpha}_{t,m}^{(k)}-\eta\cdot\mathbf{z}_{t,m}^{(k+1)}\right)\right), \quad (20)$$

where $\mathbf{I}$ denotes the identity matrix. Finally, the update of the Lagrange multiplier is obtained by solving the proximal operator:

$$\min_{\boldsymbol{\alpha}_{t,m}} \quad \boldsymbol{\alpha}_{t,m}^T\cdot\left(\mathbf{z}_{t,m}-\mathbf{B}^T\mathbf{d}_{t,m}\right)+\frac{1}{2\eta}\cdot\left\|\boldsymbol{\alpha}_{t,m}-\boldsymbol{\alpha}_{t,m}^{(k)}\right\|_2^2. \quad (21)$$

Its solution follows the closed-form expression:

$$\boldsymbol{\alpha}_{t,m}^{(k+1)}=\boldsymbol{\alpha}_{t,m}^{(k)}-\eta\cdot\left(\mathbf{z}_{t,m}^{(k+1)}-\mathbf{B}^T\cdot\mathbf{d}_{t,m}^{(k+1)}\right). \quad (22)$$

Algorithm 1 summarizes the major steps of DALM where we iteratively solve the sub-optimizations (17), (20) and (22) until reaching convergence.

Algorithm 1: Dual Augmented Lagrangian Method

1. Start from the L_1-norm optimization problem in (7) and its augmented Lagrangian formulation in (15).
2. Initialize η, $\mathbf{d}_{t,m}^{(0)}$ and $\boldsymbol{\alpha}_{t,m}^{(0)}$. Set $k = 0$.
3. Update $\mathbf{z}_{t,m}^{(k+1)}$ by using (17).
4. Update $\mathbf{d}_{t,m}^{(k+1)}$ by using (20).
5. Update $\boldsymbol{\alpha}_{t,m}^{(k+1)}$ by using (22).
6. Update $k = k + 1$.
7. Repeat Step 3~6 until reaching convergence.

IV. Implementation Details

To make the proposed DALM method efficient for practical applications, a number of implementation issues must be taken into account. In this section, we discuss these implementation details and highlight the novelty.

A. Fast Matrix Inverse

Studying Algorithm 1 reveals an important fact that the overall computational cost is dominated in general by Step 4 where we need to solve a linear equation in (20). Here, the matrix $\mathbf{B}$ is generated by DCT basis functions, as shown in (3). Hence, its rows are orthonormal [24] and the matrix

multiplication $\mathbf{B}\mathbf{B}^T$ in (20) is simply equal to an identity matrix. For this reason, Eq. (20) can be simplified as:

$$\mathbf{d}_{t,m}^{(k+1)} = \frac{1}{1+\eta} \cdot \left(\mathbf{g}_{t,m} - \mathbf{B} \cdot \left(\boldsymbol{\alpha}_{t,m}^{(k)} - \eta \cdot \mathbf{z}_{t,m}^{(k+1)} \right) \right). \quad (23)$$

Note that computing (23) is substantially less expensive than (20), since there is no need to solve any linear equation in (23).

B. Fast Matrix-Vector Multiplication

Given (17), (22) and (23), the computational cost is now dominated by the matrix-vector multiplications involving $\mathbf{B}$ and $\mathbf{B}^T$. Since the matrix $\mathbf{B}$ in (3) is defined by the DCT basis functions, we can use the fast DCT and IDCT algorithms to perform these operations, instead of explicitly calculating the matrix-vector multiplications. However, since $\mathbf{B}$ does not represent the full DCT or IDCT matrix, we must carefully process the data when applying the fast DCT or IDCT transform.

Let us denote $\mathbf{v} = B(\mathbf{u})$ as the fast 2-D IDCT transform, $\mathbf{u} = B^{-1}(\mathbf{v})$ as the fast 2-D DCT transform, and Ω as the set of spatial locations corresponding to the measured dies. As explained in Section II, the vector $\boldsymbol{\alpha}_{t,m}$ contains the 2-D DCT coefficients and the vector $\mathbf{g}_{t,m}$ contains the measurement data at the spatial locations belonging to Ω. The matrix $\mathbf{B}$ in (3), where $\mathbf{B} \cdot \boldsymbol{\alpha}_{t,m} \approx \mathbf{g}_{t,m}$, is created by down-sampling the full 2-D IDCT matrix where only the rows corresponding to the spatial locations in Ω are chosen. Likewise, the matrix $\mathbf{B}^T$ is created by down-sampling the full 2-D DCT matrix where only the columns corresponding to the spatial locations in Ω are selected.

To compute the matrix-vector multiplication $\mathbf{w} = \mathbf{B} \cdot \mathbf{u}$, we can first compute $\mathbf{v} = B(\mathbf{u})$ by the fast 2-D IDCT transform and then select the corresponding rows of $\mathbf{v}$ to form the vector $\mathbf{w}$:

$$\mathbf{w} = \mathbf{v}_{\Omega}, \quad (24)$$

where $\mathbf{v}_{\Omega}$ represents the elements $\{v_i;\ i \in \Omega\}$. On the other hand, to compute the matrix-vector multiplication $\mathbf{u} = \mathbf{B}^T \cdot \mathbf{w}$, we need to first create a vector $\mathbf{v}$:

$$\mathbf{v}_{\Omega} = \mathbf{w} \quad (25)$$

$$\mathbf{v}_{\bar{\Omega}} = \mathbf{0}, \quad (26)$$

where $\mathbf{v}_{\bar{\Omega}}$ represents the elements $\{v_i;\ i \notin \Omega\}$. Namely, for the indices belonging to the set Ω, the corresponding elements of $\mathbf{v}$ are equal to the vector $\mathbf{w}$. For the other indices, we simply fill in zeros for $\mathbf{v}$. Next, we apply the fast 2-D DCT transform $\mathbf{u} = B^{-1}(\mathbf{v})$ to calculate the matrix-vector multiplication $\mathbf{u} = \mathbf{B}^T \cdot \mathbf{w}$.

C. Test Flow

To apply VP with DALM for test cost reduction, we adopt the test flow that was proposed in [25]. It consists of two major steps: (i) pre-test analysis, and (ii) test application. During pre-test analysis, we physically measure all test items of all dies on one wafer. Based on the measurement data, our goal is to determine whether a test item is spatially correlated at the wafer level. The test item is considered to be "predictable", if a strong spatial correlation is observed. In this case, we further decide the number of dies that should be measured on a wafer to accurately predict the spatial wafer map associated with the test item by using VP. As such, the test item will only be physically measured at selected spatial locations and its values at other unmeasured locations are estimated by VP with sufficiently small escape rate and yield loss. In practice, we may repeat the pre-test analysis, if the spatial wafer maps vary due to wafer-to-wafer variations.

During test application, we first measure all test items at a subset of dies that are determined by pre-test analysis. Next, for the predictable test items, we apply VP to estimate their values for other unmeasured dies on the same wafer. Escape rate and yield loss are closely monitored during the prediction process. If the escape rate or the yield loss exceeds a pre-defined target for a specific test item, the test item is temporarily labelled as "unpredictable" for the current wafer.

Finally, the unpredictable test items are physically measured for the remaining dies on the wafer. Here, a test item is considered to be unpredictable, if it is classified as an unpredictable item during pre-test analysis or its escape rate or yield loss is not sufficiently small for the current wafer. In either case, the unpredictable test item must be measured for all dies on the current wafer. Algorithm 2 summarizes the major steps of the test flow. More details can be found in [25].

Algorithm 2: Test Cost Reduction by VP with DALM

1. Start from M wafers for a given circuit design.
2. Physically measure all dies from the first wafer. Perform pre-test analysis to determine the set of predictable test items $\{g_t;\ t \in \Phi\}$ and the optimal number of measured dies (say, N).
3. For $m = 2, 3, \cdots, M$
4. Initialize the set $\Theta = \{\}$.
5. Physically measure all test items from N randomly selected dies on the m-th wafer.
6. For each test item g_t where $t \in \Phi$, apply VP with DALM to predict the spatial variation $g_{t,m}(x, y)$ of the m-th wafer. Estimate the escape rate $ER_{t,m}$ and the yield loss $YL_{t,m}$. If $ER_{t,m}$ or $YL_{t,m}$ exceeds the pre-defined target, set $\Theta = \Theta \cup \{t\}$.
7. Physically measure the test items $\{g_t;\ t \notin \Phi \text{ or } t \in \Theta\}$ for all other dies on the m-th wafer.
8. Determine "pass" or "fail" for each die on the m-th wafer.
9. End For

V. Experimental Results

In this section, we demonstrate the efficiency of the proposed DALM method by using the wafer probe measurement data of an industrial RF transceiver. In total, nearly 1.2M dies are measured from 176 wafers of 9 lots where each wafer contains more than 6,000 dies. There are approximately 50 test items (e.g., bit error rate, power consumption, standby current, etc.) for this transceiver example.

For testing and comparison purposes, two different numerical solvers are implemented for spatial variation modeling: (i) the conventional interior-point method (IPM) [19], and (ii) the proposed DALM method (DALM). All numerical experiments are performed on a Linux server with 3.4 GHz CPU and 16 GB memory.

A. Spatial Variation Modeling

Fig. 1 shows the spatial variations for two different test items. Studying Fig. 1 reveals an important observation that a number of test items may not be spatially correlated (e.g., test item #1), while the other test items carry strong spatial correlations (e.g., test item #48). It, in turn, demonstrates the importance of our proposed pre-test analysis in Section IV.C where the objective is to identify the set of spatially correlated test items for test cost reduction.

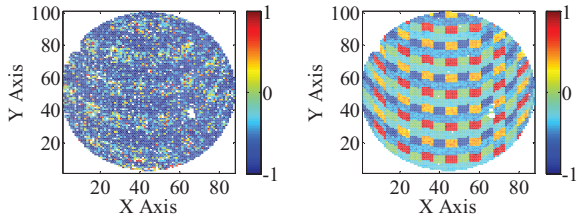

Fig. 1. Spatial variations (normalized) are shown for (a) test item #1 that is spatially uncorrelated, and (b) test item #48 that is spatially correlated.

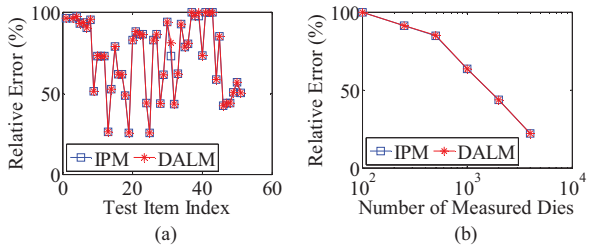

Fig. 2. Modeling error is compared between IPM and DALM for (a) all test items where 2000 dies are physically measured on a wafer, and (b) test item #48 where the number of physically measured dies varies from 100 to 4000.

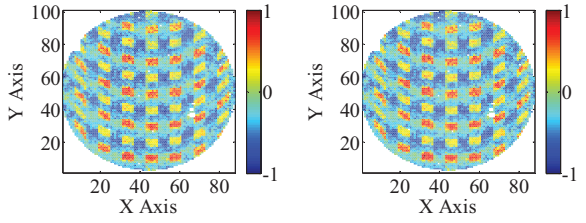

Fig. 3. Spatial variations (normalized) are predicted by (a) IPM and (b) DALM for test item #48 where 2000 dies are physically measured on a wafer.

TABLE I
COMPUTATIONAL TIME FOR IPM AND DALM TO MODEL THE SPATIAL VARIATION OF A SINGLE TEST ITEM FOR A GIVEN WAFER

Number of Measured Dies (N)	IPM	DALM	
	Runtime (Sec.)	Runtime (Sec.)	# of Iterations
100	48.3	12.2	7027
250	62.7	10.3	5664
500	84.7	8.9	5083
1000	119.9	8.1	4504
2000	171.2	7.3	3922
4000	255.2	6.7	3580

Fig. 2 compares the modeling error for IPM and DALM. In our experiment, the modeling error is defined as:

$$Error_{t,m} = \sqrt{\sum_n \left(g_{t,m}^{(n)} - \tilde{g}_{t,m}^{(n)}\right)^2 \Big/ \sum_n \left(g_{t,m}^{(n)}\right)^2}\ , \qquad (27)$$

where $g_{t,m}^{(n)}$ and $\tilde{g}_{t,m}^{(n)}$ denote the actual and predicted values of the t-th test item for the n-th die respectively, and the summation in (27) is calculated over all dies on the wafer. Fig. 3 further shows the spatial variations predicted by IPM and DALM. Note that the results in Fig. 2 and Fig. 3 are almost identical for both solvers, implying that the proposed DALM solver is as accurate as the conventional IPM.

Table I compares the computational time for IPM and DALM. Based on the results in Table I, two important observations can be made. First, DALM achieves up to 38× runtime speed-up over IPM. Hence, DALM is the preferred method for our test application where the spatial variations must be accurately estimated in real time during the testing process.

Second, and more interestingly, the computational time of DALM decreases with the number of measured dies (i.e., N). As N becomes larger, the underdetermined linear equation in (2) is better constrained and, hence, the DALM algorithm requires less number of iterations to converge, as shown in Table I. As a result, the computational time is reduced.

B. Test Application

TABLE II
PRE-TEST ANALYSIS RESULTS FOR IPM/DALM

Pre-defined Target for Escape Rate	5×10^{-3}
Pre-defined Target for Yield Loss	5×10^{-3}
Number of Measured Dies (N) per Wafer	2000
Number of Predictable Test Items	38

TABLE III
TEST COST REDUCTION BY IPM/DALM

	Full	IPM/DALM
Overall Test Cost	6.0×10^{7}	3.2×10^{7}
Escape Rate	—	1.2×10^{-3}
Yield Loss	—	2.0×10^{-3}

By applying DALM to test cost reduction, Table II summarizes the setup for our pre-test analysis and the corresponding results. Since both IPM and DALM give the same results in this example, we do not explicitly distinguish these two solvers in Table II.

Two important clarifications should be made here. First, a relatively large escape rate is allowed, because we focus on the application of wafer probe test and the "escaped" dies can be further captured during the final test after packaging. Second, most of the test items (i.e., 38 items) are considered to be predictable during our pre-test analysis. During test application, we will further monitor the escape rate and the yield loss for these 38 test items that are predictable. If the escape rate or the yield loss of a test item exceeds the pre-defined target in Table II for a specific wafer, the test item will be temporarily set as unpredictable for that wafer.

The proposed test flow (i.e., Algorithm 2) is applied to all wafers. Fig. 4 shows the total number of measured dies for each test item across all wafers. Here, two different cases are studied: (i) without test cost reduction (Full) and (ii) with test cost reduction (IPM/DALM). Note that IPM/DALM achieves significant cost reduction for most test items in this example. Table III summarizes the overall test cost, the escape rate and

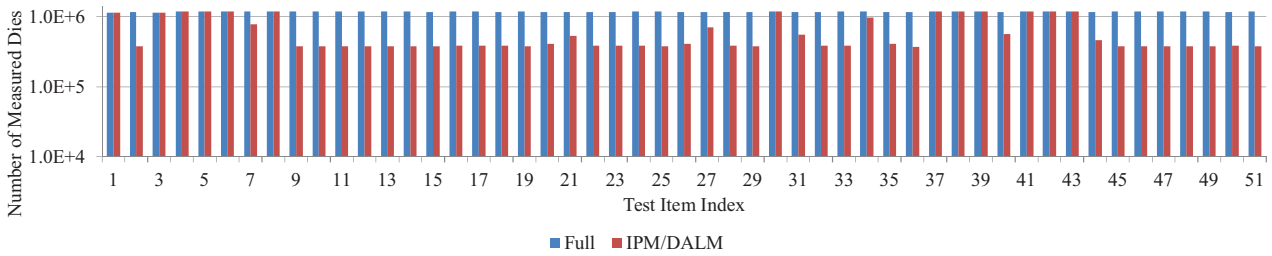

Fig. 4. The total number of measured dies across all wafers is shown for each test item without test cost reduction (Full) and with test cost reduction (IPM/DALM).

the yield loss. In this example, since the test time of each test item is not disclosed by our industrial collaborator, we simply use the total number of measurements to assess the test cost. Based on Table III, IPM/DALM achieves 1.875× reduction in test cost for this example.

VI. Conclusions

In this paper, a fast DALM method is described to efficiently solve the L_1-regularized least-squares regression problem for spatial variation modeling. DALM derives and then solves the dual formulation of the regression problem. Hence, it is computationally more efficient than directly solving the primal formulation by an interior-point method. Moreover, a number of fast numerical enhancements are further used to reduce the computational cost of DALM. Based on an industrial RF transceiver example where approximately 1.2M dies are measured from 176 wafers and 9 lots, DALM achieves up to 38× runtime speed-up over the conventional interior-point solver. In addition, the proposed test flow with DALM is able to reduce the test cost by 1.875×, while maintaining sufficiently small escape rate and yield loss. The proposed DALM method can be further applied to many other CAD problems (e.g., analog performance modeling) that involves L_1-regularized least-squares regression.

Acknowledgment

Support for this research was provided by the Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) through the Carnegie Mellon Portugal Program.

References

[1] K. Cheng and H. Chang, "Recent advances in analog, mixed-signal, and RF testing," *IPSJ T-SLDM*, vol. 3, pp. 19-46, Feb. 2010.

[2] K. Arabi, "Mixed-signal test impact to SoC commercialization," *IEEE VTS*, 2010.

[3] P. Variyam, S. Cherubal and A. Chatterjee, "Prediction of analog performance parameters using fast transient testing," *IEEE T-CAD*, vol. 21, no. 3, pp. 349-361, Feb. 2002.

[4] H. Stratigopoulos, P. Drineas, M. Slamani and Y. Makris, "Non-RF to RF test correlation using learning machines: A case study," *IEEE VTS*, 2007.

[5] R. Voorakaranam, S. Akbay, S. Bhattacharya, S. Cherubal and A. Chatterjee, "Signature testing of analog and RF circuits: algorithms and methodology," *IEEE T-CAS-I*, vol. 54, no. 5, pp. 1018-1031, May. 2007.

[6] H. Stratigopoulos and Y. Makris, "Error moderation in low-cost machine learning-based analog/RF testing," *IEEE T-CAD*, vol. 27, no. 2, pp. 339-351, Feb. 2008.

[7] H. Stratigopoulos, P. Drineas, M. Slamani and Y. Makris, "RF specification test compaction using learning machines," *IEEE T-VLSI*, vol. 18, no. 6, pp. 998-1002, Jun. 2010.

[8] E. Yilmaz, S. Ozev and K. Butler, "Per-device adaptive test for analog/RF circuits using entropy-based process monitoring," *IEEE T-VLSI*, vol. 21, no. 6, pp. 1116-1128, Jun. 2013.

[9] X. Li, R. Rutenbar and R. Blanton, "Virtual probe: a statistically optimal framework for minimum-cost silicon characterization of nanoscale integrated circuits," *IEEE ICCAD*, pp. 433-440, 2009.

[10] W. Zhang, X. Li and R. Rutenbar, "Bayesian virtual probe: minimizing variation characterization cost for nanoscale IC technologies via Bayesian inference," *IEEE DAC*, pp. 262-267, 2010.

[11] W. Zhang, X. Li, E. Acar, F. Liu and R. Rutenbar, "Multi-wafer virtual probe: minimum-cost variation characterization by exploring wafer-to-wafer correlation," *IEEE ICCAD*, pp. 47-54, 2010.

[12] H. Chang, K. Cheng, W. Zhang, X. Li and K. Butler, "Test cost reduction through performance prediction using virtual probe," *IEEE ITC*, 2011.

[13] W. Zhang, X. Li, F. Liu, E. Acar, R. Rutenbar and R. Blanton, "Virtual probe: a statistical framework for low-cost silicon characterization of nanoscale integrated circuits," *IEEE T-CAD*, vol. 30, no. 12, pp. 1814-1827, Dec. 2011.

[14] C. Hsu, F. Lin, K. Cheng, W. Zhang, X. Li, J. Carulli and K. Butler, "Test data analytics - exploring spatial and test-item correlations in production test data," *IEEE ITC*, 2013.

[15] N. Kupp, K. Huang, J. Carulli and Y. Makris, "Spatial estimation of wafer measurement parameters using Gaussian process models," *IEEE ITC*, 2012.

[16] N. Kupp, K. Huang, J. Carulli and Y. Makris, "Spatial correlation modeling for probe test cost reduction in RF devices," *IEEE ICCAD*, pp. 23-29, 2012.

[17] K. Huang, N. Kupp, J. Carulli and Y. Makris, "Handling discontinuous effects in modeling spatial correlation of wafer-level analog/RF tests," *IEEE DATE*, pp. 553-558, 2013.

[18] K. Huang, N. Kupp, J. Carulli and Y. Makris, "On combining alternate test with spatial correlation modeling in analog/RF ICs," *IEEE ETS*, 2013.

[19] S. Kim, K. Lustig, S. Boyd and D. Gorinevsky, "An interior-point method for large-scale l_1-regularized least squares," *IEEE J-STSP*, vol. 1, no. 4, pp. 606-617, Dec. 2007.

[20] R. Tomioka and M. Sugiyama, "Dual-augmented Lagrangian method for efficient sparse reconstruction," *IEEE SPL*, vol. 16, no. 12, pp. 1067-1070, Dec. 2009.

[21] R. Tomioka, T. Suzuki and M. Sugiyama, "Super-linear convergence of dual augmented Lagrangian algorithm for sparsity regularized estimation," *JMRL*, vol. 12, pp. 1537-1586, May 2011.

[22] J. Yang and Y. Zhang, "Alternating direction algorithms for l1-problems in compressive sensing", *SIAM J. Sci. Comput.*, vol. 33, no. 1, pp. 250–278, 2011.

[23] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2009.

[24] R. Gonzalez and R. Woods, *Digital Image Processing*, Prentice Hall, 2007.

[25] S. Zhang, X. Li, R. Blanton, J. Silva, J. Carulli and K. Butler, "Bayesian model fusion: enabling test cost reduction of analog/RF circuits via wafer-level spatial variation modeling," *IEEE ITC*, 2014.