

SPECTRAL FINITE-ELEMENT METHODS FOR PARAMETRIC CONSTRAINED OPTIMIZATION PROBLEMS

MIHAI ANITESCU*

Abstract. We present a method to approximate the solution mapping of parametric constrained optimization problems. The approximation, which is of the spectral finite element type, is represented as a linear combination of orthogonal polynomials. Its coefficients are determined by solving an appropriate finite-dimensional constrained optimization problem. We show that, under certain conditions, the latter problem is solvable because it is feasible for a sufficiently large degree of the polynomial approximation and has an objective function with bounded level sets. In addition, the solutions of the finite dimensional problems converge for an increasing degree of the polynomials considered, provided that the solutions exhibit a sufficiently large and uniform degree of smoothness. Our approach solves, in the case of optimization problems with uncertain parameters, the most computationally intensive part of stochastic finite-element approaches. We demonstrate that our framework is applicable to parametric eigenvalue problems. Numerical results in one-dimensional cases indicate that our method is, for those examples, superior in both accuracy and speed to black-box approaches.

Key words. spectral approximations, orthogonal polynomials, parametric problems, stochastic finite element, constrained optimization.

AMS subject classifications. 65K05, 42C05.

1. Introduction. This paper is concerned with the application of spectral finite-element methods (SFEMs) to the determination of the parametric variation of the solution of parametric constrained optimization. Parametric problems appear in a variety of circumstances, and, relevant to this work, when the parameters of the problem are uncertain [15]. Applications of parametric problems include elastoplasticity [1], radioactive waste disposal [17], elasticity problems [15], disease transmission [3], and nuclear reactor safety assessment [22].

In parametric uncertainty analysis of nonlinear equations, the problem is to characterize the dependence with respect to parameters of the solution of a nonlinear equation $F(x, \omega) = 0$, $x \in R^n$, $\omega \in \Omega \subset R^m$, $F : R^n \times R^m \rightarrow R^n$. In addition, the function $F(\cdot, \cdot)$ is smooth in both its arguments. Under the assumption of non singularity of $\nabla_x F(x, \omega)$ in a sufficiently large open set that contains (x_0, ω_0) , one can determine a smooth mapping $x(\omega)$ that satisfies $x(\omega_0) = x_0$ and $F(x(\omega), \omega) = 0$. The essence of parametric uncertainty analysis is to characterize the mapping $x(\omega)$ either by approximating it to an acceptable degree or by computing some of its integral characteristics, such as averages with appropriate weighting functions.

Perhaps the most widespread approach in carrying out this endeavor is the use of some form of the Monte Carlo method [19, 28]. In this approach, the parameter ω is interpreted as a random variable with an appropriate probability density function, and either the probability density function of $x(\omega)$ is approximated or computed, or appropriate averages $E_\omega [g(x(\omega))]$ are computed for suitable expressions of the multidimensional merit function g . Here E_ω is the expectation operator with respect with the probability density function of ω . In the Monte Carlo approach, values for

*Argonne National Laboratory, Mathematics and Computer Science Division, 9700 S Cass Avenue, Argonne IL, 60439, USA, email anitescu@mcs.anl.gov

$x(\omega)$ are produced for an appropriate set of sample points ω_i , in which case for each sample point the original nonlinear problem must be solved for its argument x .

1.1. SFEM and Stochastic Finite Element Approaches. Recently, there has been substantial interest in carrying out the analytical computation as far as possible in characterizing the mapping $x(\omega)$. The component of this endeavor that is relevant to this work is the spectral stochastic finite-element method [15, 14]. In this method, the mapping $x(\omega)$ is approximated by a Fourier-type expansion with respect to a basis of polynomials $P_0(\omega), P_1(\omega), \dots, P_{M_K}(\omega)$ that are orthogonal with respect to the probability density function of ω , that is $E_\omega(P_i(\omega)P_j(\omega)) = \delta_{ij}$, $0 \leq i, j \leq M_K$. For $x_0, x_1 \dots x_{M_K} \in R^n$, one defines the SFEM approximation $\tilde{x}(\omega) = \tilde{x}(\omega; x_0, x_1, \dots, x_{M_K}) = \sum_{i=0}^{M_K} x_i P_i(\omega)$, and its vector coefficients $x_0^*, x_1^* \dots x_{M_K}^*$ are determined from the Galerkin projection conditions

$$E_\omega(F(\tilde{x}(\omega), \omega) P_k(\omega)) = 0_n, \quad k = 0, 1, \dots, M_K. \quad (1.1)$$

This procedure results in a nonlinear system of equations that is $M_K + 1$ times larger than the original nonlinear system of equations for a given choice of the parameter ω . The advantage over the Monte Carlo method is that once this nonlinear system of equations is solved, the original nonlinear problem no longer needs to be solved.

In the *spectral stochastic finite-element method* one uses SFEM to generate an approximation of the mapping $x(\omega)$, and its probability density function is evaluated by, for example, a Monte Carlo method using the probability density function of ω and the SFEM approximation $\tilde{x}(\omega) = \sum_{k=0}^{M_K} x_k^* P_k(\omega)$, *without the need to solve any further system of nonlinear equations*. Because the polynomials are used as the generators of the space over which approximation is carried out and the parameter ω has a stochastic interpretation the expansion defined by this approximation is called the *chaos polynomial expansion* [14].

We note that solving for the coefficients of the SFEM approximation in (1.1) is the main topic of this work in the case where $F(\cdot, \cdot)$ is the function defining the first-order conditions of an optimization problem and is the main computational endeavor in stochastic finite-element approaches [15, 14]. Compared to stochastic finite-element approaches, our work does not cover the determination of the probability density function of $\tilde{x}(\omega)$ from the probability density function of the variable ω [15]. Note, however, that such calculations would not involve the model function $F(\cdot, \cdot)$ and generally have far smaller computational complexity compared to the one of obtaining $\tilde{x}(\omega)$.

Of course, the success of this method resides in the ability to suitably choose the set of polynomials P_i so that the residual decreases rapidly for relatively small values of M_K , before the size of the Galerkin projected problem explodes, a situation that occurs if one considers in the approximating set all the polynomials of degree up to K and if m is large. Nonetheless, for cases where n is huge (as are the cases originating in the discretization of partial differential equations) and m is relatively moderate, the SFEM has shown substantially more efficiency compared to the Monte Carlo approach, even when all polynomials of degree up to K were considered as generators of the approximating subspace [1]. In this work, we choose as the basis for the approximation the set of all polynomials of degree up to K , and we will defer the investigation of choosing a smaller subset to future research.

Spectral finite element approaches in the context of uncertainty quantification have been applied primarily to the problem of parametric nonlinear equations [15, 14, 1, 7]. The object of this paper is to analyze the properties of SFEM and its extensions when the original problem is a *constrained optimization problem*.

To demonstrate several of the points we make, as well as to validate our theoretical findings, we will conduct a series of numerical experiments for problems with one-dimensional parameter spaces, some of which have solutions that are not smooth. For such problems there may exist more suitable methods such as mesh-based finite-element methods with mesh adaptation. We point out, however, that such approaches are impractical for high-dimensional problems [24] which are the goal of our approach in the near future. The aim of the examples with nonsmooth solutions is to compare the performance of the Galerkin spectral finite element-approach with collocation, “black-box” approaches, and to verify the result of Theorem 3.9.

2. Background on Spectral Methods. In this section, we use the framework from [10]. The choice of orthogonal polynomials is based on the scalar product

$$\langle g, h \rangle_W = \int_{\Omega} W(\omega)g(\omega)h(\omega)d\omega,$$

where g, h are continuous functions from \mathbb{R}^m to \mathbb{R} . Here $\Omega \in \mathbb{R}^m$ is a compact set with a nonempty interior, and $W(\omega)$ is a weight function that satisfies the following.

1. $W(\omega) \geq 0, \forall \omega \in \Omega$.
2. Any multivariable polynomial function $P(\omega)$ is integrable, that is,

$$\int_{\Omega} W(\omega)|P(\omega)|d\omega < \infty.$$

We define the semi norm

$$\|g\|_W = \sqrt{\langle g, g \rangle_W}$$

on the space of continuous functions. If, in addition, $\|g\|_W = 0 \Rightarrow g = 0$, then $\|\cdot\|_W$ is a norm. We will concern ourselves only with this case, in which we denote by $L^2_W = L^2_W(\Omega)$ the completion of the space of continuous functions whose norm $\|\cdot\|_W$ is finite.

With respect to the scalar product $\langle \cdot, \cdot \rangle_W$, we can orthonormalize the set of polynomials in the variable ω . We obtain the orthogonal polynomials $P_i(\omega)$ that satisfy the following.

- $\langle P_i, P_j \rangle_W = \delta_{ij}, 0 \leq i, j$. By convention, we always take P_0 to be the constant polynomial.
- The set $\{P_i\}_{i=0,1,2,\dots}$ forms the basis of the complete space L^2_W .
- If $k_1 \leq k_2$, then $\deg(P_{k_1}) \leq \deg(P_{k_2})$. To simplify our notation, we introduce the definition

$$M_K = \max\{k | \deg(P_k) \leq K\}.$$

We define $L^2_{p,W} = \underbrace{L^2_W \otimes L^2_W \otimes \dots \otimes L^2_W}_p$. We use the notation $L^2_W = L^2_{p,W}$ when the

value of p can be inferred from the context. The Fourier coefficients of a function $f : \Omega \rightarrow \mathbb{R}^p$ are defined as

$$c_k(f) = \int_{\Omega} f P_k(\omega) W(\omega) d\omega \in \mathbb{R}^p, \quad f \in L^2_W, \quad k = 0, 1, \dots,$$

and they satisfy Bessel's identity [10]

$$f \in L_W^2 \Rightarrow \sum_{k=0}^{\infty} \|c_k(f)\|^2 = \int_{\Omega} \|f(\omega)\|^2 = \|f\|_W^2. \quad (2.1)$$

The projection of a function $f \in L_W^2$ onto the space of the polynomials of degree at most K can be calculated as [10]

$$\Pi_W^K(f) = \sum_{k=0}^{M_K} c_k(f) P_k(\omega).$$

The most common type multidimensional weight function is probably the one of the separable type, that is, $W(\omega_1, \omega_2, \dots, \omega_m) = \prod_{i=1}^m w_i(\omega_i)$. In this case, the orthogonal polynomials can be chosen to be products of orthogonal polynomials in each individual variable [6, 10]. We refer to such orthogonal polynomials as *tensor products*. The case $\Omega = [-1, 1]^m$, with $w_i(x) = \frac{1}{2}$, $i = 1, 2, \dots, m$ is the one of tensor Legendre polynomials, whereas the one with $w_i(x) = \frac{1}{\pi\sqrt{1-x^2}}$, $i = 1, 2, \dots, m$ is the one of Chebyshev polynomials [6].

Following the multidimensional Jackson theorem [13, Theorem 2], there exists a parameter C that depends only on the function f such that

$$D^\alpha f(\omega) \text{ are Lipschitz } \forall \alpha \in \mathbb{N}^m, \|\alpha\|_1 = q - 1 \implies \|f - \Pi_W^K(f)\|_W \leq C \frac{1}{K^q}. \quad (2.2)$$

Here, we denote by D^α the derivative of multiindex $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_m) \in \mathbb{N}^m$,

$$D^\alpha(f) = \frac{\partial^{\sum_{i=1}^m \alpha_i} f}{\partial \omega_1^{\alpha_1} \partial \omega_2^{\alpha_2} \dots \partial \omega_m^{\alpha_m}}.$$

If $m = 1$, then the polynomial functions are polynomials of only one variable, and we can obtain an orthonormal family that satisfies $\deg P_k = k$, and $M_K = K + 1$.

In addition, a reciprocal of (2.2) holds in certain circumstances. There exists a parameter t that depends only on $W(x)$ and on m such that

$$\max \left\{ \|f\|_\infty, \left\| \frac{\partial f}{\partial \omega_1} \right\|_\infty, \left\| \frac{\partial f}{\partial \omega_2} \right\|_\infty, \dots, \left\| \frac{\partial f}{\partial \omega_m} \right\|_\infty \right\} \leq C_S \sum_{k=0}^{\infty} \|c_k(f)\| \deg(P_k)^t < \infty. \quad (2.3)$$

Indeed, for tensor Legendre and Chebyshev polynomials such a conclusion follows by techniques described in [6] from choosing an appropriate t , computing the Sobolev norm of weak derivatives of f whose projection can be explicitly computed for either case followed by an application of Sobolev's theorem.

Finally, for some orthogonal polynomial families, the following holds.

$$\Lambda^K = \sup_{\omega \in \Omega} \sqrt{\sum_{k=0}^{M_K} (P_k(\omega))^2} \leq C_\Lambda M_K^d, \quad (2.4)$$

where d and C_Λ are parameters, depending on m , Ω , but not on K . For tensor-product Chebyshev polynomials, and $\Omega = [-1, 1]^m$, it is immediate that $C_\Lambda = 1$ and $d = \frac{m}{2}$. For tensor-product Legendre polynomials, one can choose $d = m$, following

the properties of separable weight functions [10, Proposition 7.1.5] as well as the asymptotic properties of Λ^K for the Legendre case when $m = 1$ [23, Lemma 21].

In addition, for the case where

$$\int_{\Omega} W(\omega) d(\omega) = 1,$$

we can interpret $W(\omega)$ as a probability density function (this case can be achieved for any weight function after rescaling with a constant). In that case, we may refer to ω as a random variable, and it is the case we treat in this work.

Notations The expectation of a function $f(\omega)$ of the random variable is

$$E_{\omega} [f(\omega)] = \int_{\Omega} f(\omega) W(\omega) d\omega \triangleq \langle f(\omega) \rangle.$$

The last notation is useful to compact mathematical formulas. Note that the symbol of the scalar product includes a comma ($\langle \cdot, \cdot \rangle$). We use $\|u\|$ to denote the Euclidean norm of a vector $u \in \mathbb{R}^p$. For $f : \Omega \rightarrow \mathbb{R}^p$, the quantity $\|f\|_W = \|f(\omega)\|_W$ is the L^2_W norm, defined in (2.1), whereas $\|f(\omega)\|_{\infty} = \|\|f(\omega)\|\|_{\infty}$.

When proving an inequality or equality, we will display on top of the respective sign the equation that justifies it. For example $\stackrel{(2.1)}{=}$ is an identity justified by Bessel's identity (2.1).

3. Constrained Optimization Problems. Consider the following constrained optimization (O) problem

$$(O) \quad \begin{aligned} \tilde{x}^*(\omega) &= \arg \min_x f(x, \omega) \\ \text{subject to } g(x, \omega) &= 0_p. \end{aligned}$$

Here, the function $g : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^p$. We are interested in approximating the mapping $x(\omega)$, where $\omega \in \Omega$.

3.1. SFEM formulations. An SFEM formulation can be obtained by writing the optimality conditions for the problem (O) after we introduce the Lagrange multiplier mapping $\lambda(\omega) : \Omega \rightarrow \mathbb{R}^m$, followed by the procedure outlined in [15]. The optimality conditions result in

$$\begin{aligned} \nabla_x f(x(\omega), \omega) + \lambda^T(\omega) \nabla_x g(x(\omega), \omega) &= 0_n \\ g(x(\omega), \omega) &= 0_p. \end{aligned} \quad (3.1)$$

We introduce the SFEM parametrization of the approximation

$$\tilde{x}^K(\omega) = \sum_{k=0}^{M_K} x_k P_k(\omega), \quad \tilde{\lambda}^K(\omega) = \sum_{k=0}^{M_K} \lambda_k P_k(\omega).$$

Here, we have that the coefficients of the expansion satisfy $x_k \in \mathbb{R}^n$ and $\lambda_k \in \mathbb{R}^p$, $k = 0, 1, 2, \dots$. The procedure outlined in [15] results in the following system of nonlinear equations

$$\left. \begin{aligned} \left\langle P_k(\omega) \left(\nabla_x f(\tilde{x}^K(\omega), \omega) + \left(\tilde{\lambda}^K(\omega) \right)^T \nabla_x g(\tilde{x}^K(\omega), \omega) \right) \right\rangle &= 0_n, \\ \left\langle P_k(\omega) g(\tilde{x}^K(\omega), \omega) \right\rangle &= 0_p, \end{aligned} \right\} 0 \leq k \leq M_K. \quad (3.2)$$

We could try to solve the equations (3.2) in order to obtain the SFEM approximation. Once we do that, however, we face the problem of determining whether the resulting system of nonlinear equations has a solution, and how we can determine it. One could imagine that certain results can be proved under the assumption that the solution of (O), $\tilde{x}^*(\omega)$, has sufficiently small variation. A result of this type will be shown in Subsection 3.4 though for weaker assumptions than the small variation of the solution. But more important from a practical perspective, we started with an optimization structure to our original problem (O) and, at first sight, the equations (3.2) do not have an optimization problem structure. This situation restricts the type of algorithms that we could use to solve the problem. Nonetheless, this difficulty is only superficial, as shown by the following theorem, which relates the solution of the nonlinear equations (3.2) to the solution of the following stochastic optimization problem:

$$(SO(K)) \min_{\{x_k\}_{k=0,1,\dots,M_K}} \begin{aligned} & \langle f(\tilde{x}(\omega), \omega) \rangle \\ & \langle g(\tilde{x}(\omega), \omega) P_k(\omega) \rangle = 0_p, \quad k = 0, 1, \dots, M_K. \end{aligned}$$

THEOREM 3.1. *Consider the coefficients $\hat{x}_0, \hat{x}_1, \dots, \hat{x}_{M_K}$ that are a solution of the minimization problem (SO(K)) and assume that they satisfy the KKT conditions with the Lagrange multipliers $\hat{\lambda}_0, \hat{\lambda}_1, \dots, \hat{\lambda}_{M_K}$. With these coefficients and multipliers we define the functions*

$$\hat{x}^K(\omega) = \sum_{k=0,1,\dots,M_K} \hat{x}_k P_k(\omega), \quad \hat{\lambda}^K(\omega) = \sum_{k=0,1,\dots,M_K} \hat{\lambda}_k P_k(\omega).$$

Then, $\hat{x}(\omega)$ and $\hat{\lambda}(\omega)$ satisfy the equations (3.2), assuming that f and g have Lipschitz first derivatives.

Proof The optimality conditions for (SO(K)), that are satisfied by the solution, since the constraint qualifications holds [21], result, for a fixed $k \in \{0, 1, \dots, M_K\}$, in

$$\begin{aligned} 0_n &= \nabla_{x_k} \langle f(\hat{x}^K(\omega), \omega) \rangle + \nabla_{x_k} \sum_{k'=0,1,\dots,M_K} \hat{\lambda}_{k'}^T \langle g(\hat{x}^K(\omega), \omega) P_{k'}(\omega) \rangle \\ &= \nabla_{x_k} \langle f(\hat{x}^K(\omega), \omega) \rangle + \nabla_{x_k} \left\langle \left(\sum_{k'=0,1,\dots,M_K} P_{k'}(\omega) \hat{\lambda}_{k'} \right) g(\hat{x}^K(\omega), \omega) \right\rangle \\ &= \nabla_{x_k} \left\langle f(\hat{x}^K(\omega), \omega) + \left(\hat{\lambda}^K(\omega) \right)^T g(\hat{x}^K(\omega), \omega) \right\rangle \\ &= \left\langle P_k(\omega) \left(\nabla_x f(\hat{x}^K(\omega), \omega) + \left(\hat{\lambda}^K(\omega) \right)^T \nabla_x g(\hat{x}^K(\omega), \omega) \right) \right\rangle, \end{aligned}$$

where we have used the fact that the expectation operator commutes with multiplication with a parameter. We have also used the fact that f and g have Lipschitz continuous derivatives, which allows us to interchange the derivative and the expectation operator.

The last equation represents the first set of equations in (3.2) for $\hat{x}^K(\omega)$ and $\hat{\lambda}^K(\omega)$. Since the second set of equations must be satisfied from the feasibility conditions, the proof of the theorem is complete. \square

The preceding theorem represents the main practical advance brought by this work, because it provides an alternative way of formulating the spectral finite-element

approximation when the original problem has an optimization structure. The computational advantage of the formulation (SO(K)) over the nonlinear equation formulation (3.2) is that it preserves the optimization structure and allows one to use optimization software that is guaranteed to obtain a solution of (3.2) under milder conditions than solving the nonlinear equation directly.

3.2. Assumptions. Our goal is to show that, under certain assumptions, a solution of (SO(K)) approximates a solution of (O). A key step is to ensure that the problem (SO(K)) has a feasible point whose Jacobian of the constraints is well conditioned in the neighborhood of $\tilde{x}^*(\omega)$, the solution of the problem (O). As we will later see, this result, in addition to a bounded level set condition, will be the key to ensuring that (SO(K)) is feasible and, in turn, that (SO(K)) has a solution.

An important result is the following.

THEOREM 3.2 (Kantorovich's theorem for nonsquare systems of equations, [8, 27]). *Assume that $f : X \rightarrow Y$ is defined and differentiable on a ball $\mathcal{B} = \{x \mid \|x - x_0\| \leq r\}$, and assume that its derivative $F(x)$ satisfies the Lipschitz condition on \mathcal{B} :*

$$\|F(x) - F(z)\| \leq L \|x - z\|, \forall x, z \in \mathcal{B}.$$

Here, X and Y are Banach spaces, $F(x)$ maps X onto Y , and the following estimate holds:

$$\|F(x_0)^* y\| \geq \mu \|y\| \text{ for any } y \in Y \quad (3.3)$$

with $\mu > 0$ (the star denotes conjugation). Introduce the function $H(t) = \sum_{k=1}^{\infty} t^{2^k}$, and suppose that $h = \frac{L\mu^2 \|F(x_0)\|}{2} < 1$, $\rho = \frac{2H(t)}{L\mu} \leq r$. Then the equation $F(x) = 0$ has a solution that satisfies $\|x - x_0\| \leq \rho$. **Note** This result is stated slightly differently in [27], where (3.3) is required for all $x \in \mathcal{B}$. However, the purpose in that reference is to prove a rate of convergence result for an iterative process. From Graves' theorem [8, Theorem 1.2] the Theorem 3.2 follows as stated. Note that for Kantorovich's Theorem for square systems [27] (where the spaces X and Y are the same), the condition corresponding to (3.3) is also stated only at x_0 .

We can immediately see that the nature of the constraints in (SO(K)) is quite a bit different from the one of (O). It is clear how to assume well-posedness of the constraints at the solution $\tilde{x}^*(\omega)$.

$$[\mathbf{A1}] \quad \sigma_{\min}(\nabla_x g(\tilde{x}^*(\omega), \omega)) \geq \sigma_m, \forall \omega \in \Omega.$$

Here σ_{\min} is the smallest singular value of a given matrix. It is not clear how to immediately translate [A1] into a proof of well-conditioning for the constraints of (SO(K)):

$$\langle g(\tilde{x}(\omega), \omega) P_k(\omega) \rangle = 0_p, \quad k = 0, 1, 2, \dots, M_K,$$

which we investigate in this section. We have that $\nabla_{x_i} \langle g(\tilde{x}^K(\omega), \omega) P_k \rangle$ are the blocks of the Jacobian matrix at an SFEM approximation $\tilde{x}^K(\omega)$. Since $g(x, \omega)$ has Lipschitz continuous derivatives from Assumption [A3] below and Ω is compact, we can interchange the average and the differentiation and use the chain rule to obtain that the blocks are $\langle \nabla_x g(\tilde{x}^k(\omega), \omega) P_i P_k \rangle$.

Therefore, for fixed K , the Jacobian has dimension $p(M_K + 1) \times n(M_K + 1)$.

$$J^K(\tilde{x}^K) = \begin{bmatrix} J_{00}(\tilde{x}^K) & J_{01}(\tilde{x}^K) & \cdots & J_{0M_K}(\tilde{x}^K) \\ J_{10}(\tilde{x}^K) & J_{11}(\tilde{x}^K) & \cdots & J_{1M_K}(\tilde{x}^K) \\ \vdots & \vdots & \vdots & \vdots \\ J_{M_K 0}(\tilde{x}^K) & J_{M_K 1}(\tilde{x}^K) & \cdots & J_{M_K M_K}(\tilde{x}^K) \end{bmatrix},$$

where

$$J_{ij}(\tilde{x}^K) = \langle \nabla_x g(\tilde{x}^k(\omega), \omega) P_i(\omega) P_j(\omega) \rangle \in \mathbb{R}^{p \times n}, \quad i, j = 0, 1, \dots, M_K.$$

We want to show that the matrix J^K is uniformly well-conditioned with respect to K , for K sufficiently large, at $\tilde{x}^{*,K} = \Pi_W^K(\tilde{x}^*)$. In that sense, we need to prove that its smallest singular value is bounded below. To obtain such a bound, we need a more workable expression for the minimum singular value. The minimum singular value of a matrix B of dimension $p \times n$ is the following inf – sup condition [5]:

$$\sigma_{\min} = \inf_{\lambda \in \mathbb{R}^p} \sup_{u \in \mathbb{R}^n} \frac{\lambda^T B u}{\|\lambda\| \|u\|} = \inf_{\lambda \in \mathbb{R}^p, \|\lambda\|=1} \sup_{u \in \mathbb{R}^n, \|u\|=1} \lambda^T B u.$$

To prove our results, we need to invoke several assumptions. One of the assumptions will involve a statement about functions that have bounded level sets. We say that a function $\chi : \mathbb{R} \rightarrow \mathbb{R}$ has bounded level sets if the sets $\mathcal{L}_M^\chi = \chi^{-1}((-\infty, M])$ are bounded for any $M \in \mathbb{R}$.

- [A2] Uniformly bounded level sets assumption: There exist a function $\chi(\cdot)$ that is convex and nondecreasing, and that has bounded level sets and a parameter $\gamma > 0$ such that

$$\chi(\|x\|^\gamma) \leq f(x, \omega), \quad \forall \omega \in \Omega. \quad (3.4)$$

- [A3] Smoothness assumption: The functions $f(x, \omega)$ and $g(x, \omega)$ have Lipschitz continuous first derivatives in both variables. In addition, $\nabla_x g(x, \omega)$ is uniformly Lipschitz with respect to x , that is there exists $L > 0$ such that $\|\nabla_x g(x_1, \omega) - \nabla_x g(x_2, \omega)\| \leq L(\omega) (\|x_1 - x_2\|) \leq L \|x_1 - x_2\|$, $\forall x_1, x_2 \in \mathbb{R}^n$. The last inequality follows from the fact that Ω is a compact set.
- [A4] The solution of the problem (O), $\tilde{x}^*(\omega)$ is Lipschitz continuous.
- [A5] $cC_G < \frac{1}{4}$, where

$$C_G = \sup_{\omega \in \Omega} \|\nabla_x g(\tilde{x}^*(\omega), \omega)\|,$$

$$c = \sup_{\forall K, Q \in \mathbb{N}} \|(\bar{J}^{K, Q}(\tilde{x}^*))\|.$$

Here,

$$\bar{J}^{K, Q}(\tilde{x}) = \begin{bmatrix} \bar{J}_{0, K+1}(\tilde{x}) & \bar{J}_{0, K+2}(\tilde{x}) & \cdots & \bar{J}_{0, K+Q}(\tilde{x}) \\ \bar{J}_{1, K+1}(\tilde{x}) & \bar{J}_{1, K+2}(\tilde{x}) & \cdots & \bar{J}_{1, K+Q}(\tilde{x}) \\ \vdots & \vdots & \vdots & \vdots \\ \bar{J}_{K, K+1}(\tilde{x}) & \bar{J}_{K, K+2}(\tilde{x}) & \cdots & \bar{J}_{K, K+Q}(\tilde{x}) \end{bmatrix}, \quad (3.5)$$

where

$$\bar{J}_{i,j}(\tilde{x}) = \langle G^\dagger(\omega)P_i(\omega)P_j(\omega) \rangle \in \mathbb{R}^{n \times p}, \quad i, j = 0, 1, 2, \dots,$$

and $G^\dagger(\omega) \in \mathbb{R}^{n \times p}$ (the pseudoinverse) is, following Assumptions [A3] and [A4] a matrix-valued Lipschitz mapping such that

$$\nabla_x g(\tilde{x}^*(\omega), \omega)G^\dagger(\omega) = I_p, \quad \|G^\dagger(\omega)\| \leq \frac{1}{\sigma_m}, \quad \forall \omega \in \Omega. \quad (3.6)$$

The pseudoinverse exists following Assumption [A1].

3.3. Discussion of the Assumptions. We now discuss the restrictions imposed by and implications of the assumptions [A1]–[A5].

3.3.1. Existence and Properties of the Optimal Map $\tilde{x}^*(\omega)$. Assumptions [A2] and [A3], which make no mention of the solution map $\tilde{x}^*(\omega)$, are key ingredients for the existence of such a solution map. For the purpose of this discussion, assume that $g(x, \omega)$ is feasible for any ω . Specifically assume that there exists $D_\Omega > 0$ such that for any ω there exists $\hat{x}(\omega)$ such that $\|\hat{x}(\omega)\| \leq D_\Omega$ and $g(\hat{x}(\omega), \omega) = 0$. It follows that

$$f(\hat{x}(\omega), \omega) \leq \max_{\|x\| \leq D_\Omega, \omega \in \Omega} f(x, \omega) \triangleq M_{D_\Omega},$$

where the maximum can be taken because the sphere of radius D_Ω is compact in \mathbb{R}^n , Ω is compact, and, from Assumption [A3], $f(x, \omega)$ is continuous. Define now the set

$$\hat{X}_{M_{D_\Omega}} = \{x | \exists \omega \in \Omega, f(x, \omega) \leq M_{D_\Omega}\},$$

which, following Assumption [A2], is included in a level set of χ and thus must be a compact set. Therefore, for any ω the problem

$$\min_{x \in \hat{X}_{M_{D_\Omega}}} f(x, \omega) \text{ subject to } g(x, \omega)$$

is feasible, and thus, since it represents minimization of a continuous function over a compact set, it also has at least one solution, which we denote by $\tilde{x}^*(\omega)$, which must also be a solution of the original problem (O).

Therefore, a mapping $\tilde{x}^*(\omega)$ does exist, under [A2], [A3] and the uniform feasibility assumption. In the case where $g(x, \omega) = g(x)$ (that is, g does not depend on Ω , such as is the case in the eigenvalue problem, where $g(x, \omega) = x^T x - 1$), the uniform feasibility assumption, is immediately satisfied so that a mapping $\tilde{x}^*(\omega)$ does exist.

In addition, if f, g are smooth (infinitely differentiable) in both arguments and if [A1] holds for $\tilde{x}^*(\omega)$, then for any ω there exists a unique Lagrange multiplier $\tilde{\lambda}^*(\omega)$ such that $\nabla_x \mathcal{L}(\tilde{x}^*(\omega), \tilde{\lambda}^*(\omega), \omega) = 0$ [4, 21]. Here $\mathcal{L}(x, \lambda, \omega) = f(x, \omega) + \lambda^T g(x, \omega)$ is the Lagrangian function. Finally, if the following second-order sufficient condition holds,

$$\min_{\|u\|=1, \nabla_x g(\tilde{x}^*(\omega), \omega)u=0} u^T \nabla_{xx}^2 \mathcal{L}(\tilde{x}^*(\omega), \tilde{\lambda}^*(\omega), \omega)u > 0, \quad \forall \omega \in \Omega$$

then $\tilde{x}^*(\omega)$ is a locally unique solution of (O), and it is a smooth mapping of ω [4].

3.3.2. Significance of Assumption [A5]. All the assumptions invoked here are standard fare except for [A5]. Note that it is immediate, from [A3] and from the fact that Ω is compact, that we have that $C_G < \infty$. In addition, we have from the definition of the Euclidean norm and (3.5) that

$$\|J^{K,Q}\| = \sup_{\substack{u_i \in \mathbb{R}^n, \\ i=0,1,\dots,K \\ \sum_{i=0}^K \|\lambda_i\|^2 = 1}} \sup_{\substack{\lambda_j \in \mathbb{R}^p, \\ j-K=1,2,\dots,Q \\ \sum_{j=K+1}^{K+Q} \|\lambda_j\|^2 = 1}} \begin{pmatrix} u_0 \\ u_1 \\ \vdots \\ u_K \end{pmatrix}^T J^{KQ} \begin{pmatrix} \lambda_{K+1} \\ \lambda_{K+2} \\ \vdots \\ \lambda_{K+Q} \end{pmatrix}.$$

Examining the last expression and the definition of the blocks $J_{i,j}$ of $J^{K,Q}$, we obtain that

$$\begin{aligned} \begin{pmatrix} u_0 \\ u_1 \\ \vdots \\ u_K \end{pmatrix}^T J^{KQ} \begin{pmatrix} \lambda_{K+1} \\ \lambda_{K+2} \\ \vdots \\ \lambda_{K+Q} \end{pmatrix} &= \sum_{i=0, j=K+1}^{i=K, j=K+Q} \langle u_i^T P_i(\omega) G^\dagger(\omega) P_j(\omega) \lambda_j \rangle = \\ &\left\langle \left(\sum_{i=0}^K u_i^T P_i(\omega) \right) G^\dagger(\omega) \left(\sum_{j=K+1}^{K+Q} P_j(\omega) \lambda_j \right) \right\rangle \quad \text{triangle ineq.} \\ &\leq \\ &\left\langle \|\hat{u}(\omega)\| \|G^\dagger(\omega)\| \|\hat{\lambda}(\omega)\| \right\rangle \quad (3.6), \text{ Cauchy-Schwarz} \\ &\leq \\ &\frac{1}{\sigma_m} \|\hat{u}(\omega)\|_W \|\hat{\lambda}(\omega)\|_W \stackrel{(2.1)}{=} \frac{1}{\sigma_m}, \end{aligned}$$

where

$$\hat{u}(\omega) : \Omega \rightarrow \mathbb{R}^n = \sum_{i=0}^K u_i P_i(\omega), \quad \hat{\lambda}(\omega) : \Omega \rightarrow \mathbb{R}^p = \sum_{j=K+1}^{K+Q} \lambda_j P_j(\omega).$$

In turn, this implies that

$$\|J^{K,Q}\| \leq \frac{1}{\sigma_m}, \forall K, Q \in \mathbb{N}.$$

Tracking back the definition of c in Assumption [A5], we have that $c \leq \frac{1}{\sigma_m}$.

On the other hand, requiring $C_G \frac{1}{\sigma_m} \leq \frac{1}{4}$ instead of $cC_G < \frac{1}{4}$ results in exceedingly conservative bounds. For instance, if the constraint function does not depend explicitly on ω and is linear in x , then it immediately follows that $J^{K,Q} = 0$, $\forall K, Q \in \mathbb{N}$ and thus $c = 0$, even if $\frac{1}{\sigma_m} > 0$. A perturbation argument applied to the same circumstance (the independence of ω situation) implies that the condition $cC_G \leq \frac{1}{4}$ is satisfied if the dependence of g on ω is sufficiently weak.

We conclude that [A5] is implied by sufficiently small variation of g with respect to ω (a condition that is far weaker than assuming that $\frac{1}{\sigma_m}$ is less than $\frac{1}{4C_G}$).

3.4. Solvability and Convergence Results. Notation We use the notation $\tilde{x}^{*,K}(\omega) = \Pi_W^K \tilde{x}^*(\omega)$.

Define now

$$G_2^K(\lambda, u) = \sum_{i,j=0}^{M_K} \lambda_i^T \langle P_i \nabla_x g(\tilde{x}^{*,K}(\omega), \omega) P_j \rangle u_j,$$

$$G^K = \inf_{\substack{\lambda_i \in \mathbb{R}^p, \\ i=0, 1, \dots, M_K \\ \sum_{i=0}^{M_K} \|\lambda_i\|^2 = 1}} \sup_{\substack{u_j \in \mathbb{R}^n, \\ j=0, 1, \dots, M_K \\ \sum_{j=0}^{M_K} \|u_j\|^2 = 1}} G_2^K(\lambda, u).$$

LEMMA 3.3. Define $\Gamma^K = A_0 - A_1 \|\tilde{x}^* - \tilde{x}^{*,K}\|_\infty - A_2 \|\tilde{x}^* - \tilde{x}^{*,K}\|_\infty^2$, where

$$A_0 = 1 - 4c^2 C_G^2, \quad A_1 = 8c^2 C_G L, \quad \text{and} \quad A_2 = 4c^2 L^2 + 2 \frac{L^2}{\sigma_m^2}. \quad (3.7)$$

Then, if $\Gamma^K > 0$, it follows that $c\sigma_m < 1$ and

$$G^K \geq \sqrt{\frac{4\Gamma^K}{\frac{1}{\sigma_m^2} - c^2}}.$$

Notation For simplicity, we use the notation $\tilde{x} = \tilde{x}^*$ and $\tilde{x}^K = \tilde{x}^{*,K}$.

Proof Define

$$\Theta^K = \left\{ \tilde{\lambda}(\omega) = \sum_{i=0}^{M_K} \lambda_i P_i(\omega) \left| \lambda_i \in \mathbb{R}^p, i=0, 1, \dots, M_K, \sum_{i=0}^{M_K} \|\lambda_i\|^2 = 1 \right. \right\} \quad (3.8)$$

$$\Upsilon^K = \left\{ \tilde{u}(\omega) = \sum_{i=0}^{M_K} u_i P_i(\omega) \left| u_i \in \mathbb{R}^n, i=0, 1, \dots, M_K, \sum_{i=0}^{M_K} \|u_i\|^2 = 1 \right. \right\} \quad (3.9)$$

It immediately follows from (2.1) that $\tilde{\lambda} \in \Theta^K$ implies that $\|\tilde{\lambda}(\omega)\|_W = 1$, and $\tilde{u} \in \Upsilon^K$ implies that $\|\tilde{u}(\omega)\|_W = 1$. We will use \tilde{u} and $\tilde{\lambda}$ as the functional image of $\{\lambda_k\}_{k=0,1,\dots,M_K}$, and, respectively, $\{u_k\}_{k=0,1,\dots,M_K}$. We have that

$$G_2^K(\lambda, u) = \left\langle \tilde{\lambda}(\omega)^T \nabla_x g(\tilde{x}^K(\omega), \omega) \tilde{u}(\omega) \right\rangle.$$

We define $\mathbb{R}^n \ni e_{nK} = \frac{1}{\sqrt{n(M_K+1)}} (1, 1, \dots, 1)^T$ and

$$0 \leq H^K(\tilde{\lambda}) = \sum_{i=0}^{M_K} \left| \left\langle \tilde{\lambda}^T(\omega) \nabla_x g(\tilde{x}^K(\omega), \omega) P_i(\omega) \right\rangle \right|^2. \quad (3.10)$$

We now define

$$u_i = \begin{cases} \frac{1}{\sqrt{H^K(\tilde{\lambda})}} \left\langle \tilde{\lambda}^T(\omega) \nabla_x g(\tilde{x}^K(\omega), \omega) P_i(\omega) \right\rangle & H^K(\tilde{\lambda}) \neq 0 \\ e_{nK} & H^K(\tilde{\lambda}) = 0. \end{cases}, \quad i = 0, 1, \dots, M_K$$

which results in $\tilde{u} \in \Upsilon^K$. With this choice we get that $G(\tilde{\lambda}, \tilde{u}) = \sqrt{H^K(\tilde{\lambda})}$, and, using the expression of G^K , we obtain that

$$G^K \geq \inf_{\tilde{\lambda} \in \Theta^K} \sqrt{H^K(\tilde{\lambda})}.$$

So we now proceed to bound below $H^K(\tilde{\lambda})$.

From the definition of $G^\dagger(\omega)$, we have that

$$\tilde{\lambda}(\omega)^T = \tilde{\lambda}(\omega)^T \nabla_x g(\tilde{x}(\omega), \omega) G^\dagger(\omega), \quad \forall \omega \in \Omega. \quad (3.11)$$

Define

$$\begin{aligned} \mathbb{R}^n \ni \theta_i(\tilde{\lambda}) &= \left\langle \tilde{\lambda}(\omega)^T \nabla_x g(\tilde{x}^K(\omega), \omega) P_i(\omega) \right\rangle \\ \mathbb{R}^{n \times p} \ni \mu_{ik}(G^\dagger) &= \left\langle P_i(\omega) G^\dagger(\omega) P_k(\omega) \right\rangle. \end{aligned}$$

We now discuss the well-posedness of the preceding quantities. Since $\tilde{\lambda}$ is a polynomial and, from assumptions [A3] and [A4] the function $\tilde{\lambda}^T(\omega) \nabla_x g(\tilde{x}^K(\omega), \omega)$ is continuous and Ω is compact, it follows that $\left\| \tilde{\lambda}(\omega) \nabla_x g(\tilde{x}^K(\omega), \omega) \right\|_W^2 < \infty$. This, in turn, means that $\theta_i(\tilde{\lambda})$ is well defined (the integral that defines it is absolutely convergent), and

$$\begin{aligned} \sum_{i=0}^{\infty} \left\| \theta_i(\tilde{\lambda}) \right\|^2 &\stackrel{(2.1)}{=} \left\| \tilde{\lambda}(\omega) \nabla_x g(\tilde{x}^K(\omega), \omega) \right\|_W^2 = \left\langle \left\| \tilde{\lambda}(\omega) \nabla_x g(\tilde{x}^K(\omega), \omega) \right\|^2 \right\rangle \\ &\leq \left\langle \left\| \tilde{\lambda}(\omega) \right\|^2 \left\| \nabla_x g(\tilde{x}^K(\omega), \omega) \right\|^2 \right\rangle \\ &\leq \left\| \tilde{\lambda}(\omega) \right\|_W^2 \left\| \nabla_x g(\tilde{x}^K(\omega), \omega) \right\|_{\infty}^2 \\ &= \left\| \nabla_x g(\tilde{x}^K(\omega), \omega) \right\|_{\infty}^2 \leq (C_G + L \|\tilde{x} - \tilde{x}^K\|_{\infty})^2, \end{aligned} \quad (3.12)$$

where the last inequality follows from $\left\| \tilde{\lambda}(\omega) \right\|_W^2 = 1$ (since $\tilde{\lambda} \in \Theta^K$), the triangle inequality

$$\begin{aligned} \left\| \nabla_x g(\tilde{x}^K(\omega), \omega) \right\|_{\infty} &\leq \left\| \nabla_x g(\tilde{x}(\omega), \omega) \right\|_{\infty} + \left\| \nabla_x g(\tilde{x}^K(\omega), \omega) - \nabla_x g(\tilde{x}(\omega), \omega) \right\|_{\infty} \\ &\leq C_G + L \|\tilde{x} - \tilde{x}^K\|_{\infty}, \end{aligned}$$

and the notations from Assumption [A5].

From (3.11), using the extension of $\langle h, l \rangle_W = \sum_{i=0}^{\infty} c_i(h) c_i(l)$, that holds for $h, l \in L_W^2$, to matrix-valued mappings, we obtain that, for $\forall k \leq M_K$, we have that

$$\begin{aligned} \left\langle P_k \tilde{\lambda}(\omega)^T \right\rangle &= \sum_{i=1}^{\infty} \left\langle \tilde{\lambda}(\omega)^T \nabla_x g(\tilde{x}(\omega), \omega) P_i(\omega) \right\rangle \mu_{ik}(G^\dagger) = \sum_{i=1}^{\infty} \theta_i(\tilde{\lambda})^T \mu_{ik}(G^\dagger) + \\ &\quad \left\langle P_k(\omega) \tilde{\lambda}(\omega)^T (\nabla_x g(\tilde{x}(\omega), \omega) - \nabla_x g(\tilde{x}^K(\omega), \omega)) G^\dagger(\omega) \right\rangle. \end{aligned}$$

Since $\tilde{\lambda} \in \Theta^K$ and from the preceding equation, we have that

$$1 = \sum_{k=0}^{M_K} \left\| \left\langle P_k, \tilde{\lambda}(\omega)^T \right\rangle \right\|^2 \leq 2 \sum_{k=0}^{M_K} \left\| \sum_{i=0}^{\infty} \theta_i(\tilde{\lambda})^T \mu_{ik}(G^\dagger) \right\|^2 + 2T_3 \leq 4(T_1 + T_2) + 2T_3, \quad (3.13)$$

where the last two inequalities follow from the inequality $\|a + b\|^2 \leq 2(\|a\|^2 + \|b\|^2)$ applied twice and from Bessel's identity (2.1) where

$$\begin{aligned} T_1 &= \sum_{k=0}^{M_K} \left\| \sum_{i=0}^{M_K} \theta_i(\tilde{\lambda})^T \mu_{ik}(G^\dagger) \right\|^2 & T_2 &= \sum_{k=0}^{M_K} \left\| \sum_{i=M_K+1}^{\infty} \theta_i(\tilde{\lambda})^T \mu_{ik}(G^\dagger) \right\|^2 \\ T_3 &= \sum_{k=0}^{M_K} \left\| \left\langle P_k(\omega) \tilde{\lambda}(\omega)^T (\nabla_x g(\tilde{x}(\omega), \omega) - \nabla_x g(\tilde{x}^K(\omega), \omega)) G^\dagger(\omega) \right\rangle \right\|^2. \end{aligned}$$

We now find upper bounds on T_1 , T_2 , and T_3 . We define $\tilde{\theta}(\omega) = \sum_{i=0}^{M_K} \theta_i(\tilde{\lambda}) P_i(\omega)$. We obtain that

$$\begin{aligned} T_1 &= \sum_{k=0}^{M_K} \left\| \left\langle \tilde{\theta}(\omega)^T G^\dagger(\omega) P_k \right\rangle \right\|^2 \stackrel{(2.1)}{\leq} \left\langle \left\| \tilde{\theta}(\omega)^T G^\dagger(\omega) \right\|^2 \right\rangle \leq \left\langle \left\| \tilde{\theta}(\omega) \right\|^2 \left\| G^\dagger(\omega) \right\|^2 \right\rangle \\ &\stackrel{\text{by (3.6)}}{\leq} \frac{1}{\sigma_m^2} \left\langle \left\| \tilde{\theta}(\omega) \right\|^2 \right\rangle \stackrel{(2.1)}{=} \frac{1}{\sigma_m^2} \sum_{k=0}^{M_K} \left\| \theta_k(\tilde{\lambda}) \right\|^2 \stackrel{(3.10)}{=} \frac{H^K(\tilde{\lambda})}{\sigma_m^2}. \end{aligned}$$

Using [A5], we obtain that

$$\begin{aligned} T_2 &\leq c^2 \sum_{k=M_K+1}^{\infty} \left\| \theta_k(\tilde{\lambda}) \right\|^2 = c^2 \sum_{k=0}^{\infty} \left\| \theta_k(\tilde{\lambda}) \right\|^2 - c^2 \sum_{k=0}^{M_K} \left\| \theta_k(\tilde{\lambda}) \right\|^2 \\ &\stackrel{\text{by (3.12), (3.10)}}{\leq} c^2 (C_G + L \|\tilde{x} - \tilde{x}^K\|)^2 - c^2 H^K(\tilde{\lambda}). \end{aligned}$$

Finally, using Bessel's identity (2.1), Cauchy-Schwarz, [A5], and that $\|\tilde{\lambda}(\omega)\|_W = 1$, which follows from $\tilde{\lambda} \in \Theta^K$, we obtain that

$$\begin{aligned} T_3 &\stackrel{(2.1)}{\leq} \left\| \tilde{\lambda}(\omega)^T (\nabla_x g(\tilde{x}(\omega), \omega) - \nabla_x g(\tilde{x}^K(\omega), \omega)) G^\dagger(\omega) \right\|_W^2 \\ &\leq \left\langle \left\| \tilde{\lambda}(\omega) \right\|^2 \left\| (\nabla_x g(\tilde{x}(\omega), \omega) - \nabla_x g(\tilde{x}^K(\omega), \omega)) \right\|^2 \left\| G^\dagger(\omega) \right\|^2 \right\rangle \\ &\stackrel{\text{by [A5], (3.6)}}{\leq} \left(\frac{L}{\sigma_m} \right)^2 \|\tilde{x} - \tilde{x}^K\|_\infty^2. \end{aligned}$$

Replacing the bounds obtained for T_1 , T_2 , and T_3 in (3.13), we obtain that

$$4H^K(\tilde{\lambda}) \left(\frac{1}{\sigma_m^2} - c^2 \right) \geq \Gamma^K = A_0 - A_1 \|\tilde{x} - \tilde{x}^K\|_\infty - A_2 \|\tilde{x} - \tilde{x}^K\|_\infty^2,$$

where A_0, A_1, A_2 are defined in (3.7). Since $\Gamma^K > 0$ implies that $A_0 > 0$, which in turn implies that $cC_G < \frac{1}{4}$, we get, from [A1], that $c\sigma_m < \frac{1}{4}$. The conclusion follows from the preceding displayed inequality. \square

A key point of our analysis consists of obtaining bounds between $\|\tilde{x}^K\|_\infty$ and $\|\tilde{x}^K\|_W$.

LEMMA 3.4.

$$\|\tilde{x}^K\|_W \leq \|\tilde{x}^K\|_\infty \leq \Lambda^K \|\tilde{x}^K\|_W$$

Proof It is immediate that $\|\tilde{x}^K\|_W \leq \|\tilde{x}^K\|_\infty$. We have, using Cauchy-Schwarz, that

$$\|\tilde{x}^K(\omega)\| = \left\| \sum_{k=0}^{M_K} x_k P_k(\omega) \right\| \leq \sum_{k=0}^{M_K} \|x_k\| \|P_k(\omega)\| \leq \sqrt{\sum_{k=0}^{M_K} \|x_k\|^2} \sqrt{\sum_{k=0}^{M_K} |P_k(\omega)|^2}.$$

From the definition of Λ^K , (2.4), we obtain that

$$\sup_{\omega \in \Omega} \|\tilde{x}^K(\omega)\| \leq \Lambda^K \|\tilde{x}^K(\omega)\|_W,$$

which proves the claim. \square

LEMMA 3.5.

$$\|J^K(\tilde{x}_1(\omega)) - J^K(\tilde{x}_2(\omega))\| \leq L\Lambda^K \|\tilde{x}_1(\omega) - \tilde{x}_2(\omega)\|_W$$

Notation We will denote $\tilde{x}_1 = \tilde{x}_1(\omega)$, $\tilde{x}_2 = \tilde{x}_2(\omega)$.

Proof By algebraic manipulations and notations similar to the ones in Lemma 3.3, we obtain that

$$\begin{aligned} & \|J^K(\tilde{x}_1) - J^K(\tilde{x}_2)\| &= \\ & \sup_{\substack{\lambda_i \in \mathbb{R}^p, u_j \in \mathbb{R}^n \\ i=0, 1, \dots, M_K \\ \sum_{i=0}^{M_K} \|\lambda_i\|^2 = 1 \\ \sum_{j=0}^{M_K} \|u_j\|^2 = 1}} \sum_{i,j=0}^{M_K} \lambda_i^T \langle P_i (\nabla_x g(\tilde{x}_1, \omega) - \nabla_x g(\tilde{x}_2, \omega)) P_j \rangle u_j &= \\ & \sup_{\tilde{\lambda} \in \Theta^K, \tilde{u} \in \Upsilon^K} \left\langle \tilde{\lambda}(\omega)^T (\nabla_x g(\tilde{x}_1, \omega) - \nabla_x g(\tilde{x}_2, \omega)) \tilde{u}(\omega) \right\rangle &\stackrel{\text{by [A5]}}{\leq} \\ L \|\tilde{x}_2 - \tilde{x}_1\|_\infty \sup_{\tilde{\lambda} \in \Theta^K, \tilde{u} \in \Upsilon^K} \left\langle \|\tilde{\lambda}(\omega)\| \|\tilde{u}(\omega)\| \right\rangle &\stackrel{\text{Cauchy-Schwarz}}{\leq} L \|\tilde{x}_2 - \tilde{x}_1\|_\infty \stackrel{\text{by Lemma 3.4}}{\leq} \\ & L\Lambda^K \|\tilde{x}_2 - \tilde{x}_1\|, \end{aligned}$$

which completes the claim. \square

LEMMA 3.6. *The objective function of the problem (SO(K)) has bounded level sets.*

Proof Take $\tilde{x}^K(\omega) = \sum_{k=0}^{M_K} x_k P_k(\omega)$. Consider the level set of height M of $\tilde{f}^K(\tilde{x}^K) = \langle f(\tilde{x}^K(\omega), \omega) \rangle$,

$$\mathcal{L}_K(M) = \left\{ (x_0, x_1, \dots, x_{M_K}) \in \mathbb{R}^{mM_K} \mid \tilde{f}^K(\tilde{x}) \leq M \right\}.$$

Using Assumption [A2], we obtain that, if $(x_0, x_1, \dots, x_{M_K}) \in \mathcal{L}_K(M)$, then

$$\begin{aligned} M &> \tilde{f}(\tilde{x}^K) = \langle f(\tilde{x}^K(\omega), \omega) \rangle \stackrel{[\text{A2}]}{\geq} \left\langle \chi \left(\|\tilde{x}^K(\omega)\|^\gamma \right) \right\rangle \\ &\stackrel{\text{by Jensen's inequality}}{\geq} \chi \left(\left\langle \|\tilde{x}^K(\omega)\|^\gamma \right\rangle \right) \implies \left\langle \|\tilde{x}^K(\omega)\|^\gamma \right\rangle \in \mathcal{L}_M^\chi. \end{aligned} \quad (3.14)$$

We denote

$$L_K^\gamma = \min_{\sum_{k=0}^{M_K} \|x_k\|^2 = 1} \left\langle \|\tilde{x}^K(\omega)\|^\gamma \right\rangle.$$

Since the unit ball $\mathcal{B}^K \in \mathbb{R}^{n(M_K+1)}$, defined as

$$\mathcal{B}^K = \left\{ (x_0, x_1, \dots, x_{M_K}) \in \mathbb{R}^{n(M_K+1)} \mid \sum_{k=0}^{M_K} \|x_k\|^2 = 1 \right\},$$

is a compact set, the quantity L_K^γ is well defined.

It also immediately follows that $L_K^\gamma > 0, \forall K > 0$. Indeed, if there existed a K for which $L_K^\gamma = 0$, it would follow that for some choice of x_0, x_1, \dots, x_K , such that $\sum_{k=0}^{M_K} \|x_k\|^2 = 1$, we have that $\tilde{x}^K(\omega) = 0, \forall \omega \in \Omega$, which contradicts the fact that $P_k(\omega)$ are linearly independent because they are a subset of a basis.

In return (3.14) results in $\chi \left(L_K^\gamma \left(\sum_{k=0}^{M_K} \|x_k\|^2 \right) \right) \leq M$, which, in turn, results in $\left(\sum_{k=0}^{M_K} \|x_k\|^2 \right) \leq \frac{\chi^{-1}(M)}{L_K^\gamma}$. Since we assumed that the function χ has bounded level sets, the conclusion follows. \square

Note Lemma 3.6 ensures that the solution of the systems of nonlinear equations that defines the spectral spectral finite element method [15, 14] does exist, at least for the case where the system of nonlinear equations is derived from the optimality conditions of an unconstrained optimization problem. The same result can be obtained for constrained problems from Lemma 3.3 for all K when $g(x, \omega)$ is linear in x and does not depend on ω , since [A5] is satisfied with $c = 0$. To our knowledge, this is a new result for the case where the the variation of $\tilde{x}^*(\omega)$ is not necessarily small.

LEMMA 3.7. *Assume that*

- (a) $\lim_{K \rightarrow \infty} \|\tilde{x}^* - \Pi_W^K \tilde{x}^*\|_\infty = 0$ and
- (b) $\lim_{K \rightarrow \infty} \Lambda^K \|\tilde{x}^* - \Pi_W^K \tilde{x}^*\|_W = 0$.

For any $r > 0$, there exists K_0 such that (SO(K)) has a feasible point $\bar{x}^K(\omega)$ that satisfies $\|\bar{x}^K - \Pi_W^K \tilde{x}^*\|_W \leq r, \forall K \geq K_0$.

Proof We seek to apply Kantorovich's Theorem 3.2. With the notations in the assumptions [A2]–[A5], and from the definition of G^K preceding Lemma 3.3 and from the definition of Λ^K (2.4), it follows, using Lemma 3.5, that the conditions of the theorem are satisfied at $\tilde{x}^{*,K} = \Pi_W^K(\tilde{x}^*)$ provided the following two conditions hold:

$$(i) \ h = \frac{L\Lambda^K (G^K)^2 g^K}{2} < 1, \quad (ii') \ \rho = \frac{2H(h)}{L\Lambda^K G^K} < r,$$

where $g^K = \sqrt{\sum_{k=0}^{M_K} \|\langle g(\tilde{x}^{*,K}(\omega), \omega) P_k \rangle\|^2}$. Note that if $h < \frac{1}{2}$, we have that $h \leq H(h) \leq 2h$. Therefore a sufficient condition for the condition (ii) to hold is

$$(ii) \ 4G^K g^K < r.$$

We have that

$$(g^K)^2 = \sum_{k=0}^{M_K} \|\langle g(\tilde{x}^{*,K}(\omega), \omega) P_k \rangle\|^2 \stackrel{(2.1)}{\leq} \|g(\tilde{x}^{*,K}(\omega), \omega)\|_W^2.$$

From Leibnitz-Newton and assumption [A3] we get

$$g(\tilde{x}^{*,K}(\omega), \omega) = g(\tilde{x}^*(\omega), \omega) + \int_0^1 \nabla_x g(\bar{x}(t, \omega), \omega) (\tilde{x}^{*,K}(\omega) - \tilde{x}^*(\omega)) dt,$$

where $\bar{x}(t, \omega) = t\tilde{x}^{*,K}(\omega) + (1-t)\tilde{x}^*(\omega)$. Since $\tilde{x}^*(\omega)$ is a solution of (O), for any $\omega \in \Omega$, we get $g(\tilde{x}^*(\omega), \omega) = 0, \forall \omega \in \Omega$. Using Assumptions [A5], we obtain the following

$$\begin{aligned} \|g(\tilde{x}^{*,K}(\omega), \omega)\| &\leq \|\tilde{x}^{*,K}(\omega) - \tilde{x}^*(\omega)\| \int_0^1 (C_G + L \|\bar{x}(t, \omega) - \tilde{x}^*(\omega)\|) dt \\ &\leq \left(C_G + \frac{L}{2} \|\tilde{x}^{*,K}(\omega) - \tilde{x}^*(\omega)\| \right) \|\tilde{x}^{*,K}(\omega) - \tilde{x}^*(\omega)\|. \end{aligned}$$

In turn, this implies that

$$g^K \leq \|g(\tilde{x}^{*,K}(\omega), \omega)\|_W \leq \left(C_G + \frac{L}{2} \|\tilde{x}^{*,K} - \tilde{x}\|_\infty \right) \|\tilde{x}^{*,K} - \tilde{x}^*\|_W. \quad (3.15)$$

From assumption (a) of this theorem, we have that $\exists K_0$ such that, $\forall K \geq K_0$,

$$L \|\tilde{x}^* - \tilde{x}^{*,K}\|_\infty \leq C_G, \quad A_1 \|\tilde{x}^* - \tilde{x}^{*,K}\|_\infty + A_2 \|\tilde{x}^* - \tilde{x}^{*,K}\|_\infty^2 \leq \frac{A_0}{2},$$

where A_0 , A_1 , and A_2 are defined in (3.7). With the notations of Lemma 3.3, $\Gamma^K \geq \frac{A_0}{2}$, and thus from Assumption [A5], which ensures that $A_0 > 0$, we get that $G^K \geq A_3 \triangleq \sqrt{\frac{2A_0}{\frac{1}{\sigma_m^2} - c}} > 0$. Therefore, for $K \geq K_0$ and from (i),(ii), and (3.15) we get that the conditions of Kantorovich's theorem 3.2 are satisfied provided that

$$2LA_3^2 C_G \Lambda^K \|\tilde{x}^* - \tilde{x}^{*,K}\|_W \leq 1, \quad 8A_3 C_G \|\tilde{x}^* - \tilde{x}^{*,K}\|_W \leq r.$$

From assumptions (a) and (b), it follows that these conditions are satisfied, by eventually choosing a larger K_0 , for all $K \geq K_0$. Therefore, Kantorovich's theorem 3.2 applies to give the conclusion. The proof is complete. \square

THEOREM 3.8. *Assume that $\tilde{x}^*(\omega)$ is smooth (infinitely differentiable). Then there exists K_0 such that $(SO(K))$ has a solution, $\forall K \geq K_0$.*

Proof The key of the proof is that we are able to choose q as large as necessary in (2.2) for $f = \tilde{x}^*$. Choose $q = t + m + 2$. We obtain from (2.3) that

$$\|\tilde{x}^* - \Pi_W^K \tilde{x}^*\|_\infty \leq mC_S \sum_{k=M_{K+1}+1}^{\infty} \|c_k(f)\| \deg(P_k)^t.$$

Since the number of polynomials of degree at most K is $\binom{m+K}{m}$ [10] we obtain from (2.2) and (2.1) that $\|c_k(f)\| \leq CQ^{-q}$, and from the preceding displayed equation and (2.3), that

$$\begin{aligned} \|\tilde{x}^* - \Pi_W^K \tilde{x}^*\|_\infty &\leq mCC_S \sum_{Q=K+1}^{\infty} \binom{m+Q}{m} Q^{-(t+m+2)} Q^t \\ &\leq \frac{CC_S}{(m-1)!} \sum_{Q=K+1}^{\infty} Q^{-2} \left(\frac{m+Q}{Q} \right)^m \xrightarrow{K \rightarrow \infty} 0 \end{aligned}$$

and thus

$$\lim_{K \rightarrow \infty} \|\tilde{x}^* - \Pi_W^K \tilde{x}^*\|_\infty = 0. \quad (3.16)$$

In addition, from (2.4) and (2.2) we obtain that

$$\Lambda^K \|\tilde{x}^* - \Pi_W^K \tilde{x}^*\|_W \leq C \frac{1}{K^q} \binom{m+K}{m}^d \leq C \frac{1}{m! K^{q-dm}} \left(\frac{m+K}{K} \right)^{md}.$$

Therefore, if we choose $q \geq md + 1$, we get that $\Lambda^K \|\tilde{x}^* - \tilde{x}^{*,K}\|_W \xrightarrow{K \rightarrow \infty} 0$. From (3.16), conditions (a) and (b) of Lemma 3.7 are satisfied. We apply Lemmas 3.7 and

3.6 to obtain that problem (SO(K)) is feasible for $K \geq K_0$ and has bounded level sets and thus has a solution [21]. The proof is complete. \square

THEOREM 3.9. *Let $m = 1$ and $W(x) = \sqrt{1-x^2}^{-1}$ (the Chebyshev polynomials case). Then $SO(K)$ has a solution for all $K \geq K_0$.*

Proof From [A4], $\tilde{x}^*(\omega)$ is continuous and has bounded variation; therefore $\|\tilde{x}^*(\omega) - \Pi_W^K \tilde{x}^*(\omega)\|_\infty \rightarrow 0$ as $K \rightarrow \infty$ [18, Theorem 1]. Also, from [A4], (2.4), and (2.2) we obtain that

$$\Lambda^K \|\tilde{x}^*(\omega) - \Pi_W^K \tilde{x}^*(\omega)\|_W \leq K^{\frac{1}{2}} C \frac{1}{K} \xrightarrow{K \rightarrow \infty} 0.$$

Conditions (a) and (b) of Lemma 3.7 therefore are satisfied. Therefore Lemmas 3.7 and 3.6 apply to give that the problem (SO(K)) is feasible for $K \geq K_0$ and its objective function has bounded level sets. Therefore (SO(K)) is solvable [21], and the proof is complete. \square

Discussion Theorem 3.8 completely addresses the issue of solvability of (SO(K)) in the case of smooth solution functions, independent of the dimension of the problem. The result for nonsmooth solution functions Theorem 3.9 is restrictive in terms of both dimensions and polynomial type, and its extension is deferred to future research.

Finally, we approach the issue of limits of solutions of (SO(K)) for increasing K . For convergence as $K \rightarrow \infty$ we need to invoke stronger assumptions, that allow us to guarantee the existence of convergent subsequences.

THEOREM 3.10. *Assume that the conditions of Theorem 3.8 are satisfied, that the sequence of solutions of the problem (SO(K)) satisfies the Kuhn-Tucker conditions, and that there exists a $C_X > 0$ such that the solution and multiplier sequences (λ_k^K, x_k^K) satisfy*

$$\sum_{k=0}^{M_K} \|\lambda_k^K\| \deg(P_k)^t < C_X; \quad \sum_{k=0}^{M_K} \|x_k^K\| \deg(P_k)^t < C_X,$$

where t is the parameter from (2.3). Define $\tilde{\lambda}^K(\omega) = \sum_{k=0}^{M_K} \lambda_k^K P_k(\omega)$ and $\tilde{x}^K(\omega) = \sum_{k=0}^{M_K} x_k^K P_k(\omega)$. Then the sequence $(\tilde{x}^K(\omega), \tilde{\lambda}^K(\omega))$ has a uniformly convergent subsequence. Any limit $(\hat{x}(\omega), \hat{\lambda}(\omega))$ of such a subsequence satisfies the nonlinear system of equations (3.1).

Proof From (2.3) it follows that the sequence $(\tilde{\lambda}^K(\omega), \tilde{x}^K(\omega))$ satisfies

$$\forall \omega_1, \omega_2 \in \Omega \left\{ \begin{array}{l} \|\tilde{\lambda}^K(\omega_1)\|, \|\tilde{x}^K(\omega_1)\| \leq C_S C_X, \\ \|\tilde{\lambda}^K(\omega_1) - \tilde{\lambda}^K(\omega_2)\|, \|\tilde{x}^K(\omega_1) - \tilde{x}^K(\omega_2)\| \leq C_S C_X \|\omega_1 - \omega_2\|. \end{array} \right.$$

Therefore the families $\tilde{\lambda}^K(\omega), \tilde{x}^K(\omega)$ are equicontinuous and equibounded. We can apply the Arzela-Ascoli theorem [16, Theorem 6.41] to determine that there exists a uniformly convergent subsequence, $\tilde{x}^{K_l}, \tilde{\lambda}^{K_l}$ with a corresponding limit pair. Let $(\hat{x}, \hat{\lambda})$ be such a limit function pair, which must also be continuous because the convergence of the subsequence of Lipschitz functions is continuous. Using Theorem (3.1), we get that $\tilde{x}^{K_l}, \tilde{\lambda}^{K_l}$ satisfies the equation (3.2), for $l \geq 0$. Using assumptions [A3] and [A4], we can take the limit in that equation and obtain that

$$\begin{aligned} \left\langle P_k(\omega) \left(\nabla_x f(\hat{x}(\omega), \omega) + \left(\hat{\lambda}(\omega) \right)^T \nabla_x g(\hat{x}(\omega), \omega) \right) \right\rangle &= 0_n, \quad k \geq 0, \\ \langle P_k(\omega) g(\hat{x}(\omega), \omega) \rangle &= 0_p, \quad k \geq 0. \end{aligned}$$

From Bessel's identity (2.1), we get that

$$\left\| \left(\nabla_x f(\hat{x}(\omega), \omega) + \hat{\lambda}(\omega)^T \nabla_x g(\hat{x}(\omega), \omega) \right) \right\|_W^2 + \|g(\hat{x}(\omega), \omega)\|_W^2 = 0,$$

which, in turn, proves our claim. The proof is complete. \square

Discussion Of course, it would be important to prove the convergence of the approximating sequences $\tilde{x}^K(\omega)$ and $\tilde{\lambda}^K(\omega)$ without assuming that they exhibit sufficient smoothness in the limit. For this initial investigation, we provide this limited result, and we defer the issue of extending it to further research. A promising approach seems to be to quantify the uniform validity with ω of the second order sufficient conditions for problem (O) and infer the smoothness in the limit from it.

4. Applications and Numerical Examples. Motivating our investigation was the study of parametric eigenvalue problems as they appear in neutron diffusion problems in nuclear reactor criticality analysis [9]. We thus investigate how our developments apply to eigenvalue problems.

4.1. Parametric Eigenvalue Problems. In the following, we study our formulation for two parametric eigenvalue problems, of sizes $n = 2$ and $n = 1000$. In the formulation of the problem for both cases is $(Q + \omega D_Q)x(\omega) = \lambda(\omega)x(\omega)$, where Q and D_Q are matrices of size n , $\lambda(\omega)$ and $x(\omega)$ are the smallest eigenvalue and the corresponding eigenvector of the matrix $(Q + \omega D_Q)$. Our theory is applied via the interpretation of the problem as $x(\omega) = \arg \min_{x(\omega)^T x(\omega) = 1} x(\omega)^T (Q + \omega D_Q) x(\omega)$, where $\lambda(\omega)$ is the Lagrange multiplier of the constraint, all for a fixed value of ω . Here, $\omega \in [-1, 1]$, and the spectral finite element problem is constructed by using either Legendre or Chebyshev polynomials [10].

As in our theoretical developments, the problem to be solved has $n \times (M_K + 1)$ unknowns and $M_K + 1$ constraints, and, with the notation $\Phi^K = \{0, 1, \dots, M_K\}$, can be stated as

$$\begin{aligned} \min_{\{x_i\}_{i \in \Phi^K}} \quad & \left\langle \left(\sum_{i=0}^{M_K} x_i P_i(\omega) \right)^T (Q + \omega D_Q) \left(\sum_{i=0}^{M_K} x_i P_i(\omega) \right) \right\rangle \\ \text{s.t. } \forall k \in \Phi^K \quad & \left\langle \left(\sum_{i=0}^{M_K} x_i P_i(\omega) \right)^T \left(\sum_{i=0}^{M_K} x_i P_i(\omega) \right) P_k(\omega) \right\rangle = \langle P_k(\omega) \rangle. \end{aligned}$$

The problem is set up by computing the terms involved after breaking up the parentheses, computing the terms $\langle \omega P_i(\omega) P_j(\omega) \rangle = L_{i,j}$ and $\langle P_i(\omega) P_j(\omega) P_k(\omega) \rangle = \hat{L}_{i,j,k}$. This procedure was carried out by numerical quadrature in MATLAB, after which, the resulting problem became

$$\begin{aligned} \min \quad & \sum_{i=0}^{M_K} x_i Q x_i + \sum_{i,j=0}^{M_K} L_{ij} x_i D_Q x_j \\ \text{s.t.} \quad & \sum_{i,j=0}^{M_K} \hat{L}_{ijk} x_i x_j = E_\omega [P_k(\omega)] \quad k = 0, 1, \dots, M_K. \end{aligned}$$

The problem was coded in AMPL [12] and solved by using the KNITRO interior-point solver [25] which was started with x_0 a vector with all entries $\frac{1}{\sqrt{N}}$, and $x_i = 0_n$, for $i = 0, 1, 2, \dots, M_K$. Once the problem was solved, the parametric approximation of the solution and of the multiplier were constructed as $\tilde{x}(\omega) = \sum_{i=0}^{M_K} x_k P_k(\omega)$, $\lambda(\omega) =$

$$\sum_{i=0}^{M_K} \lambda_k P_k(\omega).$$

It is immediate that the problem satisfies assumptions [A1] and [A3]. Assumption [A2] is satisfied only if the resulting matrix is positive definite for any value of ω , which can be ensured if one adds a suitable fixed multiple of the identity to the matrix. In that case [A2] is satisfied with $\gamma = 1$ and $\chi(r) = r^2$. Since the effect of that is only to shift the λ values, we can assume without loss of generality that [A2] is satisfied. Assumption [A4] holds for both examples and can be verified once the solution is computed. Assumption [A5] is a difficult assumption to verify numerically and, like any small variation assumptions, is bound to be too conservative.

4.2. Problems with Inequality Constraints. It is well known that we can transform an inequality constraint $g_1(x, \omega) \leq 0$ into an equality constraint by using a slack s_1 and representing the inequality as $g_1(x, \omega) + s_1^2 = 0$ [2]. The resulting problem can be represented as (O), and our approach can be used to solve it. To generate the problem (SO(K)), we use a parameterization for the slacks $s_1(\omega) = \sum_{k=0}^K P_k(\omega) s_{1,k}$. When we enforce the constraints of (SO(K)), we get expressions similar to the constraints of the eigenvalue problem in the preceding section, which means that the effect of s_1 on the constraints can be represented finitely in the spectral basis. Therefore, if the functions of the inequality constraints can be represented finitely in the spectral basis, the introduction of slacks will not destroy that. This means our approach and Theorem 3.1 applies to inequality constraints as well, once we have formulated them as slacks. Theorem 3.8 cannot be expected to apply because the solution $\tilde{x}^*(\omega)$ is not smooth in general when inequality constraints are present. Theorem 3.9 may apply but it is limited to the case $m = 1$. For the convergence analysis of problems with inequality constraints, further analysis is necessary.

4.3. The $n = 2$ Problem. For this problem, we chose $Q = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$, $D_Q = \begin{bmatrix} 1 & 0.4 \\ 0.4 & 0.2 \end{bmatrix}$, and we use only Legendre polynomials. We have computed the minimum eigenvalue and the corresponding eigenvector as a function of the parameter ω , both by solving the eigenvalue problem at 100 equally spaced points in the interval between minus 1 and 1, and by using our constrained optimization formulation the spectral finite element method. The results of the two approaches have been plotted in Figure 4.1 for the angle between the eigenvectors obtained by the two approaches and in Figure 4.2 for the eigenvalue. We call here, in Figures 4.1-4.2, and subsequently, the first approach simulation and the second approach ‘‘SFEM’’. It can be seen that the error for the cosine of the angle between eigenvalues is in the seventh decimal place, and the eigenvalue results are virtually indistinguishable. Note that the size of the variation for which we were computing the eigenvector reaches half the size of the maximum element in the original matrix, so the variation is far from being considered small. The results show the soundness of our approach and provide good evidence for convergence. In addition, the solution of the problem seems to be smooth, so the conditions for both Theorems 3.8 and 3.10, as well as their conclusions, appear to be satisfied. In this case the calculation was done with Legendre polynomials with the degree at most 4. Therefore, in terms of the notation that we have used in the theoretical sections, we have that $K = 4$ and $M_K = 4$.

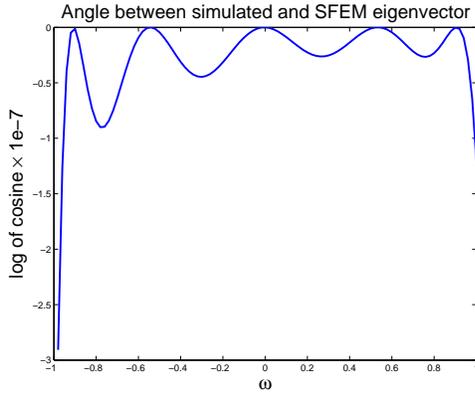


FIG. 4.1. The case $n = 2$. The eigenvector angle error

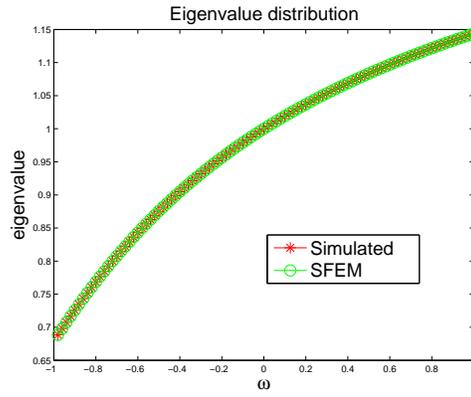


FIG. 4.2. The case $n = 2$. The eigenvalue.

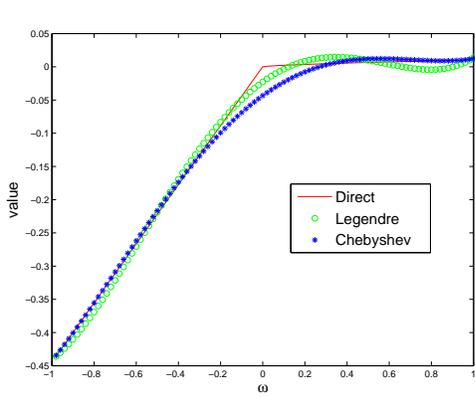


FIG. 4.3. The case $n = 1000$. The eigenvalue.

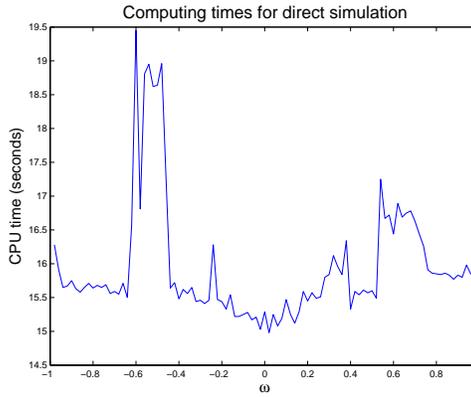


FIG. 4.4. CPU times for the 100 $n = 1000$ eigenvalue calculations by ω value.

4.4. The $n = 1000$ Problem. For this problem, we chose

$$Q_{i,j} = \begin{cases} 1 & i = j = 1 \\ 1 & i = j = n \\ -1 & |i - j| = 1 \\ 2 & 1 < i = j < n \\ 0 & \text{otherwise} \end{cases}, \quad D_{Q_{i,j}} = \begin{cases} 2 \frac{i}{n} \frac{n-i}{n} \cos\left(\frac{i}{n}\right) & i = j \\ 0 & \text{otherwise} \end{cases}.$$

This problem mimics a one-dimensional criticality analysis of the neutron flux in a nuclear reactor [9]. Again, we computed a minimum eigenvalue and the corresponding eigenvector as a function of the parameter ω both by solving the eigenvalue problem at 100 equally spaced points in the interval between minus 1 and 1, as well as by using our constrained optimization formulation of the spectral finite element method. In this case the calculation was done with both Chebyshev and Legendre polynomials with the degree at most 4. Therefore, in terms of the notation that we have used in the theoretical sections, we have that $K = 4$ and $M_K = 4$.

The results are displayed in the Figures 4.3-4.5 for the match between the eigen-

values, as well as the angle between the eigenvectors. We see that the eigenvalues match very well, with a relative error below 5% (with respect to the vector infinity norm). The simulation was run on a Linux workstation, with an Intel XEON 2 GHz CPU, with 512 KB of cache and 1 GB of RAM. Note, however, that the optimization problem was solved in 9.17 seconds for the Legendre polynomials and 14.00 seconds for the Chebyshev polynomials, whereas the direct calculation of the eigenvalue at 100 points took 1,595.10 seconds to compute on the same machine. We provide the CPU times of the eigenvalue calculation at the 100 ω values in Figure 4.4.

The timings statement that we have used compare the result of a Matlab simulation with compiled software, which may initially look suspicious. Note, however, that the objection that MATLAB is much slower for the simulation approach does not apply here, since we have timed only the call to the eigenvalue function in MATLAB, which is an external call to a compiled function.

A more substantial criticism of our comparison may be that we used the `eig` function, which computes all eigenvalues as opposed to only the minimal one at the 100 sample points, which is what is truly needed. One could wonder whether we do not actually waste too much computation because of that. There exist algorithms that determine only the minimum eigenvalue by some iterative procedure [26]. To investigate such an alternative, we have used the `eigs` Matlab function, which uses an Arnoldi iteration, to compute the minimum eigenvalue for $\omega = 0$. The Matlab `eigs` function is an interface to the large-scale eigenvalue package ARPACK [20]. After more than 300 Arnoldi iterations using the sparse form of the matrix and 60 CPU seconds, the algorithm did not converge. At least in the realm of the options immediately available to us, it is not clear how to take advantage of the fact that we need to compute only one eigenvalue in a way that will be faster than the QR algorithm implemented by `eig`. Another possibility, arising from our numerical results, is that for the problem with $n = 1000$ it may be more efficient to solve the optimization problem for fixed ω by applying KNITRO to the minimization formulation of the eigenvalue problem. This is an intriguing possibility, given the absence in the literature, to our knowledge, of any investigation in the direction of using projected CG, the algorithm implemented by KNITRO, for determining the minimum eigenvalue. We plan to explore it fully in future research.

Arguably, choosing 100 sample points at which to compute the minimal eigenvalue is somewhat artificial. This choice was based on our experience in selecting the number of points needed to represent the eigenvalue graph (Figure 4.3) with an acceptable precision. A perhaps fairer comparison would have been the amount of time needed to produce an approximation of the eigenvalue mapping with the same level of accuracy with our approach and with the direct simulation approach. Of course, once we did that, we would have to specify the interpolation procedure that we use to construct the approximating graph from the sampled values, which is again dependent on various choices beyond the number of samples (such as interpolation used and the choice of the sampling points). We do carry out such a comparison when we compare below the outcome insofar as eigenvectors. On the other hand, from Figure 4.4 one eigenvalue calculation is still far more expensive than the amount of time it took KNITRO to find the solution to the problem $SO(K)$, so our technique would still be quicker by more than a factor of 5 in that case. In any event, we have made available at the publication website of the author, <http://www.mcs.anl.gov/~anitescu/PUBLICATIONS/reports.html>, all the scripts and logs of our simulations, for the reader interested in experimenting with these issues.

In fairness, we do not expect such a striking disparity between competing times to hold in general. Experience with the spectral nonlinear equations formulation does suggest, however, that the effort needed in solving the problem $SO(K)$, a problem that has $M_K + 1$ times larger variable space compared to the original problem (O) for one choice of the parameter ω , is far less than solving the problem O for $M_K + 1$ values of the parameter ω [15, 14]. Because $SO(K)$ has, as shown by Theorem 3.1, the same solution space as its nonlinear equation formulation in the case when the original problem arises from the extremality conditions of a parametric optimization problem, such an inference is warranted. Whether this expectation, which was justified by our results, indeed extends to much larger problems will be the subject of future research.

At a first glance to the left panel of Figure 4.5, our approach did much worse in calculating the behavior of the eigenvector, in effect, the variable of our optimization problem. The figure seems to show errors in the cosine as large as 60%. A deeper investigation revealed that the cusps in the figure have to do with the degeneracy of the eigenvalue problem at those ω values. Indeed, if instead we are evaluating the residual error $\left\| (Q + \omega D_Q) \tilde{x}(\omega) - \tilde{\lambda}(\omega) \tilde{x}(\omega) \right\|$, we see in the right panel of Figure 4.5 that that residual is always below 0.035 and 97% of the time below 0.02 for the Legendre case and is always below 0.025 for the Chebyshev case. By comparison, if one would compute the exact minimum eigenvalue at the points minus 1, 1 and at the coordinates of the three cusps and used a linear interpolation with these nodes and the minimum eigenvectors obtained by simulation (denoted by the “Black box” in Figure 4.5) we see that the error would actually be quite a bit worse, by about a factor of two, and on average by a factor of four. The procedure described is essentially the one of collocation [29] with a piecewise linear basis function. Note that the number of basis functions associated with the interpolation procedure described is 5, the same as the number of basis functions considered by our SFEM approach that has generated the results considered here.

Clearly, our choice of collocation points is probably close to the worst-case one for collocation. On the other hand, one has very little a priori information on how to choose these points, and the distribution chosen is far from being pathological, judging from how well it covers the $[-1, 1]$ interval (for example, in terms of the discrepancy [24]). While such comparisons must be carried out on much larger classes of problems, we find here evidence *that the optimization based SFEM approach may be much more robust than black-box algorithms*, at least for parametric eigenvalue problems. We call a “black-box” algorithm for parametric analysis a non-intrusive algorithm that uses only input-output information of the non-parametric problem (in our example, an eigenvalue solver), in order to generate the parametric approximation. Such algorithms are perhaps the easiest to implement for parametric analysis and uncertainty quantification [11]. Our example shows that such algorithms may encounter difficulties for a small dimension of the parameter space for problems of the type presented here, in addition to the well-documented difficulties for a large dimension of a parameter space [11].

Concerning the validation of the theoretical results, we note that the conditions of Theorem 3.9 are satisfied for the Chebyshev polynomials case, though the Legendre polynomials also seems to provide good approximating properties. The latter is relevant since the Legendre polynomials are the choice in the widespread case of uniform distribution.

5. Conclusions. We have shown that, in the study of the parametric dependence of problems that originate in optimization problems, the spectral finite element

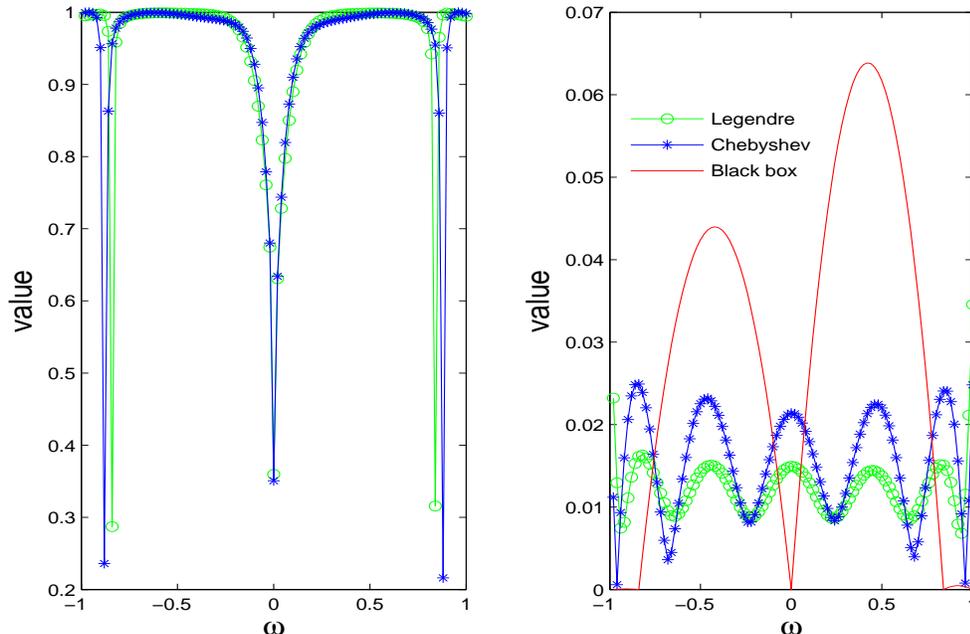


FIG. 4.5. The case $n = 1000$: eigenvector angle and residual.

(SFEM) method can be formulated as an optimization problem. The major advantage of our approach is that the resulting nonlinear problem has a solution that can be found by optimization algorithms.

The formulation will include constraints if the original problem had any, and the spectral finite element approximation to the parametric dependence of the Lagrange multipliers is obtained implicitly from the solution, rather than explicitly as one would expect from the typical spectral finite element formulation. We have shown that, under certain assumptions, the SFEM problem is well-posed and that the sequence of SFEM approximations of increasing degree converges to a solution of the parametric problem. In particular, if the constraints are linear, a solution of the SFEM approach exists without a small variation assumption of the solution $\tilde{x}^*(\omega)$ of the parametric problem (O).

We note that our approach is applicable to the most computationally intensive part of spectral stochastic finite-element methods [15], in the case where parametric nonlinear equations represent the optimality conditions of a parametric optimization problem. That is, our work presents and analyzes a method for computing the parametric solution map.

In the case where our approach is used for studying the parametric dependence of the solution of minimum eigenvalue problems, we have shown that our method can be orders of magnitude faster compared to the simulation-based exploration of the parameter space. In addition, we have evidence that the method may be quite a bit more accurate than worst-case choices of simulation based on black-box exploration of the parameter space. The resulting problem is not convex, and it is difficult to guarantee that the global minimum can be actually found by the software. Nonetheless the software that we used KNITRO showed no difficulty in actually determining the

minimum value.

Several issues remain to be analyzed. These include being able to guarantee that the minimum found is actually a global minimum, determining efficient ways of choosing the polynomial basis functions for a large number of dimensions of the parameter space, efficiently solving the larger coupled optimization problem, showing that the limit of solutions of $(SO(K))$ is sufficiently smooth rather than assuming it in Theorem 3.10, and providing convergence results for inequality-constrained problems and problems without smooth solutions.

Acknowledgments. We are grateful to the editor and the anonymous referee for the suggestions that have substantially improved the quality of the paper, especially insofar the presentation of the assumptions and the discussion of their implications. We are grateful to Michel Deville, Paul Fischer, and Yvon Maday, for discussions on orthogonal polynomials and to Paul Hovland, Pino Palmotti, and Wong-Sik Yang for discussions about stochastic spectral finite elements, uncertainty quantification, and their applications. Mihai Anitescu was supported by the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Advanced Scientific Computing Research, Office of Science, U.S. Dept. of Energy, under Contract DE-AC02-06CH11357.

REFERENCES

- [1] M. ANDERS AND M. HORI, *Three-dimensional stochastic finite element method for elasto-plastic bodies*, International Journal for Numerical Methods In Engineering, 51 (2001), pp. 449–478.
- [2] D. P. BERTSEKAS, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.
- [3] S. BLOWER AND H. DOWLATABADI, *Sensitivity and uncertainty analysis of complex models of disease transmission: an HIV model, as an example*, International Statistical Review, 62(2) (1994), pp. 229–243.
- [4] F. BONNANS AND A. SHAPIRO, *Perturbation Analysis of Optimization Problems*, Springer-Verlag, New York, 2000.
- [5] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.
- [6] C. CANUTO AND A. QUARTERONI, *Approximation results for orthogonal polynomials in Sobolev spaces*, Mathematics of Computation, 38 (157) (1982), pp. 67–86.
- [7] M. K. DEB, I. M. BABUSKA, AND J. T. ODEN, *Solution of stochastic partial differential equations using Galerkin finite element techniques*, Computer Methods in Applied Mechanics and Engineering, 190(48) (2001), pp. 6359–6372.
- [8] A. L. DONTCHEV, *The Graves theorem revisited*, Journal of Convex Analysis, 3(1) (1996), pp. 45–53.
- [9] J. DUDERSTADT AND L. HAMILTON, *Nuclear Reactor Analysis*, John Wiley and Sons, 1976.
- [10] C. F. DUNKL AND Y. XU, *Orthogonal Polynomials of Several Variables*, Cambridge University Press, Cambridge, U.K., 2001.
- [11] M. S. ELDERED, A. A. GIUNTA, B. G. VAN BLOEMEN WAANDERS, S. F. WOJTKIEWICZ, W. E. HART, AND M. P. ALLEVA, *Dakota, a multilevel parallel object-oriented framework for design optimization, parameter estimation, uncertainty quantification, and sensitivity analysis*, SAND Report SAND2001-3796, Sandia National Laboratories, april 2002.
- [12] R. FOURER, D. M. GAY, AND B. W. KERNIGHAN, *AMPL: A Modeling Language for Mathematical Programming*, Brooks/Cole Publishing Company, Pacific Grove, CA, 2002.
- [13] A. GANZBURG, *Multidimensional Jackson theorems*, Siberian Mathematical Journal, 22 (2) (1981), pp. 223–231.
- [14] R. GHANEM AND J. RED-HORSE, *Propagation of uncertainty in complex physical system using a stochastic finite element approach*, Physica D, 133 (1999), pp. 137–144.
- [15] R. GHANEM AND P. SPANOS, *The Stochastic Finite Element Method: A Spectral Approach*, Springer, New York, 1991.
- [16] D. GRIFFEL, *Applied Functional Analysis*, Dover, Mineola, NY, 2002.

- [17] J. HELTON, *Uncertainty and sensitivity analysis techniques for use in performance assessment for radioactive waste disposal*, Reliability Engineering & Systems Safety, 42(2-3) (1993), pp. 327–367.
- [18] I. N. KOVALENKO, A. N. NAKONECHNYI, AND A. B. ROMANOV, *Chebyshev approximation of the distribution function of nonnegative random variables*, Cybernetics and Systems Analysis, 32(2) (1996).
- [19] G. KUCZERA AND E. PARENT, *Monte Carlo assessment of parameter uncertainty in conceptual catchment models: the Metropolis algorithm*, Journal of Hydrology, 211 (1) (1998), pp. 69–85.
- [20] R. LEHOUCQ, D. SORENSEN, AND C. YANG, *ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*, SIAM, Philadelphia, PA, 1997.
- [21] O. L. MANGASARIAN, *Nonlinear Programming*, McGraw-Hill, Englewood Cliffs, NJ, 1969.
- [22] M. G. MORGAN, M. HENRION, AND M. SMALL, *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*, Cambridge University Press, Cambridge, UK, 1992.
- [23] P. NEVAI, *Orthogonal Polynomials*, vol. 18 (213) of Memoirs of the American Mathematical Society, The American Mathematical Society, Providence, 1979.
- [24] H. NIEDERREITER, *Random Number Generation and Quasi-Monte Carlo methods*, Society for Industrial and Applied Mathematics, 1992.
- [25] J. NOCEDAL, R. H. BYRD, AND M. E. HRIBAR, *An interior point algorithm for large scale nonlinear programming*, SIAM J. Optimization, 9(4) (1999), pp. 877–900.
- [26] B. N. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice Hall, Englewood Cliffs, N. J., 1980.
- [27] B. POLYAK, *Newton Kantorovich method and its global convergence*, Journal of Mathematical Sciences, 133(4) (2006), pp. 1513–1523.
- [28] K. M. THOMPSON, D. E. BURMASTER, AND E. A. CROUCH, *Monte Carlo techniques for quantitative uncertainty analysis in public health risk assessments*, Risk Analysis, 12(1) (1992), pp. 53–63.
- [29] D. XIU AND J. S. HESTHAVEN, *High-order collocation methods for differential equations with random inputs*, SIAM Journal on Scientific Computing, 27 (2005), pp. 1118–1139.

The submitted manuscript has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory ("Argonne"). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. The U.S. Government retains for itself, and others acting on its behalf, a paid-up, nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government.