SIAM J. NUMER. ANAL.
Vol. 46, No. 3, pp. 1483–1502

UNIFORM CONVERGENCE OF A NONLINEAR ENERGY-BASED MULTILEVEL QUANTIZATION SCHEME*

QIANG DU† AND MARIA EMELIANENKO‡

Abstract. A popular vector quantization scheme can be constructed by centroidal Voronoi tessellations (CVTs) which also have many other applications in diverse areas of science and engineering. The development of efficient algorithms for their construction is a key to the successful applications of CVTs in practice. This paper studies the details of a new optimization-based multilevel algorithm for the numerical computation of CVTs. The rigorous proof of its uniform convergence in one space dimension and the results of computational simulations are provided. They substantiate recent claims on the significant speedup demonstrated by the new scheme in comparison with traditional methods.

Key words. optimal quantization, centroidal Voronoi tessellations, Lloyd's method, multilevel method, uniform convergence

AMS subject classifications. 65D18, 65H10, 65Y20, 68U10

DOI. 10.1137/050648699

1. Introduction.

Optimal vector quantization is used in many applications such as image and data compression, pattern recognition, and image rendering [24]. A vector quantizer maps N-dimensional vectors in the domain $\Omega \subset \mathbb{R}^N$ into a finite set of vectors $\{\mathbf{z}_i\}_{i=1}^k$. Each vector $\mathbf{z}_i$ is called a code vector or a *codeword*, and the set of all the codewords is called a codebook. A special quantization scheme is given by the Voronoi tessellation which associates with each codeword $\mathbf{z}_i$ (also called a *generator*) a nearest neighbor region that is called a Voronoi region $\{V_i\}_{i=1}^k$. That is, for each i, V_i consists of all points in the domain Ω that are closer to $\mathbf{z}_i$ than to all the other generating points, and a Voronoi tessellation refers to the tessellation of a given domain by the Voronoi regions $\{V_i\}_{i=1}^k$ associated with a set of given generating points $\{\mathbf{z}_i\}_{i=1}^k \subset \Omega$ [1, 34].

With a suitably defined distortion measure, an optimal quantizer can be described as a centroidal Voronoi tessellation. For a given density function ρ defined on Ω, we may define the centroids, or mass centers, of regions $\{V_i\}_{i=1}^k$ by

$$\mathbf{z}_i^* = \Big(\int_{V_i} \mathbf{y}\rho(\mathbf{y})\,d\mathbf{y}\Big)\Big(\int_{V_i} \rho(\mathbf{y})\,d\mathbf{y}\Big)^{-1}. \tag{1.1}$$

Then an *optimal quantization* may be constructed through a *centroidal Voronoi tessellation* (CVT) for which the generators of the Voronoi tessellation are the centroids of their respective Voronoi regions; in other words, $\mathbf{z}_i = \mathbf{z}_i^*$ for all i. Such a connection between CVTs and optimal quantization schemes has been explored extensively in the literature [13].

*Received by the editors January 3, 2006; accepted for publication (in revised form) December 17, 2007; published electronically April 4, 2008. This research was supported in part by the U.S. NSF through grants DMS-0409297 and ITR-0205232.

http://www.siam.org/journals/sinum/46-3/64869.html

†Department of Mathematics, Pennsylvania State University, University Park, PA 16802 (qdu@math.psu.edu).

‡Department of Mathematical Sciences, George Mason University, Fairfax, VA 22030 (memelian@gmu.edu).

Given a set of points $\{\mathbf{z}_i\}_{i=1}^k$ and a tessellation $\{V_i\}_{i=1}^k$ of the domain, we may define the *energy functional* or the *distortion value* for the pair $(\{\mathbf{z}_i\}_{i=1}^k, \{V_i\}_{i=1}^k)$ by

$$\mathcal{G}\Big(\{\mathbf{z}_i\}_{i=1}^k, \{V_i\}_{i=1}^k\Big) = \sum_{i=1}^k \int_{V_i} \rho(\mathbf{y})|\mathbf{y}-\mathbf{z}_i|^2\, d\mathbf{y}\,. \tag{1.2}$$

The minimizer of $\mathcal{G}$, that is, the optimal quantizer, necessarily forms a CVT which illustrates the optimization property of the CVT [13]. The terms optimal quantizer and CVT are thus to be used interchangeably in what follows. For more studies on optimal quantization schemes, we refer to [23, 24, 38]. We note that, besides providing an optimal least squares vector quantizer design in electrical engineering applications, the CVT concept also has applications in diverse areas such as astronomy, biology, image and data analysis, resource optimization, sensor networks, geometric design, and numerical partial differential equations [5, 8, 13, 14, 15, 16, 17, 18, 26, 27, 33, 39, 40]. We refer to [13] for a more comprehensive review of the mathematical theory and diverse applications of CVTs.

In the seminal work of Lloyd on least squares quantization [32], one of the algorithms proposed for computing optimal quantizers is a fixed point–type iterative algorithm consisting of the following simple steps: Starting from an initial quantization (a Voronoi tessellation corresponding to an old set of generators), a new set of generators is defined by the mass centers of the Voronoi regions. This process is continued until a certain stopping criterion is met. For algorithms on the computation of Voronoi tessellations, we refer to [1, 19, 21]. It is easy to see that the Lloyd algorithm is an energy descent iteration of the energy functional (1.2), which gives strong indications to its practical convergence. We refer to [11, 20] for some discussions on the recent development of a rigorous convergence theory.

Lloyd's algorithms and their variants have been proposed and studied in many contexts for different applications [15, 24, 31, 36]. A particular extension using parallel and probabilistic sampling was given in [27] which allows efficient and mesh-free implementation of the Lloyd algorithm. Still, the Lloyd algorithm is at best linearly convergent. Moreover, it slows down as the number of generators increases.

For modern applications of the CVT concept in large scale scientific and engineering problems such as data communication, vector quantization, and mesh generation, efficient algorithms for computing the CVTs play crucial roles.

Recently, a new multilevel approach to the optimal quantization problem has been developed [10, 12]. The new multilevel scheme offers considerable speedup over traditional methods such as the celebrated Lloyd iteration. It can be combined with other techniques to accelerate the computation of CVTs and the optimal vector quantizers. In this paper, we present a rigorous mathematical theory for the new algorithm. We focus on the main characteristics of this multilevel scheme, namely, the uniform convergence with respect to the problem size. Though the result is shown in one space dimension only, such a result is the first of its type in the vector quantization field and on the computation of CVTs. Proofs for higher-dimensional cases are more involved, and they are now being worked out in our ongoing study [12].

The rest of the paper is organized as follows. In section 2, the optimization-based nonlinear multilevel algorithm is introduced. In section 3, the uniform convergence theory is established in the one-dimensional case. Numerical results demonstrating its superiority over traditional methods are given in section 4. Final conclusions are made in section 5.

2. Optimization-based nonlinear multilevel algorithm. Since the original concept of CVTs is related to the solution of a nonlinear optimization problem, and the monotone energy descent property is preserved by the Lloyd iteration [13], we may thus investigate whether monotone energy reduction can be achieved in a multilevel procedure which would also improve the performance of the simple-minded iteration.

The problem of constructing a CVT is nonlinear in nature, and hence standard linear multigrid theory cannot be directly applied. There are still several ways one could implement a nonlinear multilevel scheme in this context (see [9, 10, 28, 29]). A Newton-type acceleration method, studied earlier in [9], is based on some global linearization as the outer loop, coupled with other fast solvers in the inner loop. Alternatively, we now study an approach that overcomes the difficulties of the nonlinearity by essentially relying on direct energy minimization without any type of global linearization.

The optimization formulation of CVTs is not new [13]. CVTs and optimal quantizers are known to be closely related to the problem of minimizing the functional $\mathcal{G}\big(\{\mathbf{z}_i\}_{i=1}^k, \{V_i\}_{i=1}^k\big)$ with $\{V_i\}_{i=1}^k$ forming the Voronoi tessellation corresponding to the generators $\{\mathbf{z}_i\}_{i=1}^k$. In fact, we may note that the vector of generators of a CVT forms a critical point of the energy, and vice versa [13]. This is one of the important characterizations of CVTs which will be used in the later discussion. In our study here, we follow the ideas presented in the literature on the extension of multigrid ideas to nonlinear optimization problems (see [6, 22, 25, 30, 37] and the references cited therein). However, we observe that a direct application of multilevel schemes to the minimization of the original energy is not the best strategy. We illustrate later in this section that an alternative minimization formulation can be introduced to make the problem of computing CVTs more amenable for the multilevel framework.

2.1. Preconditioned formulation. Let $\mathbf{V} = \{V_i\}_{i=1}^k$ denote a Voronoi tessellation corresponding to some choice of iteration points $\mathbf{Z}$ in Ω. For any Voronoi cell $V_i \subset \Omega$, let us denote by R_i the mass of V_i, i.e.,

$$R_i := \int_{V_i} \rho(\mathbf{y})\, \mathrm{d}\mathbf{y}.$$

The fact that ρ is positive almost everywhere implies that $R_i > 0$ for all $i = 1, \ldots, k$.

We define the set of iteration points $\mathbf{W}$ by

$$\mathbf{W} = \{(w_i)|_{i=0}^{k+1} | \quad 0 = w_0 \le w_i \le w_{i+1} \le w_{k+1} = 1 \quad \forall\, 0 \le i \le k\}\,, \tag{2.1}$$

which includes the generators and the end points. The so-called Lloyd map is the function $\mathbf{T} : \mathbf{W} \to \mathbf{W}$ that takes a tuple of generators into the tuple of the centroids of its Voronoi regions (exactly the iteration used in the Lloyd algorithm). More precisely, let $\mathbf{T} = (T_1, \ldots, T_k)$; then for any $z \in \mathbf{W}$,

$$T_i(z) = \frac{\int_{V_i(z)} \mathbf{y}\rho(\mathbf{y})\, \mathrm{d}\mathbf{y}}{\int_{V_i(z)} \rho(\mathbf{y})\, \mathrm{d}\mathbf{y}}$$

for $i = 1, \ldots, k$. The Lloyd map is continuous from $\mathbf{W}$ to $\mathbf{W}$ [13, 11].

For any $z \in \mathbf{W}$, the *Lloyd energy* is defined as

$$\mathcal{H}(z) = \mathcal{G}(z, \mathbf{V}(z)) = \sum_{i=0}^{n} \int_{V_i(z)} \rho(\mathbf{y})|\mathbf{y} - z_i|^2\, \mathrm{d}\mathbf{y}.$$

The Lloyd energy function is also continuous on $\mathbf{W}$. The derivative of $\mathcal{H}$ is given [13] by

$$\frac{\partial \mathcal{H}}{\partial z_i}(z) = 2R(V_i(z))(z_i - \mathbf{T}_i(z)) \,. \tag{2.2}$$

To perform rigorous analysis of our nonlinear multigrid scheme, it is desirable if the energy functional is locally convex and satisfies some spectral equivalence properties in the appropriate norm. In general, we cannot expect these properties to hold. However, relating our problem to an equivalent optimization problem through a technique that mimics the role of a dynamic nonlinear preconditioner turns out to be very effective. More precisely, denote by P a nonsingular matrix playing the role of a preconditioner. It is easy to deduce that $P\nabla\mathcal{H} = 0$ at a CVT. Hence we arrive at an equivalent formulation of the minimization problem: $\min ||P\nabla\mathcal{H}||^2$, with respect to the standard Euclidean norm. In particular, we can choose $P = (\mathbf{I} - S)^{-1/2}R^{-1}$, where R is a diagonal matrix $R = \mathtt{diag}(R_1, R_2, \ldots, R_k)$, where, as before, the R_i's are the masses of the corresponding Voronoi cells. As we will show later, this preconditioning strategy renders the problem convex and shares some spectral equivalence properties in an appropriately chosen norm for a class of densities, say of the form $\rho(x) = 1 + \epsilon g(x)$. For such densities, the energy functional $\mathcal{H}$ may not be locally convex; however, the equivalent preconditioned strategy successfully eliminates this difficulty. Notice that for this type of density we have $\frac{\partial \mathbf{T}}{\partial z} = S + \epsilon\widetilde{S}(z)$, where

$$S = \begin{pmatrix} 1/4 & 1/4 & 0 & & 0 & 0 & 0 \\ 1/4 & 1/2 & 1/4 & \ldots & 0 & 0 & 0 \\ 0 & 1/4 & 1/2 & & 0 & 0 & 0 \\ & \ldots & & \ldots & & \ldots & \\ 0 & 0 & 0 & & 1/2 & 1/4 & 0 \\ 0 & 0 & 0 & \ldots & 1/4 & 1/2 & 1/4 \\ 0 & 0 & 0 & & 0 & 1/4 & 1/4 \end{pmatrix}$$

is the derivative of the Lloyd map corresponding to the case of the constant density. This preconditioned formulation leads to a multilevel algorithm based on the following nonlinear optimization problem:

$$\min_{z\in\mathbf{W}} \widetilde{\mathcal{H}}(z), \text{ where } \widetilde{\mathcal{H}}(z = \{z_i\}_{i=0}^{k+1}) = ||P\nabla\mathcal{H}(\{\mathbf{z}_i\}_{i=1}^{k})||^2 \,. \tag{2.3}$$

The functional $\widetilde{\mathcal{H}}$ may be regarded as a preconditioned energy, and with the above choice of the preconditioner P, we have

$$\widetilde{\mathcal{H}}(z) = ||2(\mathbf{I} - S)^{-1/2}R^{-1}R(z - \mathbf{T}(z))||^2 = 2[z - \mathbf{T}(z)]^T[(\mathbf{I} - S)^{-1}(z - \mathbf{T}(z))] \tag{2.4}$$

and the gradient can be computed as

$$\begin{aligned} \nabla\widetilde{\mathcal{H}}(z) &= 4\left(\mathbf{I} - \frac{\partial \mathbf{T}}{\partial z}\right)(\mathbf{I} - S)^{-1}(z - \mathbf{T}(z)) = 4(\mathbf{I} - \epsilon\widehat{S}(z))(z - \mathbf{T}(z)) \\ &= 4(z - \mathbf{T}(z)) - 4\epsilon\beta(z), \end{aligned} \tag{2.5}$$

where $\widehat{S}(z) = \widetilde{S}(z)(I - S)^{-1}$ and $\beta(z) = \widehat{S}(z)(z - \mathbf{T}(z))$.

A key observation is that as R varies with respect to the generators, the above transformation or *dynamic preconditioning* keeps the modified objective functional convex in a suitably large neighborhood of the minimizer and therefore makes the new formulation more amenable to analysis than the original problem.

2.2. Space decomposition. Suppose that without loss of generality we start with an odd number of generators $k = 2^{n_J} - 1$. Now, we define the set of iteration points $\mathbf{W}$ by (2.1).

While the set of iteration points will be dynamically updated at each step of the algorithm, we want to construct a domain decomposition that will not be affected by these changes. Namely, we can denote by $\mathbf{X} = \mathbf{X}^J$ the set of uniformly distributed grid nodes in the unit interval, and by $\mathbf{X}^i$ the sequence of grid nodes $\mathbf{X}^i = \{x_j^i\}_{j=0}^{n_i}$, obtained via uniform subdivision $x_j^i = j/2^{n_i}$. A set of nonoverlapping grid points at the ith level would consist of $\overline{\mathbf{X}}^i = \mathbf{X}^i/\mathbf{X}^{i-1}$. Next we define by $\mathcal{T} = \mathcal{T}_J$ a mesh corresponding to $\mathbf{X}$ and consider a sequence of nested uniform finite element meshes $\mathcal{T}_1 \subset \mathcal{T}_2 \subset \cdots \mathcal{T}_J$, where $\mathcal{T}_i$ consists of all finite element meshes $\{\tau_j^i\}_{j=1}^{n_i-1}$ with mesh parameters $h_i = h_{i-1}/2$, such that $\cup_{j=1}^{n_i} \tau_j^i = \Omega = [0,1]$ and $n_i = 2n_{i-1}$. Corresponding to each partition $\mathcal{T}_i$ $(i = 1, \ldots, J)$, there is a finite element space $\mathbf{Y}_i$ defined by

$$\mathbf{Y}_i = \{v \in H_0^1(\Omega) \ \mid \ v|_\tau \in \mathcal{P}_1(\tau) \ \forall \tau \in \mathcal{T}_i\} \ .$$

For each $\mathbf{Y}_i$, there corresponds a nodal basis $\{\psi_j^i\}_{j=1}^{n_i}$, such that $\psi_j^i(x_k^i) = \delta_{jk}$, where δ_{jk} is the usual Kronecker delta function and $\{x_k^i\}_{k=1}^{n_i}$ is the set of all nodes of the elements of $\mathcal{T}_i$ with $x_1^J = 0, x_{n_J}^J = 1$. Define the corresponding one-dimensional subspaces $\mathbf{Y}_{i,j} = \text{span}\{\psi_j^i\}$. Then the decomposition can be regarded as

$$\mathbf{Y}_J = \sum_{i=1}^{J} \sum_{j=1}^{n_i} \mathbf{Y}_{i,j} = \bigoplus_{i=1}^{J} \overline{\mathbf{Y}}_i,$$

where $\overline{\mathbf{Y}}_i = \mathbf{Y}_i/\mathbf{Y}_{i-1}$ for $i > 1$ and $\overline{\mathbf{Y}}_1 = \mathbf{Y}_1$, following the standard multigrid construction given in [4]. Now clearly for each $\psi_j^i \in \mathbf{Y}_i$, we can find a vector $w_j^i = \{\bar{\psi}_{jm}^i\}_{m=1}^{n_J-1} \in \mathbb{R}^{n_J}$, such that $\psi_j^i(x) = \sum_{m=1}^{n_J-1} \bar{\psi}_{jm}^i \psi_m^J(x)$ for $x \in \Omega$.

Figure 2.1 gives an illustration of this decomposition. Notice that this configuration is independent of the positions of generators and relies entirely on the grids $\mathcal{T}_i$ that are chosen to be uniform.

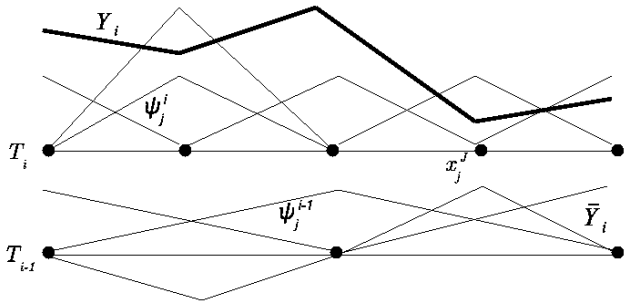

FIG. 2.1. *The structure of the space decomposition subordinate to the uniform mesh* $\mathbf{X}$ *in* $[0,1]$.

Using the linear interpolation coefficients $w_j^i = \{\bar{\psi}_{jm}^i\}_{m=1}^{n_j} \in \mathbb{R}^{n_J}$ defined above, we can decompose the space of iteration points as follows:

$$\mathbf{W} = \sum_{i=1}^{J} \mathbf{W}_i = \bigoplus_{i=1}^{J} \overline{\mathbf{W}}_i,$$

where $\mathbf{W}_i = \text{span}\{w_j^i\}$ and $\overline{\mathbf{W}}_i = \mathbf{W}_i/\mathbf{W}_{i-1}$ for $i > 1$. The vector spaces defined this way provide all the information necessary to fully describe the iteration procedure given below.

Note that once a mesh coarsening procedure is described, the above setup naturally applies to higher-dimensional cases, where $\mathbf{X}$ consists of grids points in a higher-dimensional domain and is discussed in more detail in [12]. Here we will restrict our attention to the one-dimensional case. It may then be noted that the set of basis functions $[w_1^i, \ldots, w_{n_i}^i]^T \in \mathbb{R}^{n_i \times k}$ used at each iteration can be pregenerated using the recursive procedure $[w_1^J, \ldots, w_{nJ}^J]^T = I_{k\times k}$ and $[w_1^{J-s}, \ldots, w_{nJ-s}^{J-s}]^T = (\Pi_{i=1}^s P_{J-i})$, where P_i is the basis transformation from space $\mathbf{W}_{i+1}$ to $\mathbf{W}_i$ which plays the role of a restriction operator.

2.3. Description of the algorithm. Using the above notation, we design Algorithm 2.2, which is a multilevel successive subspace correction algorithm [6, 41]. Each step of the procedure outlined below involves solving a system of nonlinear equations, which plays the role of relaxation. We can use the Newton iteration to solve this nonlinear system. Solution at current iterate is updated after each nonlinear solve by the Gauss–Seidel-type procedure, and hence the resulting scheme is successive in nature. The algorithm uses the procedure *CoarseGridSolve*, which, as the name indicates, refers to finding the solution at the coarsest level. In our implementation, this procedure consists of applying the Lloyd method to the $2^m - 1$ coarsest generators for a few steps (given by $imax$) or until saturation. In the implementation, we mostly use $m = 2$ so that there are only three coarsest generators.

ALGORITHM 2.1. $\mathbf{U}$ = CoarseGridSolve($\mathbf{U}$,tol).
Input: A set of k iteration points $\mathbf{U} = \{u_i\}_{i=1}^k$ *on* Ω.
Method:
Define the set of coarsest grid nodes: $z = \{z_j\}_{j=1}^{2^m-1}$, *where* $z_j = u_{\frac{j(k+1)}{2^m}}$.
For `iter` $= 1 : imax$
Perform Lloyd iterations with $2^m - 1$ *generators:*
1. *Construct the Voronoi regions* $\{V_j(z)\}_{j=1}^{2^m-1}$ *of* Ω *associated with z;*
2. *Determine the centroids* $\mathbf{T}(z)$ *of the Voronoi regions* $V_j(z)$*;*
these centroids form the new set of points z;
3. *If* $||z - \mathbf{T}(z)|| <$ tol, *return z and terminate the cycle;*
otherwise, goto step 1.

endfor.
Update $\mathbf{U}$ *by* $u_{\frac{j(k+1)}{2^m}} = z_j, j = 1, 2, \ldots, 2^m - 1$.

In general, other efficient optimization methods, as well as Newton's method, can be used in order to quickly damp the error, since the number of unknowns on the coarsest grid remains relatively small.

The "slash" cycle can be defined as follows.

ALGORITHM 2.2. Successive correction $V(\nu)$ scheme.
Input:
k, number of generators; $u_1 = \{z_i\}_{i=1}^k$, *the ends plus the initial set of generators.*
tol, *tolerance for the coarse grid solution accuracy.*
Output after n cycles:
$u_n = \{z_i\}_{i=0}^{k+1}$, *the ends plus the set of generators for CVT.*
Method: For $n > 1$, *given* u_n, *do*
1. *For* $i = 1 : J$
$\bar{u}_{n+\frac{i-1}{J}} = u_{n+\frac{i-1}{J}}$

Perform ν smoothings:

$\bar{u}_{n+\frac{i-1}{J}} = \bar{u}_{n+\frac{i-1}{J}} + \alpha^0_{j\,l} w^i_j \in \overline{\mathbf{W}}_i$ *sequentially for* $1 \le j \le n_i$,

such that $\widetilde{\mathcal{H}}(\bar{u}_{n+\frac{i-1}{J}} + \alpha^0_{j\,l} w^i_j) = \min_{\alpha_{jl}} \widetilde{\mathcal{H}}(\bar{u}_{n+\frac{i-1}{J}} + \alpha_{j\,l} w^i_j)$,

Update the solution at ith level:

$$u_{n+\frac{i}{J}} = \bar{u}_{n+\frac{i-1}{J}} = u_{n+\frac{i-1}{J}} + e^i_n, \text{ where } e^i_n = \sum_{l=1}^{\nu_1} \sum_{j=1}^{n_i - 1} \alpha^0_{j\,l} w^i_j$$

endfor.

2. *On the coarsest level,* $u_{n+1} \leftarrow$*CoarseGridSolve(*u_{n+1},tol*).*
3. *Set* $n = n + 1$ *repeat procedure* 1 *to* 2 *until some stopping criterion is met.*

Here, the parameter ν denotes the number of Gauss–Seidel iterations (smoothings) used at each level. Although it is enough to have $\nu = 1$ in theory, larger values are used in practice due to the inexact nonlinear solver. The values $\nu \le 3$ usually suffice for the optimization to reach saturation. In the above description, no postsmoothings are used to make the analysis more transparent. A complete $V(\nu_1, \nu_2)$ cycle with ν_1 presmoothings and ν_2 postsmoothings can be defined and analyzed in a similar fashion. The algorithm essentially depends only on the proper space decompositions and its relation to the set of generators, and thus it is applicable in any dimension. The more general forms, including detailed description of the grid coarsening procedure, will be discussed in our subsequent works.

3. The uniform convergence theory. The uniform convergence of the new multilevel scheme can be rigorously proved, at least, for a large class of density functions in one-dimensional space. This is the first step toward a more comprehensive theoretical analysis of this type of multilevel scheme in general. Here, let us first establish some important properties of the energy functional defined in (2.3). Then we introduce our main convergence results. Without further notice, all the analysis in this section is restricted to the one-dimensional case only, and we further assume without loss of generality that the domain is simply the unit interval $\Omega = (0, 1)$, which can always be transformed into any other intervals.

3.1. Notation. First, for each $u, v \in \mathbf{W}$, we adopt the following H^1 inner product:

$$\langle u, v \rangle_{\mathbf{W}} = \frac{1}{k} \sum_{i=1}^{k+1} (u_i - u_{i-1})(v_i - v_{i-1}).$$

Next, if $y = u - v$, we have $y_0 = y_{k+1} = 0$ and

$$||y||^2_{\mathbf{W}} = \frac{1}{k} \sum_{i=1}^{k+1} (y_i - y_{i-1})^2$$

defines a $\mathbf{W}$-norm that we will be using throughout the rest of the paper.

To simplify the presentation, let us introduce the following notation: for $i = 1, \ldots, k$,

$$u^-_i = \frac{u_i + u_{i-1}}{2}, \qquad u^+_{i-1} = \frac{u_{i-1} + u_i}{2}$$

with $a_i = u_i - u_{i-1}, b_i = w_i - w_{i-1}, x_i = u_i - w_i$, and $u_0 = w_0 = 0$, $u_{k+1} = w_{k+1} = 1$ being the fixed ends of the interval.

3.2. Technical lemmas. In the discussion that follows we say that a functional F satisfies the *convexity* and *continuity* properties in $\mathbf{W}$ if there exist constants $K > 0$, $L > 0$ satisfying

$$(\nabla F(w) - \nabla F(v), w - v) \geq K||w - v||^2_{\mathbf{W}} \quad \forall w, v \in \mathbf{W} \,, \tag{3.1}$$

$$(\nabla F(w) - \nabla F(v), w - v) \leq L||w - v||^2_{\mathbf{W}} \quad \forall w, v \in \mathbf{W} \,, \tag{3.2}$$

where $(\cdot, \cdot)$ denotes the regular Euclidean inner product.

Let us first turn our attention to the case of a constant density. In this simple case, with the above notation, we get the following result for the preconditioned energy functional.

PROPOSITION 3.1. *Let $\rho(x) = 1$ be the density function on $[0, 1]$. Then the following relation holds for all $u, w \in \mathbf{W}$:*

$$(\nabla\widetilde{\mathcal{H}}(u) - \nabla\widetilde{\mathcal{H}}(w), u - w) = \frac{1}{2}\sum_{i=1}^{k+1}(a_i - b_i)^2 \,.$$

Proof. Following (2.5), we have

$$\frac{\partial\widetilde{\mathcal{H}}}{\partial u_i} = 2(u_i - T_i) = 2\Big(u_i - \frac{u_i^+ + u_i^-}{2}\Big) = \frac{1}{2}(a_i - a_{i+1}) \,.$$

Thus,

$$\begin{aligned}(\nabla\widetilde{\mathcal{H}}(u) - \nabla\widetilde{\mathcal{H}}(w), u - w) &= \frac{1}{2}\sum_{i=1}^{k}(u_i - w_i)(a_i - a_{i+1} - b_i + b_{i+1}) \\ = \frac{1}{2}\sum_{i=1}^{k}(u_i - w_i)(a_i - b_i) - \frac{1}{2}\sum_{i=1}^{k}(u_{i-1} - w_{i-1})(a_i - b_i) &= \frac{1}{2}\sum_{i=1}^{k+1}(a_i - b_i)^2 \,. \qquad \square\end{aligned}$$

COROLLARY 3.1. *For a constant density, the functional $F = k^{-1}\widetilde{\mathcal{H}}$ satisfies the continuity and convexity conditions with $K = L = 1/2$ for all points in $\mathbf{W}$.*

Note that this is a simple consequence of the fact $\sum_{i=1}^{k+1}(a_i - b_i)^2 = k||u - w||^2_{\mathbf{W}}$.

One can extend the above result to a broader class of density functions. First, the following auxiliary lemma can be verified.

LEMMA 3.2. *Let $M_g = \sup_{x\in[0,1]}|g(x)|$ and $M' = \sup_{x\in[0,1]}|g'(x)|$ on $[0, 1]$, and define*

$$Q_i(u) = \frac{\int_{u_i^-}^{u_i^+}(2x - (u_i^+ + u_i^-))g(x)dx}{(u_i^+ - u_i^-) + \epsilon\int_{u_i^-}^{u_i^+} g(x)dx} \,. \tag{3.3}$$

Then

$$|Q_i(u)| \leq \frac{4M_g|u_i^+ - u_i^-|}{1 - 2\epsilon M_g} \tag{3.4}$$

and

$$|Q_i(u) - Q_i(w)| \leq C_Q\alpha_i \,, \tag{3.5}$$

where $\alpha_i = |u_i^+ - w_i^+| + |u_i^- - w_i^-|$ and $C_Q = M_g + M' + \epsilon M_g^2 + 3\epsilon M_g M'$.

Proof. Denote

$$Q_i(u) = \frac{\int_{u_i^-}^{u_i^+} (2x - (u_i^+ + u_i^-))g(x)dx}{(u_i^+ - u_i^-) + \epsilon \int_{u_i^-}^{u_i^+} g(x)dx} = \frac{N_i}{D_i} .$$

In order to represent the above expression in a more convenient form, we employ the following change of variables argument:

$$\int_\alpha^\beta f(x)\,dx = (\beta - \alpha)\int_{-1}^{1} f\Big(\frac{\alpha+\beta}{2} + \frac{s}{2}(\beta-\alpha)\Big)\,ds . \tag{3.6}$$

Introducing $-1 \le s \le 1$, we can rewrite $x_u(s) = \frac{u_i^+ + u_i^-}{2} + \frac{s}{2}(u_i^+ - u_i^-)$ and $x_w(s) = \frac{w_i^+ + w_i^-}{2} + \frac{s}{2}(w_i^+ - w_i^-)$. Then for the numerator we have

$$\begin{aligned} N_i(u) &= \int_{u_i^-}^{u_i^+} (2x - (u_i^+ + u_i^-))g(x)dx \\ &= (u_i^+ - u_i^-)\int_{-1}^{1} s(u_i^+ - u_i^-)g(x_u(s))\,ds \\ &= (u_i^+ - u_i^-)^2 \int_{-1}^{1} g(x_u(s))s\,ds, \end{aligned}$$

while the denominator is equal to

$$D_i(u) = (u_i^+ - u_i^-)\Big(1 + \epsilon\int_{-1}^{1} g(x_u(s))\,ds\Big) .$$

For the ratio $Q_i = N_i/D_i$ we have

$$Q_i(u) = \frac{N_i(u)}{D_i(u)} = \frac{(u_i^+ - u_i^-)\int_{-1}^{1} g(x_u(s))s\,ds}{1 + \epsilon\int_{-1}^{1} g(x_u(s))\,ds} .$$

We thus can use the bound on g to get (3.4). Now, for simplicity let us redefine the modified numerator and denominator as

$$\widetilde{N}_i(u) = (u_i^+ - u_i^-)\int_{-1}^{1} g(x_u(s))s\,ds , \quad \widetilde{D}_i(u) = 1 + \epsilon\int_{-1}^{1} g(x_u(s))\,ds .$$

In the new notation,

$$\begin{aligned} |Q_i(u) - Q_i(w)| &\le \frac{1}{\widetilde{D}_i(u)\widetilde{D}_i(w)}\Big|\widetilde{N}_i(u)\widetilde{D}_i(w) - \widetilde{N}_i(w)\widetilde{D}_i(u)\Big| \\ &\le \frac{1}{2}\Big|\widetilde{N}_i(u) - \widetilde{N}_i(w)\Big|\Big(\widetilde{D}_i(u) + \widetilde{D}_i(w)\Big) \\ &\quad + \frac{1}{2}\Big(\widetilde{N}_i(u) + \widetilde{N}_i(w)\Big)\Big|\widetilde{D}_i(u) - \widetilde{D}_i(w)\Big| . \end{aligned}$$

Notice further that $|x_u(s) - x_w(s)| \le |u_i^+ - w_i^+| + |u_i^- - w_i^-| = \alpha_i$. It follows that

$$\left|\widetilde{D}_i(u) - \widetilde{D}_i(w)\right| = \epsilon\left|\int_{-1}^{1} (g(x_u(s)) - g(x_w(s)))\, ds\right| \le \epsilon M_g' \alpha_i \ ,$$

$$\begin{aligned}\left|\widetilde{N}_i(u) - \widetilde{N}_i(w)\right| &= \Big|\int_{-1}^{1} ((u_i^+ - u_i^-)g(x_u(z)) - (w_i^+ - w_i^-)g(x_w(s)))s\, ds\Big| \\ &\le 2(M_g + M_g')\alpha_i \ .\end{aligned}$$

Finally, since $|\widetilde{D}_i| \le 1 + \epsilon M_g$ and $\widetilde{N}_i \le 2M_g$, we have

$$|Q_i(u) - Q_i(w)| \le \Big((M_g + M')(1 + \epsilon M_g) + 2\epsilon M_g M'\Big)\alpha_i = C_Q \alpha_i,$$

which leads to (3.5). This proves the lemma. ∎

In a similar fashion, by employing the change of variables argument, we can easily prove the following.

LEMMA 3.3. *With the same assumptions (and notation) as the previous lemma,*

$$|\mathbf{T}_i(u) - \mathbf{T}_i(w)| \le C_T \alpha_i \quad \textit{and} \quad \sum_j |\widetilde{S}_{ij}(u) - \widetilde{S}_{ij}(w)| \le C_S \alpha_i.$$

Proof. The proofs for both $\mathbf{T}$ and $\widetilde{S}$ are analogous and follow along the lines of Lemma 3.2. For instance, direct calculation yields

$$\begin{aligned}\mathbf{T}_i(u) &= u_i - \Big(\int \rho(x)dx\Big)^{-1}\Big(\int (u_i - x)dx + \epsilon\int (u_i - x)g(x)dx\Big) \\ &= u_i - \frac{(u_i^+ - u_i^-)(2x_i - u_i^+ - u_i^-) + 2\epsilon\int (u_i - x)g(x)dx}{2(u_i^+ - u_i^-) + 2\epsilon\int g(x)dx} \\ &= \frac{u_i^+ + u_i^-}{2} - \frac{2\epsilon\int (u_i - x)g(x)dx - \epsilon(u_i^+ - u_i^-)\int g(x)dx}{2(u_i^+ - u_i^-) + 2\epsilon\int g(x)dx} \\ &= \frac{1}{2}\Big(u_i^+ + u_i^-\Big) + \epsilon\Big(\frac{\int (2x - (u_i^+ + u_i^-))g(x)dx}{2(u_i^+ - u_i^-) + 2\epsilon\int g(x)dx}\Big) \ ,\end{aligned}$$

where the integrals all refer to the integral over $[u_i^-, u_i^+]$. By the definition (3.3), we have

$$\mathbf{T}_i(u) = \frac{u_i^+ + u_i^-}{2} + \frac{1}{2}\epsilon Q_i \ . \tag{3.7}$$

One may then use estimates on Q_i to derive similar estimates for $\mathbf{T}_i$. We omit the rest of the details. ∎

With the help of Lemmas 3.2 and 3.3 we can finally derive the following general result.

PROPOSITION 3.2. *For any $\rho(x) = 1 + \epsilon g(x)$ with ϵ suitably small, there exist constants $C_l > 0$ and $C_u > 0$ dependent only on k, such that for any $u, w \in \mathbf{W}$ and with the notation defined earlier*

$$kC_l(k)||u - w||_{\mathbf{W}}^2 \le (\nabla\widetilde{\mathcal{H}}(u) - \nabla\widetilde{\mathcal{H}}(w), u - w) \le kC_u(k)||u - w||_{\mathbf{W}}^2 \ .$$

Proof. Recall from (2.5) that $\frac{\partial \widetilde{\mathcal{H}}}{\partial u_i} = 4[(u_i - \mathbf{T}_i(u)) - \epsilon\beta_i(u)]$. By (3.7), we get

$$\frac{\partial \widetilde{\mathcal{H}}}{\partial u_i}(u) = \Big(a_i - a_{i+1}\Big) - 2\epsilon(Q_i(u) + 2\beta_i(u))$$

and hence

$$(\nabla\widetilde{\mathcal{H}}(u) - \nabla\widetilde{\mathcal{H}}(w), u - w) = \sum_{i=1}^{k} (a_i - b_i)^2 - 2\epsilon(Q(u) - Q(w), u - w)$$

$$- 4\epsilon(\beta(u) - \beta(w), u - w)\ . \tag{3.8}$$

The first term in (3.8) comes from the constant part of the density and hence complies with the results of the previous theorem.

For the second term, from the Cauchy inequality and Lemma 3.2, we have

$$|\epsilon(Q(u) - Q(w), u - w)| \le \sum_{i=1}^{k} \epsilon|Q_i(u) - Q_i(w)|\,|u_i - w_i| \le C_q\epsilon k\|u - w\|_{\mathbf{W}}^2 \tag{3.9}$$

for some constant C_q.

Finally, for the last term in (3.8), we note that by (2.5) and (3.7)

$$\beta(u) = \widehat{S}(u)(u - \mathbf{T}(u)) = \widehat{S}(u)\left[(I - S)u - \frac{1}{2}\epsilon Q(u)\right] = \widetilde{S}(u)\left[u - \frac{\epsilon}{2}(I - S)^{-1}Q(u)\right].$$

Hence, $(u - w)^T(\beta(u) - \beta(w)) = B_1 - B_2 + B_3 - B_4$, where

$$B_1 = (u - w)^T\widetilde{S}(u)(u - w), \qquad B_2 = \frac{\epsilon}{2}(u - w)^T\widetilde{S}(u)(I - S)^{-1}[Q(u) - Q(w)],$$

$$B_3 = (u - w)^T[\widetilde{S}(u) - \widetilde{S}(w)]w, \quad B_4 = \frac{\epsilon}{2}(u - w)^T[\widetilde{S}(u) - \widetilde{S}(w)](I - S)^{-1}Q(u).$$

Using the fact that $\widetilde{S}(u)$ is uniform bounded, we get from the Cauchy inequality that

$$|B_1| \le C_{b1}k\|u - w\|_{\mathbf{W}}^2$$

for some constant C_{b1}. Similarly for B_2, we apply the Cauchy inequality along with Lemma 3.2 and the spectral bound on $(1 - S)^{-1}$ (which is of the order k^2) to get

$$|B_2| \le \epsilon C_{b2}k^3\|u - w\|_{\mathbf{W}}^2$$

for some constant C_{b2}. As for B_3, we may use the estimates on $\widetilde{S}(u) - \widetilde{S}(w)$ in Lemma 3.3 and apply the Cauchy inequality to get

$$|B_3| \le C_{b3}k\|u - w\|_{\mathbf{W}}^2$$

for some constant C_{b3}. In addition, for ϵ small, it is easy to see that $Q_i(u)$ is uniformly bounded (independent of k). Hence we use the Cauchy inequality, the estimates for $\widetilde{S}(u) - \widetilde{S}(w)$ in Lemma 3.3, along with the same spectral bound on $(1 - S)^{-1}$ we used in B_2 estimation to get

$$|B_4| \le C_{b4}\epsilon k^3\|u - w\|_{\mathbf{W}}^2$$

for some constant C_{b4}. Hence, we get for some constant C_β, independent of k,

$$|\epsilon(\beta(u)-\beta(w),u-w)| \le C_\beta k(\epsilon+\epsilon^2k^2)||u-w||^2_{\mathbf{W}} \ . \tag{3.10}$$

Combining (3.5), (3.8), (3.9), and (3.10), we get

$$\begin{aligned} kC_l||u-w||^2_{\mathbf{W}} &= k[1-2\epsilon C_q-4C_\beta(\epsilon+\epsilon^2k^2)]\|u-w\|^2_{\mathbf{W}} \\ &\le (\nabla\widetilde{\mathcal{H}}(u)-\nabla\widetilde{\mathcal{H}}(w),u-w) \\ &\le k[1+2\epsilon C_q+4C_\beta(\epsilon+\epsilon^2k^2)]\|u-w\|^2_{\mathbf{W}} \\ &= kC_u||u-w||^2_{\mathbf{W}} \ , \end{aligned}$$

with constants $C_l = 1-2\epsilon C_q-4C_\beta(\epsilon+\epsilon^2k^2)]$ and $C_u = 1+2\epsilon C_q+4C_\beta(\epsilon+\epsilon^2k^2)$.

So, if ϵ is small such that ϵk is also small, independent of k, then we have the proposition. ∎

Note that it follows from Proposition 3.2 that the functional $F = k^{-1}\widetilde{\mathcal{H}}$ preserves the continuity and convexity for the class of densities considered here. Consequently, the preconditioned energy is guaranteed to be convex as long as the perturbations to a constant density $\rho(x) = 1+\epsilon g(x)$ remain small enough with $\epsilon = o(k^{-1})$. In other words, we obtain the following.

COROLLARY 3.4. *For any $\rho(x) = 1+\epsilon g(x)$ with ϵ suitably small, the functional $F = k^{-1}\widetilde{\mathcal{H}}$ satisfies the continuity and convexity conditions with $K = C_l, L = C_u$ for all points in $\mathbf{W}$.*

In addition to the convexity and continuity properties, we also need the following conditions on the space decomposition to be satisfied.

Condition A. For all $v \in \mathbf{W}$, there exists $v_i \in \overline{\mathbf{W}}_i$ such that $\sum_{i=1}^J v_i = v$, and $(\sum_{i=1}^J ||v_i||^2_{\mathbf{W}})^{1/2} \le C_1||v||_{\mathbf{W}}$ for some C_1 independent of k.

Condition B. *Strengthened Cauchy–Schwarz.* For all $w_{ij} \in \mathbf{W}, u_i \in \overline{\mathbf{W}}_i, v_j \in \overline{\mathbf{W}}_j \Rightarrow$

$$\sum_{i,j=1}^J (\nabla F(w_{ij}+u_i)-\nabla F(w_{ij}),v_j) \le C_2\left(\sum_{i=1}^J ||u_i||^2_{\mathbf{W}}\right)^{1/2}\left(\sum_{j=1}^J ||v_j||^2_{\mathbf{W}}\right)^{1/2}$$

for some $C_2 \le cL$, where c is independent of k.

These conditions are verified below. First, the following result follows from a simple accounting and the properties of the uniform domain decomposition we defined in section 2.2. Here and later supp denotes the support set of a function.

LEMMA 3.5. *For the one-dimensional* "hat" *basis, for all $u \in \overline{\mathbf{W}}_j$ and all $v \in \overline{\mathbf{W}}_l$ such that $j > l$, the following scaling of the supports relationship holds:*

$$\operatorname{supp}(u) \cap \operatorname{supp}(v) = 2^{-|j-l|}\operatorname{supp}(v).$$

In addition, denote the connected set of common indices $\mathcal{I} = \{0 \le i \le k+1|\ x_i^J \in \operatorname{supp}(u) \cap \operatorname{supp}(v)\}$ and define a restricted norm to be

$$||v||^2_{\mathbf{W},\operatorname{supp}(u)\cap\operatorname{supp}(v)} = \frac{1}{k}\sum_{i\in\mathcal{I}}(v_i-v_{i-1})^2 \ .$$

Then the following norm equivalence up to scaling is satisfied:

$$||v||_{\mathbf{W},\operatorname{supp}(u)\cap\operatorname{supp}(v)} = 2^{-|j-l|}||v||_{\mathbf{W}}. \tag{3.11}$$

The result of this lemma follows by counting the number of common nodes in the intersection of the supports of functions at different levels of the hierarchical decomposition and from the properties of the piecewise linear functions.

With the help of Lemma 3.5, we can show that for the chosen decomposition, both Conditions A and B are satisfied.

THEOREM 3.6. *For the nested subspace decomposition with the choice of the "hat" basis functions,* $(\sum_{i=1}^{J} ||v_i||_{\mathbf{W}}^2)^{1/2} = ||v||_{\mathbf{W}}$, *so that* $C_1 = 1$. *Moreover, for* $F = k^{-1}\widetilde{\mathcal{H}}$, C_2 *can be estimated as* $C_2 = L \cdot \max_j (\sum_{l=1}^{J} 2^{-|j-l|}) \le 2L$.

Proof. Notice that "hat" functions form an orthogonal basis, so

$$\left(\sum_{i=1}^{J} ||v_i||_{\mathbf{W}}^2 \right)^{1/2} = ||v||_{\mathbf{W}}$$

follows easily from calculation. As for the C_2, without loss of generality fix $j > l$ and take any $w \in \mathbf{W}$ and $u \in \overline{\mathbf{W}}_j$, $v \in \overline{\mathbf{W}}_l$. Now, adopting a similar argument for those in the proofs of Propositions 3.1 and 3.2 and taking notice that the contributions are confined in the support regions of u and v, we get

$$\begin{aligned}
&(\nabla\widetilde{\mathcal{H}}(u+w) - \nabla\widetilde{\mathcal{H}}(w), v) \\
&= \sum_{i=1}^{k+1} \{a_i\, b_i - 2\epsilon[(Q_i(u+w) - Q_i(w))v_i + 2(\beta_i(u+w) - \beta_i(w))v_i]\} \\
&= \sum_{i\in\mathcal{I}} (v_i - v_{i-1})(u_i - u_{i-1}) - 2\epsilon[(Q_i(u+w) - Q_i(w))v_i + 2(\beta_i(u+w) - \beta_i(w))v_i] \\
&\le k||u||_{\mathbf{W},\mathrm{supp}(u)\cap\mathrm{supp}(v)}||v||_{\mathbf{W},\mathrm{supp}(u)\cap\mathrm{supp}(v)} \\
&\quad + 2\epsilon C_q||u||_{\mathbf{W},\mathrm{supp}(u)\cap supp(v)}||v||_{\mathbf{W},\mathrm{supp}(u)\cap\mathrm{supp}(v)} \\
&\quad + 4kC_\beta(\epsilon + \epsilon^2 k^2)||u||_{\mathbf{W},\mathrm{supp}(u)\cap\mathrm{supp}(v)}||v||_{\mathbf{W},\mathrm{supp}(u)\cap\mathrm{supp}(v)} \\
&\le kL||u||_{\mathbf{W}} \cdot ||v||_{\mathbf{W},\mathrm{supp}(u)\cap\mathrm{supp}(v)},
\end{aligned}$$

where $\mathcal{I}$ is the same as in Lemma 3.5, $L = 1 + 2\epsilon C_q + 4C_\beta(\epsilon + \epsilon^2 k^2)$ is the continuity constant defined in Corollary 3.4, and C_β, C_q are as defined in Proposition 3.2.

Now by Lemma 3.5, $||v||_{\mathbf{W},\mathrm{supp}(u)\cap\mathrm{supp}(v)} = 2^{-|j-l|}||v||_{\mathbf{W}}$. It follows that, since for any symmetric matrix $||Ax|| \le \rho(A)||x|| \le \max_i (\sum_{j=1}^{n} |a_{ij}|)||x||$,

$$\begin{aligned}
\sum_{i,j=1}^{J} (\nabla\widetilde{\mathcal{H}}(w_{ij} + u_i) - \nabla\widetilde{\mathcal{H}}(w_{ij}), v_j) &\le kL \sum_{i,j=1}^{J} 2^{-|i-j|}||u_i||_{\mathbf{W}}||v_j||_{\mathbf{W}} \\
&\le kL \left(\max_j \sum_{i=1}^{J} 2^{-|i-j|} \right) \left(\sum_{i=1}^{J} ||u_i||_{\mathbf{W}}^2 \right)^{1/2} \left(\sum_{j=1}^{J} ||v_j||_{\mathbf{W}}^2 \right)^{1/2}.
\end{aligned}$$

Henceforth, we have $C_2 = L \cdot \max_j \sum_{i=1}^{J} 2^{-|i-j|} \le 2L$.

Note that although the proof above is presented for the special case of the "hat" basis that we used in our numerical implementation, similar arguments can be used to show that Conditions A and B hold for other suitable bases. Indeed, since the energy part of the argument does not depend on the decomposition, Conditions A and B will remain true for any basis as long as

$$\left(\sum_{i=1}^{J} ||v_i||_{\mathbf{W}}^2 \right)^{1/2} \le C_1||v||_{\mathbf{W}}$$

and

$$||v||_{\mathbf{W},\mathrm{supp}(v)\cap\mathrm{supp}(u)} \le C\cdot 2^{-|j-l|}||v||_{\mathbf{W}}$$

hold for some constants C, C_1 independent of k.

3.3. Uniform convergence theorem. Finally, putting together Conditions A and B and using convexity and continuity of $F = k^{-1}\widetilde{\mathcal{H}}$ in $\mathbf{W}$, we are ready to prove the following uniform convergence result.

THEOREM 3.7. *Under Conditions* A *and* B *on space decomposition, Algorithm* 2.1 *converges uniformly in* $\mathbf{W}$ *for any density of type* $\rho(x) = 1 + \epsilon g(x)$ *with sufficiently small* ϵ. *Moreover,* $d_n = \widetilde{\mathcal{H}}(u_n) - \widetilde{\mathcal{H}}(u)$ *satisfies*

$$d_n \le r d_{n-1}, \quad r \in (0,1),$$

where $r = \frac{C}{1+C}$ *and* $C = C_1^2 C_2^2 L/K^3$.

Before proceeding to its proof, let us first state a consequence.

COROLLARY 3.8. *In the case of a "hat" basis, the constants* C_1 *and* C_2 *can be estimated as* $C_1 = 1$ *and* $C_2 = 2L$, *so, for example, when* $\rho(x) = 1$, $C = 4$.

The proof of this result is similar to the one given in [37] and relies on the following inequality:

$$F(u) - F(v) - (\nabla F(v), u - v) = \int_0^1 (u - v, \nabla F(u + t(v-u)) - \nabla F(v))dt$$

$$\ge K\int_0^1 ||(1-t)(u-v)||_{\mathbf{W}}^2 \frac{dt}{1-t} = \frac{K}{2}||u-v||_{\mathbf{W}}^2 . \tag{3.12}$$

3.4. Proof of Theorem 3.7. Denote u to be the exact solution of (2.3). Consider u_n as the approximate solution after one J-level iteration of the Algorithm 2.2. At the ith level, since the supports of the basis functions w_j^i are disjoint in $\overline{\mathbf{W}}_i$, we have

$$u_{n+\frac{i}{J}} = u_{n+\frac{i-1}{J}} + e_n^i, \quad \widetilde{\mathcal{H}}(u_{n+\frac{i-1}{J}} + e_n^i) \le \widetilde{\mathcal{H}}(u_{n+\frac{i-1}{J}} + v_i) \quad \forall v_i \in \overline{\mathbf{W}}_i,$$

where in the notation of the algorithm, $e_n^i = \sum_{l=1}^{\nu_1}\sum_{j=1}^{n_i}\alpha_{jl}^0 w_j^i \in \overline{\mathbf{W}}_i$.

First notice that since the minimizer $u_{n+\frac{i}{J}}$ satisfies $(\nabla\widetilde{\mathcal{H}}(u_{n+\frac{i}{J}}), v) = 0$ for all $v \in \overline{\mathbf{W}}_i$, it follows from (3.12) that

$$\begin{aligned}
\widetilde{\mathcal{H}}(u_n) - \widetilde{\mathcal{H}}(u_{n+1}) &= \sum_{i=1}^{J}(\widetilde{\mathcal{H}}(u_{n+\frac{i-1}{J}}) - \widetilde{\mathcal{H}}(u_{n+\frac{i}{J}})) \\
&\ge \sum_{i=1}^{J}\left((\nabla\widetilde{\mathcal{H}}(u_{n+\frac{i}{J}}), u_{n+\frac{i-1}{J}} - u_{n+\frac{i}{J}}) + \frac{K}{2}||u_{n+\frac{i-1}{J}} - u_{n+\frac{i}{J}}||_{\mathbf{W}}^2\right) \\
&= \frac{K}{2}\sum_{i=1}^{J}||e_n^i||_{\mathbf{W}}^2 .
\end{aligned}$$

Next, let us use Condition A to decompose $u_{n+1} - u = \sum_{i=1}^{J} v_i$. Then

$$\begin{aligned}(\nabla\widetilde{\mathcal{H}}(u_{n+1}) - \nabla\widetilde{\mathcal{H}}(u), u_{n+1} - u) &= (\nabla\widetilde{\mathcal{H}}(u_{n+1}), u_{n+1} - u)\\ &= \sum_{i=1}^{J} (\nabla\widetilde{\mathcal{H}}(u_{n+1}), v_i) = \sum_{i=1}^{J} (\nabla\widetilde{\mathcal{H}}(u_{n+1}) - \nabla\widetilde{\mathcal{H}}(u_{n+\frac{i}{J}}), v_i)\\ &= \sum_{i=1}^{J}\sum_{j\geq i+1}^{J} (\nabla\widetilde{\mathcal{H}}(u_{n+\frac{j}{J}}) - \nabla\widetilde{\mathcal{H}}(u_{n+\frac{j-1}{J}}), v_i)\\ &\leq C_2\Big(\sum_{j=1}^{J} ||e_n^j||_{\mathbf{W}}^2\Big)^{1/2}\Big(\sum_{i=1}^{J} ||v_i||_{\mathbf{W}}^2\Big)^{1/2}.\end{aligned}$$

Hence

$$\begin{aligned}(\nabla\widetilde{\mathcal{H}}(u_{n+1}) - \nabla\widetilde{\mathcal{H}}(u), u_{n+1} - u) &\leq C_1C_2\Big(\sum_{j=1}^{J} ||e_n^j||_{\mathbf{W}}^2\Big)^{1/2} ||u_{n+1} - u||_{\mathbf{W}}\\ &\leq C_1C_2\left(\frac{2}{K}(\widetilde{\mathcal{H}}(u_n) - \widetilde{\mathcal{H}}(u_{n+1}))\right)^{1/2} ||u_{n+1} - u||_{\mathbf{W}}.\end{aligned}$$

Denote $d_n = \widetilde{\mathcal{H}}(u_n) - \widetilde{\mathcal{H}}(u)$; then $\widetilde{\mathcal{H}}(u_n) - \widetilde{\mathcal{H}}(u_{n+1}) = d_n - d_{n+1}$, and it follows from the inequality above that

$$\left(\frac{2}{K}(d_n - d_{n+1})\right)^{1/2} \geq \frac{(\nabla\widetilde{\mathcal{H}}(u_{n+1}) - \nabla\widetilde{\mathcal{H}}(u), u_{n+1} - u)}{C_1C_2||u_{n+1} - u||_{\mathbf{W}}}.$$

Thus,

$$\begin{aligned}d_n - d_{n+1} &\geq \frac{K}{2}\left(\frac{(\nabla\widetilde{\mathcal{H}}(u_{n+1}) - \nabla\widetilde{\mathcal{H}}(u), u_{n+1} - u)}{C_1C_2||u_{n+1} - u||_{\mathbf{W}}}\right)^2\\ &\geq \frac{K}{2}(C_1C_2)^{-2}K^2||u_{n+1} - u||_{\mathbf{W}}^2 \geq \frac{K^3}{C_1^2C_2^2L}d_{n+1}.\end{aligned}$$

The last step of the argument uses (3.2) and an inequality similar to (3.12) for the upper bound:

$$d_{n+1} = \widetilde{\mathcal{H}}(u_{n+1}) - \widetilde{\mathcal{H}}(u) \leq \frac{L}{2}||u_{n+1} - u||_{\mathbf{W}}^2.$$

As a consequence, we get

$$d_{n+1} \leq \frac{C_1^2C_2^2L}{K^3}(d_n - d_{n+1}) \Rightarrow d_{n+1} \leq \frac{C}{1+C}d_n, \quad \text{where } C = \frac{C_1^2C_2^2L}{K^3}.$$

Since for any basis satisfying Conditions A and B, $C_1^2C_2^2L/K^3 = O(L^3/K^3)$, the convergence is uniform as long as the ratio L/K does not depend on k. By looking at the L and K estimates obtained in Proposition 3.2, it is easy to see that this condition is satisfied for any smooth perturbation of the constant density $1 + \epsilon g(x)$ with ϵ of the order of $O(1/k)$. This concludes the proof of the main theorem. ∎

It follows that, for a suitable choice of decomposition in one dimension, the asymptotic convergence factor of our multilevel algorithm is independent of the size of the problem and the number of grid levels, which gives a significant speedup compared to other methods, like the traditional Lloyd iteration. This claim is further substantiated by the following numerical examples. Although for the sake of simplicity we have presented the detailed theory only for the case of the $V(\nu)$ multigrid cycle with no postsmoothings, same conclusions can be drawn for the case of the full $V(\nu_1, \nu_2)$ cycle, and the results of the numerical experiments in both cases are outlined below.

4. Numerical examples. We now report some numerical results obtained using the new multilevel algorithm.

First we compare the results of our $V(1,1)$ multigrid implementation with the usual Lloyd iteration for the one-dimensional problem. Then we present some results for a two-dimensional test problem in a parallelogram domain. The results are obtained with the MATLAB 6.5 implementation of the new algorithm. The test runs were performed on a PC with a Pentium IV processor and 512MB RAM.

The one-dimensional implementation is very straightforward. Here, we take the unit interval and test a couple of different density functions, like $\rho(x) = 1$ and $\rho(x) = 1 + x$. We plot the convergence factor $\rho \approx |z_{n+1} - z_n|/|z_n - z_{n-1}|$ for each $V(1,1)$ cycle with respect to the total number of generators (grid points) involved.

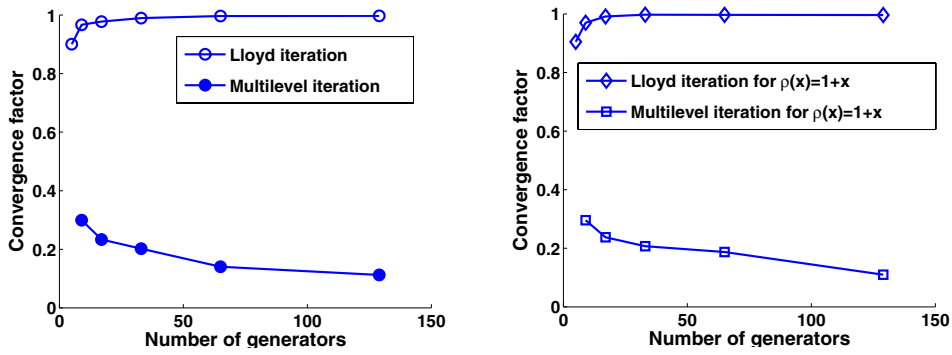

FIG. 4.1. *Plot of the convergence factor vs. the number of generators for the regular Lloyd (upper) and the multilevel (lower curves) iterations for* $\rho(x) = 1$ *(left) and* $\rho(x) = 1 + 0.1x$ *(right).*

Figure 4.1 substantiates the fact that the speed of convergence for the proposed scheme remains nearly constant as the number of generators increases. In Figure 4.2, the computational time needed for the $V(1,0)$ implementation of the multilevel method to reach 10^{-12} accuracy is given for $\rho(x) = 1$ and $\rho(x) = 1 + x$, respectively. The graph shows that in the one-dimensional case, the computational time *scales almost linearly* with the problem size.

The data in Table 4.1 shows the stabilization of the number of multigrid cycles $V(\nu_1, \nu_2)$ needed to reduce the error to $\epsilon = 10^{-12}$ in the constant density case. While there is a visible difference between the number of iterations required for $V(1,1)$ and $V(2,2)$ cycles, respectively, saturation occurs if the values of the relaxation parameters are increased, which is why $\nu_{1,2} \leq 2$ in most of our calculations. The geometric rate of the energy and error reduction asserted by Theorem 3.7 are confirmed by the experiments. Indeed, Figure 4.3 shows the convergence history of a $V(1,1)$ cycle against the total number of relaxations for the $k = 129$ case.

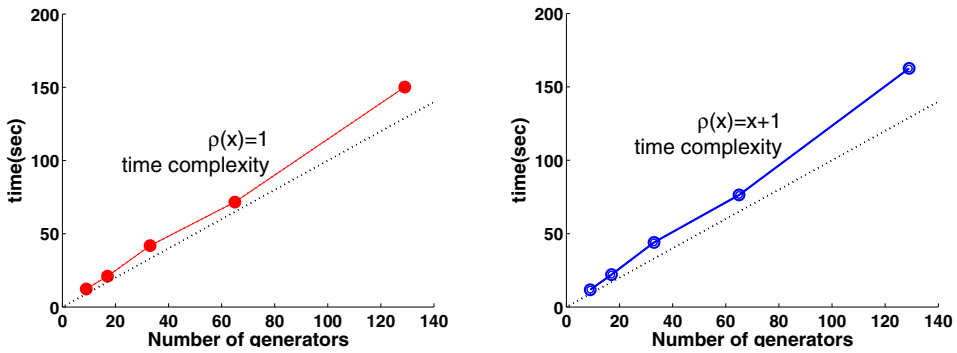

FIG. 4.2. *Computational time vs. problem size for the one-dimensional implementation.*

TABLE 4.1
The number of $V(\nu_1, \nu_2)$ cycles needed vs. the number of generators.

$k/V(\nu_1,\nu_2)$	$V(1,0)$	$V(0,1)$	$V(1,1)$	$V(2,0)$	$V(0,2)$	$V(2,2)$
3	7	8	6	6	7	4
5	11	11	8	8	8	6
9	13	14	9	9	9	7
17	18	18	12	12	12	8
33	21	20	13	12	13	8
65	21	22	12	12	12	8
129	21	21	12	12	12	8
257	20	23	12	12	13	7
513	20	22	12	11	13	7
1025	19	22	11	11	13	7

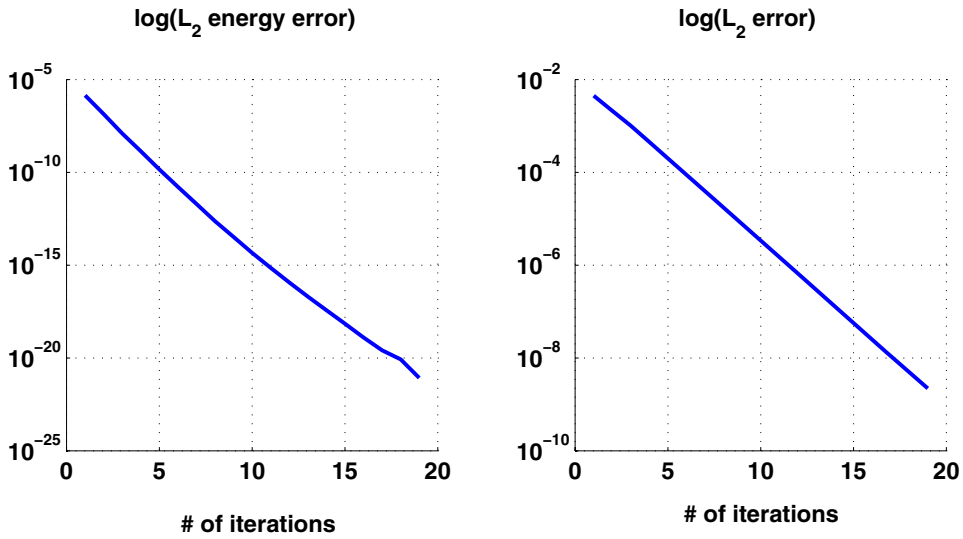

FIG. 4.3. *The energy reduction (left) and the convergence history (right) for* 129 *generators in the log-normal scale.*

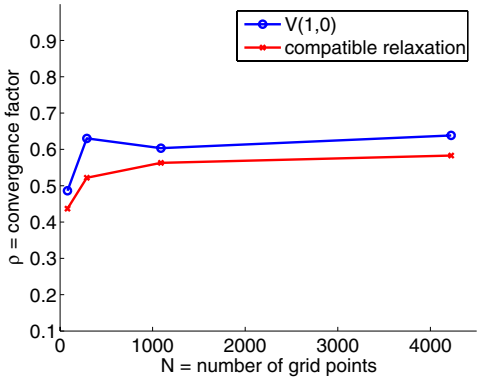

FIG. 4.4. *Comparison of the convergence factors for the compatible relaxation and a* $V(1,0)$ *cycle for* $\rho = 1$ *on a two-dimensional parallelogram-shaped domain.*

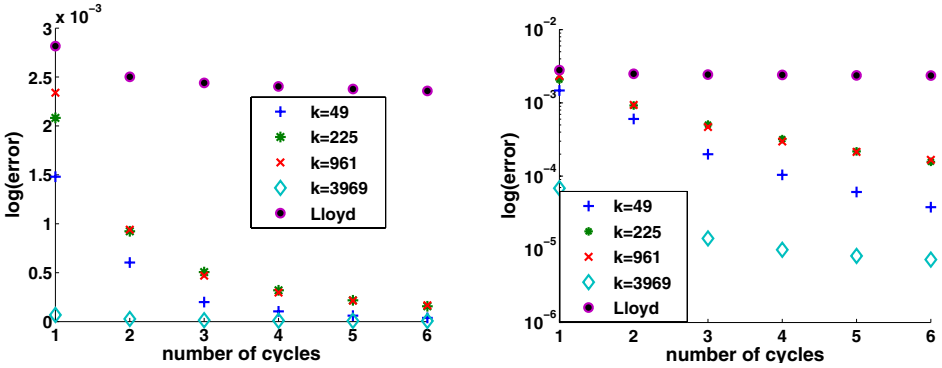

FIG. 4.5. *Convergence history for the two-dimensional multigrid scheme for* $\rho = 1$ *on a parallelogram-shaped domain compared to the Lloyd scheme shown on the top curve in the normal scale (left) and log-normal scale (right).*

Next, the convergence factors for a two-dimensional problems with $\rho = 1$ on a parallelogram domain are compared in Figure 4.4. The lower line in the graph shows the convergence factor for the compatible relaxation, that is, a relaxation performed on the grid with the exact solution given at the coarse nodes (see [2] for discussions on the compatible relaxation). This factor can serve as a lower bound on the convergence factor of the full multigrid cycle, and the quality of the coarse grid influences the distance between the two graphs. As we can see from Figure 4.4, the result of the compatible relaxation in this case is in good agreement with the convergence factor of the whole $V(1,0)$ cycle given on top, which is an indication of a good quality for the coarsening procedure.

The convergence history plots are given in Figure 4.5. The top curves in both graphs depict the error reductions given by the Lloyd iteration, while those below correspond to the convergence of the multigrid scheme for various problem sizes. One can clearly see that the slopes of error reduction on the logarithmic scale do not depend on the number of generators. Hence, even though our theoretical results here are proved only in one dimension, the proposed algorithm preserves its characteristics in certain higher-dimensional implementations.

5. Conclusion. Recently, several methods have been proposed for accelerating the convergence of the classical Lloyd iteration commonly used in the context of quantization and in the construction of centroidal Voronoi tessellations [9, 10]. These algorithmic advances are important for making the computation of optimal quantizers more efficient and for many other successful applications of CVTs. A few possible extensions that use multilevel techniques to accelerate the convergence of the CVTs have been suggested. One of such extensions uses some algebraic multigrid solvers in the spirit of [3, 7, 35] as preconditioners to accelerate the solution of the linear system at every Newton iteration [9], while the other adopts an energy-based nonlinear multigrid approach with the use of a dynamic nonlinear preconditioning [10]. In this paper, we focus on the latter approach, and for the first time, a rigorous analysis of its convergence properties is presented for a class of one-dimensional density functions, with the results of several numerical experiments given for both one- and two-dimensional cases. A more detailed analysis in the higher-dimensional cases as well as more efficient and robust implementations of the energy-based nonlinear multigrid approach are presently under investigation [12]. We conclude by commenting that there is obviously a great potential in using such multigrid methods to accelerate the optimal quantizer design and more generally the computation of CVTs.

Acknowledgment. The authors would like to thank M. Gunzburger of FSU and L. Zikatanov and J. Xu of PSU for interesting discussions on the subject, and anonymous referees for constructive criticism that helped to make current presentation more consistent.

REFERENCES

[1] F. Aurenhammer, *Voronoi diagrams. A survey of a fundamental geometric data structure*, ACM Comput. Surveys, 23 (1990), pp. 345–405.
[2] A. Brandt, *General highly accurate algebraic coarsening*, Electron. Trans. Numer. Anal., 10 (2000), pp. 1–20.
[3] M. Brezina, A. J. Cleary, R. D. Falgout, V. E. Henson, J. E. Jones, T. A. Manteuffel, S. F. McCormick, and J. W. Ruge, *Algebraic multigrid based on element interpolation (AMGe)*, SIAM J. Sci. Comput., 22 (2000), pp. 1570–1592.
[4] W. L. Briggs, V. E. Henson, and S. F. McCormick, *A Multigrid Tutorial*, 2nd ed., SIAM, Philadelphia, 2000.
[5] M. Cappellari and Y. Copin, *Adaptive spatial binning of integral-field spectroscopic data using Voronoi tessellations*, Monthly Notices Roy. Astronom. Soc., 342 (2003), pp. 345–354.
[6] T. Chan and I. Sharapov, *Subspace correction multilevel methods for elliptic eigenvalue problems*, Numer. Linear Algebra Appl., 1 (1993), pp. 1–7.
[7] Q. Chang and Z. Huang, *Efficient algebraic multigrid algorithms and their convergence*, SIAM J. Sci. Comput., 24 (2002), pp. 597–618.
[8] J. Cortes, S. Martinez, T. Karata, and F. Bullo, *Coverage control for mobile sensing networks*, IEEE Trans. Robotics Automation, 20 (2004), pp. 243–255.
[9] Q. Du and M. Emelianenko, *Acceleration of algorithms for the computation of centroidal Voronoi tessellations*, Numer. Linear Algebra Appl., 13 (2006), pp. 173–192.
[10] Q. Du and M. Emelianenko, *Uniform convergence of a multilevel energy-based quantization scheme*, in Domain Decomposition Methods in Science and Engineering XVI, Lect. Notes Comput. Sci. Eng. 55, O. B. Widlund and D. E. Keyes, eds., Springer, Berlin, 2007, pp. 533–541.
[11] Q. Du, M. Emelianenko, and L. Ju, *Convergence properties of the Lloyd algorithm for computing the centroidal Voronoi tessellations*, SIAM J. Numer. Anal., 44 (2006), pp. 102–119.
[12] Q. Du, M. Emelianenko, and L. Zikatanov, *An Energy-Based Multilevel Quantization Scheme in Multidimension*, preprint, 2007.
[13] Q. Du, V. Faber, and M. Gunzburger, *Centroidal Voronoi tessellations: Applications and algorithms*, SIAM Rev., 41 (1999), pp. 637–676.

[14] Q. Du and M. Gunzburger, *Grid generation and optimization based on centroidal Voronoi tessellations*, Appl. Math. Comput., 133 (2002), pp. 591–607.
[15] Q. Du, M. Gunzburger, and L. Ju, *Meshfree, probabilistic determination of points, support spheres, and connectivities for meshless computing*, Comput. Methods Appl. Mech. Engrg., 191 (2002), pp. 1349–1366.
[16] Q. Du, M. Gunzburger, and L. Ju, *Constrained centroidal Voronoi tessellations for surfaces*, SIAM J. Sci. Comput., 24 (2003), pp. 1488–1506.
[17] Q. Du and D. Wang, *Tetrahedral mesh generation and optimization based on centroidal Voronoi tessellations*, Internat. J. Numer. Methods Engrg., 56 (2002), pp. 1355–1373.
[18] Q. Du and D. Wang, *Recent progress in robust and quality Delaunay mesh generation*, J. Comput. Appl. Math., 195 (2006), pp. 8–23.
[19] R. Dwyer, *Higher-dimensional Voronoi diagrams in linear expected time*, Discrete Comput. Geom., 6 (1991), pp. 343–367.
[20] M. Emelianenko, L. Ju, and A. Rand, *Nondegeneracy and weak global convergence of the Lloyd algorithm in* $\mathbb{R}^d$, SIAM J. Numer. Anal., 46 (2008), pp. 1423–1441.
[21] S. Fortune, *A sweepline algorithm for Voronoi diagrams*, Algorithmica, 2 (1987), pp. 153–174.
[22] E. Gelman and J. Mandel, *On multilevel iterative methods for optimization problems*, Math. Programming, 48 (1990), pp. 1–17.
[23] A. Gersho, *Asymptotically optimal block quantization*, IEEE Trans. Inform. Theory, 25 (1979), pp. 373–380.
[24] R. Gray and D. Neuhoff, *Quantization*, IEEE Trans. Inform. Theory, 44 (1998), pp. 2325–2383.
[25] W. Hackbusch and A. Reusken, *Analysis of a damped nonlinear multilevel method*, Numer. Math., 55 (1989), pp. 225–246.
[26] S. Hiller, H. Hellwig, and O. Deussen, *Beyond stippling methods for distributing objects on the plane*, Computer Graphics Forum, 22 (2003), pp. 515–522.
[27] L. Ju, Q. Du, and M. Gunzburger, *Probabilistic methods for centroidal Voronoi tessellations and their parallel implementations*, Parallel Comput., 28 (2002), pp. 1477–1500.
[28] Y. Koren and I. Yavneh, *Adaptive multiscale redistribution for vector quantization*, SIAM J. Sci. Comput., 27 (2006), pp. 1573–1593.
[29] Y. Koren, I. Yavneh, and A. Spira, *A multigrid approach to the* 1*-D quantization problem*, IEEE Trans. Inform. Theory, 51 (200), pp. 2993–2998.
[30] R. Kornhuber, *Adaptive Monotone Multigrid Methods for Nonlinear Variational Problems*, Teubner, Stuttgart, 1997.
[31] Y. Linde, A. Buzo, and R. Gray, *An algorithms for vector quantizer design*, IEEE Trans. Commun., 28 (1980), pp. 84–95.
[32] S. Lloyd, *Least square quantization in PCM*, IEEE Trans. Inform. Theory, 28 (1982), pp. 129–137.
[33] A. Mendes and I. Themido, *Multi-outlet retail site location assessment*, Int. Trans. Oper. Res., 11 (2004), pp. 1–18.
[34] A. Okabe, B. Boots, and K. Sugihara, *Spatial Tessellations; Concepts and Applications of Voronoi Diagrams*, Wiley, Chichester, UK, 1992.
[35] J. W. Ruge and K. Stüben, *Algebraic multigrid*, in Multigrid Methods, S. F. McCormick, ed., Frontiers in Appl. Math. 3, SIAM, Philadelphia, 1987, pp. 73–130.
[36] J. Sabin and R. Gray, *Global convergence and empirical consistency of the generalized Lloyd algorithm*, IEEE Trans. Inform. Theory, 32 (1986), pp. 148–155.
[37] X. Tai and J. Xu, *Global convergence of subspace correction methods for some convex optimization problems*, Math. Comp., 71 (2002), pp. 105–124.
[38] A. Trushkin, *On the design of an optimal quantizer*, IEEE Trans. Inform. Theory, 39 (1993), pp. 1180–1194.
[39] S. Valette and J. Chassery, *Approximated centroidal Voronoi diagrams for uniform polygonal mesh coarsening*, Computer Graphics Forum, 23 (2004), pp. 381–390.
[40] C. Wager, B. Coull, and N. Lange, *Modelling spatial intensity for replicated inhomogeneous point patterns in brain imaging*, J. R. Stat. Soc. Ser. B Stat. Methodol., 66 (2004), pp. 429–446.
[41] J. Xu and L. Zikatanov, *The method of alternating projections and the method of subspace corrections in Hilbert space*, J. Amer. Math. Soc., 15 (2002), pp. 573–597.