INDUSTRIAL AND SYSTEMS ENGINEERING

ISE

A Preconditioned Semismooth Newton Method For Box-Constrained Elliptic Optimal Control Problems With Random Coefficients

ZHENDONG LI[1], AKWUM ONWUNTA[1], AND HAILING WANG[2]

[1]Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, PA, USA

[2]Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

ISE Technical Report 26T-020

LEHIGH UNIVERSITY

A PRECONDITIONED SEMISMOOTH NEWTON METHOD FOR BOX-CONSTRAINED ELLIPTIC OPTIMAL CONTROL PROBLEMS WITH RANDOM COEFFICIENTS

ZHENDONG LI*, AKWUM ONWUNTA*, AND HAILING WANG†

Abstract. We consider the optimal control problem of an elliptic PDE with random diffusion under pointwise box constraints on the control. The constraints make the optimality conditions nonsmooth, with an active set that is not known in advance, and the random diffusion couples the state and the adjoint across all stochastic modes. To design a solution algorithm for this problem, we rely on the Fenchel duality viewpoint. Fenchel duality absorbs the box constraint into the objective and yields an exact reduced problem in the elliptic adjoint variable alone. After a stochastic Galerkin–finite element discretization, eliminating the state leaves a strongly monotone, piecewise smooth equation, which we solve directly by a matrix-free semismooth Newton method with the box projection retained exactly and with no outer iteration, penalty or smoothing parameter. Next, we introduce and analyze a current-set preconditioner, which adds the active-set correction of the current Newton step to a mean-diffusion reference operator and is inverted exactly by a rank-1 update. For implementation, we use a cheaper preconditioner that keeps only a fraction of free control values. Numerical experiments show second-order convergence of the state, and Newton and PCG counts that stay bounded under mesh refinement and growth of the stochastic dimension.

Key words. stochastic optimal control, Fenchel duality, semismooth Newton method, control constraints, stochastic Galerkin method, preconditioning, uncertainty quantification

MSC codes. 49M15, 49N15, 65K10, 65N30, 65F08, 65C30

1. Introduction.

Optimal control problems governed by partial differential equations (PDEs) are ubiquitous in computational science and engineering, and both a mature theory and a large body of numerical methods are available in the deterministic case [35, 14]. Pointwise control bounds $a \le u \le b$ change the picture. The first-order conditions are no longer a system of equations but a variational inequality, or equivalently a nonsmooth equation built from the metric projection onto the admissible box [20, 15]. Three consequences follow. First, although the reduced objective is strongly convex, the resulting Karush–Kuhn–Tucker (KKT) system is nondifferentiable, so classical Newton methods do not apply. Second, the set of indices at which the bounds are attained—the active set—is not known in advance and must be discovered as part of the solution process. Third, the standard remedies—penalization, interior-point regularization, smoothing—replace the exact constraint by an approximation governed by an additional parameter, and recover it only in the limit.

In the deterministic problem, the primal–dual active set (PDAS) strategy of Bergounioux, Ito and Kunisch [9] treats the bounds exactly by alternating between a prediction of the active set and the solution of an equality-constrained subproblem. Hintermüller, Ito and Kunisch [18] subsequently showed that PDAS is precisely a semismooth Newton (SSN) method for the projection-based reformulation, which explains its local superlinear and mesh-independent convergence [11, 28, 30, 36, 19].

In most applications the PDE coefficients are uncertain and are modelled by random fields and lead to optimal control problems under uncertainty [27]. The random variable may be discretized by sampling, by stochastic collocation [10, 21, 22, 34], or by stochastic Galerkin (SG) projection onto a global orthogonal polynomial

*Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, PA (zhl923@lehigh.edu, ako221@lehigh.edu).

†Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong (hailing.wang@polyu.edu.hk).

basis [16, 39, 23, 26, 4, 24, 32]. We use SG, which yields a single coupled system: the discrete state operator is a sum of Kronecker products, and state and adjoint interact across all stochastic modes.

Combining pointwise constraints with random coefficients is harder than treating either feature alone. For control constraints, the projection onto the box is available in closed form and need not be regularized, yet stochastic Galerkin discretizations of such problems have been studied mainly from the approximation side: Shen et al. [33] derive a priori error estimates for obstacle-type control constraints, and Winkler and Yücel [38] combine error estimates for a Dirichlet boundary control problem with a primal–dual active set solver and block-diagonal preconditioning. There, the control is deterministic and acts on ∂D, so the active set stays decoupled from the stochastic tensor structure. Solvers for stochastic control problems are developed in [8, 6, 1, 25, 3, 2, 7]. None of them treats control constraints.

An alternative is the Fenchel duality viewpoint, where the dual absorbs the constraint into the objective. Clason and Kunisch [12] found the duals of certain unconstrained elliptic control problems structurally more favourable than the primal problems, and Wang et al. [37] derived the Fenchel dual of a linear parabolic problem with box constraints on the control, in which the constraints are embedded in an unconstrained, differentiable dual objective, and solved it by conjugate gradients and by semismooth Newton. They also identified the obstacle that reappears here: semismooth Newton applied to the primal problem treats the box and the PDE constraint simultaneously, so the varying active set changes the structure of the Newton equation and the preconditioner has to be adjusted at every step. In the stochastic setting, this obstacle is sharper: control values are indexed by pairs of a spatial node and a quadrature point, so the active-set term of the Jacobian matrix has no Kronecker structure. A preconditioner containing it exactly must be rebuilt at every step, whereas a mean-based Kronecker preconditioner is cheap but blind to the location of the free set.

In this paper, we develop a dual-based SSN framework for optimal control problems governed by elliptic PDEs with random coefficients and pointwise box constraints on the control. We derive the Fenchel dual in Bochner spaces over the parameter domain, which absorbs the control constraints into an unconstrained dual objective in the adjoint variable, and establish the corresponding optimality system and primal–dual recovery relations. After stochastic Galerkin–finite element discretization, we derive the exact finite-dimensional dual and solve the resulting reduced equation directly by a matrix-free semismooth Newton–PCG method, in which the box projection is retained exactly, and no outer iteration, penalty, or smoothing parameter is required. To handle the active set, which changes at every Newton step, we analyze a current-set preconditioner P_W that places the current active-set correction beside a mean-based reference operator and applies its inverse exactly through the Sherman–Morrison–Woodbury (SMW) identity [17]. For numerical implementation, we use a cheap approximation P_k of the preconditioner P_W, which keeps only the fraction of free control values, and its application needs one sparse Cholesky factor of a spatial matrix, which is reused across Newton steps.

The remainder of the paper is organized as follows. Section 2-3 presents the problem and derives the Fenchel dual and the optimality system. Section 4 gives the SG-FE discretization and its exact dual. Section 5 develops the direct SSN–PCG method and the spectral analysis of the current-set preconditioner P_W. Section 6 specifies the practical preconditioner P_k and reports the numerical experiments. Section 7 draws conclusions and discusses future work.

2. Problem formulation and well-posedness.

2.1. The optimal control problem. Let $D \subset \mathbb{R}^d$, $d \in \{1,2,3\}$, be a bounded Lipschitz domain with Poincaré constant C_P, that is, $\|v\|_{L^2(D)} \le C_P \|\nabla v\|_{L^2(D)^d}$ for all $v \in H_0^1(D)$, let $\xi = (\xi_1, \dots, \xi_K)\colon (\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}^K$, $K < \infty$, be a random vector with image $\Gamma_\xi := \xi(\Omega)$, and let $(\Gamma_\xi, \mathcal{B}, \rho)$ be the induced parameter probability space, where $\mathcal{B}$ is the Borel σ-algebra on Γ_ξ and $\rho := \mathbb{P} \circ \xi^{-1}$ is the law of ξ. Let y denote the state and u the control, and set

$$V = H_0^1(D), \qquad H = L^2(D), \qquad \mathbb{V} = L^2_\rho(\Gamma_\xi; V), \qquad \mathbb{H} = L^2_\rho(\Gamma_\xi; H),$$

where the norms and inner products on $\mathbb{V}$ and $\mathbb{H}$ include the expectation with respect to ρ. We equip V with the norm $\|v\|_V := \|\nabla v\|_{L^2(D)^d}$, so that $\|v\|_{\mathbb{V}}^2 = \int_{\Gamma_\xi} \|\nabla v(\xi)\|_{L^2(D)^d}^2 \, d\rho(\xi)$ and $\|v\|_{\mathbb{H}} \le C_P \|v\|_{\mathbb{V}}$ for $v \in \mathbb{V}$. Given a desired state $y_d \in \mathbb{H}$, a regularization parameter $\gamma > 0$, finite bounds $a < b$, and a random diffusion coefficient κ, the problem studied in this paper is

$$\begin{aligned} \min_{y,\,u} \quad & \frac{1}{2}\int_{\Gamma_\xi}\int_D \big|y(x,\xi) - y_d(x,\xi)\big|^2 dx\, d\rho(\xi) + \frac{\gamma}{2}\int_{\Gamma_\xi}\int_D \big|u(x,\xi)\big|^2 dx\, d\rho(\xi) \\ \text{subject to} \quad & -\nabla\cdot\big(\kappa(x,\xi)\nabla y(x,\xi)\big) = u(x,\xi) \quad \text{in } D\times\Gamma_\xi, \\ & y(x,\xi) = 0 \quad \text{on } \partial D\times\Gamma_\xi, \\ & a \le u(x,\xi) \le b \quad (dx\otimes d\rho)\text{-a.e.} \end{aligned} \tag{2.1}$$

Both integrals are taken with respect to the product of the Lebesgue measure on D and the probability measure ρ; so, the objective in (2.1) is exactly

$$\frac{1}{2}\|y - y_d\|_{\mathbb{H}}^2 + \frac{\gamma}{2}\|u\|_{\mathbb{H}}^2,$$

and we use this compact form from Section 3 onwards.

We note here that the bounds are imposed pointwise in the spatial and the parameter variables, so the admissible set

$$\mathcal{C} = \{u \in \mathbb{H} : a \le u(x,\xi) \le b \text{ for } (dx\otimes d\rho)\text{-a.e. } (x,\xi)\} \tag{2.2}$$

is a subset of the full Bochner space rather than of a deterministic space of controls. Since D and the parameter space carry finite measure, $\mathcal{C}$ is nonempty, bounded, closed, and convex in $\mathbb{H}$.

2.2. The state operator and the reduced problem. Assume that κ is jointly measurable, symmetric, and uniformly elliptic:

$$0 < \kappa_- \le \kappa(x,\xi) \le \kappa_+ < \infty, \tag{2.3}$$

for almost every (x,ξ), with deterministic constants κ_- and κ_+. The weak form of the state equation in (2.1) is governed by the bilinear form

$$\mathfrak{a}(y,v) = \int_{\Gamma_\xi}\int_D \kappa(x,\xi)\nabla y(x,\xi)\cdot\nabla v(x,\xi)\, dx\, d\rho(\xi).$$

THEOREM 2.1 (Lax–Milgram [35]). *Let W be a real Hilbert space and let $\mathfrak{t} : W \times W \to \mathbb{R}$ be a bilinear form for which there are constants $M \ge \alpha > 0$ such that*

$$|\mathfrak{t}(w,z)| \le M\|w\|_W\|z\|_W \quad \text{and} \quad \mathfrak{t}(w,w) \ge \alpha\|w\|_W^2 \qquad (w,z \in W).$$

Then, for every $\ell \in W'$, there is a unique $z \in W$ such that

$$\mathfrak{t}(z, w) = \ell(w) \qquad (w \in W),$$

and $\|z\|_W \leq \alpha^{-1} \|\ell\|_{W'}$.

PROPOSITION 2.2. *Under (2.3), for every $u \in \mathbb{H}$ there is a unique $y = Su \in \mathbb{V}$ such that*

$$\mathfrak{a}(y, v) = (u, v)_{\mathbb{H}} \qquad (v \in \mathbb{V}). \tag{2.4}$$

The solution map

$$S : \mathbb{H} \longrightarrow \mathbb{V} \hookrightarrow \mathbb{H}$$

is bounded, linear, positive, injective, and self-adjoint as an operator on $\mathbb{H}$.

Proof. Boundedness and coercivity of $\mathfrak{a}$ on $\mathbb{V}$ follow from (2.3). Moreover,

$$|(u, v)_{\mathbb{H}}| \leq \|u\|_{\mathbb{H}} \|v\|_{\mathbb{H}} \leq C_P \|u\|_{\mathbb{H}} \|v\|_{\mathbb{V}}.$$

Theorem 2.1 therefore gives a unique solution and

$$\|Su\|_{\mathbb{V}} \leq \frac{C_P}{\kappa_-} \|u\|_{\mathbb{H}}, \qquad \|Su\|_{\mathbb{H}} \leq \frac{C_P^2}{\kappa_-} \|u\|_{\mathbb{H}}.$$

Linearity of S follows from uniqueness. For $u_1, u_2 \in \mathbb{H}$ and $c_1, c_2 \in \mathbb{R}$, the bilinearity of $\mathfrak{a}$ and of $(\cdot, \cdot)_{\mathbb{H}}$ shows that $c_1 Su_1 + c_2 Su_2$ solves (2.4) with right-hand side $c_1 u_1 + c_2 u_2$. By uniqueness, $S(c_1 u_1 + c_2 u_2) = c_1 Su_1 + c_2 Su_2$, that is, S is linear. For $f, g \in \mathbb{H}$, utilizing the symmetry of this form, we obtain from the equation corresponding to Sg (with Sf) and the equation corresponding to Sf (with Sg) that

$$(Sf, g)_{\mathbb{H}} = (g, Sf)_{\mathbb{H}} = \mathfrak{a}(Sg, Sf) = \mathfrak{a}(Sf, Sg) = (f, Sg)_{\mathbb{H}}.$$

Hence $S^* = S$. Moreover, $(Sf, f)_{\mathbb{H}} = \mathfrak{a}(Sf, Sf) \geq 0$, so S is positive. Finally, if $Sf = 0$, then (2.4) gives $(f, v)_{\mathbb{H}} = 0$ for all $v \in \mathbb{V}$. Since $C_c^\infty(D)$ is dense in $L^2(D)$ and H-valued simple functions are dense in $\mathbb{H}$, the space $\mathbb{V}$ is dense in $\mathbb{H}$. Then $f = 0$, and S is injective. ∎

Substituting $y = Su$ eliminates the state from (2.1) and leaves the reduced primal problem

$$\min_{u \in \mathcal{C}} J(u) = \frac{1}{2} \|Su - y_d\|_{\mathbb{H}}^2 + \frac{\gamma}{2} \|u\|_{\mathbb{H}}^2, \tag{2.5}$$

which is the form used throughout the paper.

PROPOSITION 2.3 (Existence and uniqueness). *Problem (2.5) has a unique minimizer $u^* \in \mathcal{C}$. Consequently, with the state equation in (2.1) and the weak form (2.4), $(Su^*, u^*) \in \mathbb{V} \times \mathcal{C}$ is the unique solution of (2.1).*

Proof. The extended functional

$$j(u) = \frac{1}{2} \|Su - y_d\|_{\mathbb{H}}^2 + \frac{\gamma}{2} \|u\|_{\mathbb{H}}^2 + \mathcal{I}_{\mathcal{C}}(u),$$

where $\mathcal{I}_{\mathcal{C}}$ is the indicator function of $\mathcal{C}$ (0 on $\mathcal{C}$ and $+\infty$ elsewhere), and it is proper, coercive, convex, and weakly lower semicontinuous. The direct method gives a minimizer. The term $\gamma \|u\|_{\mathbb{H}}^2 / 2$ is strictly convex, so the minimizer is unique. By Proposition 2.2, the feasible pairs of (2.1) are exactly (Su, u) with $u \in \mathcal{C}$, and the objective of (2.1) at such a pair equals $J(u)$; this gives the second assertion. ∎

3. Fenchel dual and the optimality system. The reduced problem (2.5) is a constrained minimization over $\mathcal{C}$. This section removes the constraint by dualization. We first compute the two convex conjugates, then invoke Fenchel–Rockafellar duality to obtain a dual problem in a residual variable, and finally change variables to the elliptic adjoint. The last form is the one the algorithm of Section 5 actually solves, and the recovery relations obtained at the end of the section are what return the state and the control once the adjoint is known.

3.1. The dual problem. To apply Fenchel–Rockafellar duality, we define $F : \mathbb{H} \to (-\infty, \infty]$ and $G : \mathbb{H} \to \mathbb{R}$ by

$$F(u) = \frac{\gamma}{2} \|u\|_{\mathbb{H}}^2 + \mathcal{I}_{\mathcal{C}}(u), \qquad G(z) = \frac{1}{2} \|z - y_d\|_{\mathbb{H}}^2 .$$

Problem (2.5) can then be written as the unconstrained problem

$$\inf_{u \in \mathbb{H}} \{F(u) + G(Su)\}.$$

The box constraint has been absorbed into F through the indicator $\mathcal{I}_{\mathcal{C}}$, so that both terms are now extended-real-valued convex functions on $\mathbb{H}$ and no explicit feasible set remains. The next lemma shows that conjugation turns the indicator into a finite, differentiable function: this is the precise sense in which duality trades a constraint for smoothness.

LEMMA 3.1 (Conjugates). *For $q, v \in \mathbb{H}$,*

$$G^*(q) = \frac{1}{2} \|q\|_{\mathbb{H}}^2 + (q, y_d)_{\mathbb{H}}, \qquad F^*(v) = \int_{\Gamma_\xi} \int_D \psi_\gamma(v(x, \xi))\, dx\, d\rho,$$

where

$$\psi_\gamma(s) = \begin{cases} as - \dfrac{\gamma}{2} a^2, & s < \gamma a, \\ \dfrac{s^2}{2\gamma}, & \gamma a \le s \le \gamma b, \\ bs - \dfrac{\gamma}{2} b^2, & s > \gamma b. \end{cases} \tag{3.1}$$

Moreover,

$$\psi_\gamma'(s) = \mathrm{Proj}_{[a,b]}(s/\gamma). \tag{3.2}$$

Proof. For G^*, the substitution $w = z - y_d$ gives

$$\sup_z \{(q, z)_{\mathbb{H}} - \tfrac{1}{2} \|z - y_d\|_{\mathbb{H}}^2\} = (q, y_d)_{\mathbb{H}} + \sup_w \{(q, w)_{\mathbb{H}} - \tfrac{1}{2} \|w\|_{\mathbb{H}}^2\}.$$

For F^*, maximize the scalar function

$$sr - \frac{\gamma}{2} r^2$$

over $r \in [a, b]$. Its maximizer is $\mathrm{Proj}_{[a,b]}(s/\gamma)$, which gives (3.1). Pointwise integration is justified by normal-integrand conjugacy; it also follows directly because the measurable pointwise maximizer belongs to $\mathcal{C}$. ∎

Note that ψ_γ is continuously differentiable with a piecewise linear derivative (3.2): the nonsmoothness of the primal constraint has been reduced to a single projection, which is exactly the structure the semismooth Newton method of Section 5 exploits.

Strong duality needs a qualification condition. We use the following classical form, in which the qualification is the finiteness and continuity of G at a single point of the image of $\operatorname{dom} F$; see [13, 31, 5] for proofs and for more general statements.

THEOREM 3.2 (Fenchel–Rockafellar [13, Chap. III], [5, Chap. 15]). *Let X, Y be Banach spaces, let $T : X \to Y$ be bounded and linear, and let $F : X \to (-\infty, \infty]$, $G : Y \to (-\infty, \infty]$ be proper and convex. If there exists $x_0 \in \operatorname{dom} F$ such that G is finite and continuous at Tx_0, then*

$$\inf_{x \in X} \{F(x) + G(Tx)\} = \sup_{q \in Y'} \{-F^*(-T^*q) - G^*(q)\}.$$

When the primal value is finite, the supremum is attained.

The tracking term G is a finite quadratic, hence continuous on all of $\mathbb{H}$, and $\operatorname{dom} F = \mathcal{C}$ is nonempty because the constant control $(a + b)/2$ is admissible. No interior-point or Slater-type condition on the box is needed.

THEOREM 3.3 (Residual-variable Fenchel dual). *Under (2.3), the Fenchel dual of (2.5) is*

$$\max_{q \in \mathbb{H}} \left\{ - \int_{\Gamma_\xi \times D} \psi_\gamma(-Sq)\, d\rho\, dx - \frac{1}{2} \|q\|_{\mathbb{H}}^2 - (q, y_d)_{\mathbb{H}} \right\}. \tag{3.3}$$

There is no duality gap, and the maximum is attained at a unique $q^* \in \mathbb{H}$.

Proof. Apply Theorem 3.2 with

$$X = Y = \mathbb{H}, \qquad T = S,$$

identifying $\mathbb{H}'$ with $\mathbb{H}$ through the Riesz map, so that $T^* = S^*$ is the Hilbert-space adjoint. The quadratic G is finite and continuous everywhere, and the constant midpoint control $(a + b)/2$ belongs to $\operatorname{dom} F = \mathcal{C}$. Thus the qualification holds. Since $S^* = S$, Lemma 3.1 gives (3.3). The map $q \mapsto F^*(-Sq)$ is finite and convex, and $-\frac{1}{2}\|q\|_{\mathbb{H}}^2 - (q, y_d)_{\mathbb{H}}$ is 1-strongly concave. Hence the dual objective is 1-strongly concave and its maximizer is unique. ∎

3.2. The elliptic adjoint form and primal–dual recovery. Problem (3.3) is already unconstrained, but it is not yet a convenient computational form: the solution operator S sits inside the nonlinearity $\psi_\gamma(-Sq)$, so every evaluation of the objective requires an elliptic solve. Substituting $p = -Sq$ moves S out of the nonlinear term and replaces it by the differential operator A in the quadratic term, which is what makes the discretized problem sparse and the Jacobian applications of Section 5 matrix-free. For this substitution to be a genuine reparametrization—creating no spurious maximizers, losing none of the existing ones, and not tacitly shrinking the feasible set—we must know that $p \mapsto -Ap$ maps $D(A)$ bijectively onto $\mathbb{H}$. Establishing that is the sole purpose of the next lemma, and it is the only place where the graph domain (3.4) is needed.

Let A denote the elliptic operator associated with $\mathfrak{a}$. As an unbounded operator on $\mathbb{H}$, its natural graph domain is

$$D(A) = \{p \in \mathbb{V} : Ap \in \mathbb{H}\}, \tag{3.4}$$

where $Ap \in \mathbb{H}$ means that there is an element of $\mathbb{H}$, denoted Ap, with

$$\mathfrak{a}(p, v) = (Ap, v)_{\mathbb{H}} \qquad (v \in \mathbb{V}).$$

This element is unique since $\mathbb{V}$ is dense in $\mathbb{H}$ (see the proof of Proposition 2.2).

LEMMA 3.4. *The map $A : D(A) \to \mathbb{H}$ is a bijection with inverse S. Consequently $p \mapsto q = -Ap$ is a bijection from $D(A)$ onto $\mathbb{H}$.*

Proof. For every $f \in \mathbb{H}$, Proposition 2.2 gives a unique $p = Sf \in \mathbb{V}$ satisfying $Ap = f$. Hence $p \in D(A)$, so A is surjective and S is its right inverse. If $Ap = 0$, coercivity of $\mathfrak{a}$ gives $p = 0$, so A is injective. No $H^2(D)$ regularity is used or asserted. ∎

THEOREM 3.5 (Elliptic-adjoint PDE dual). *The residual dual (3.3) is equivalent, up to the constant $-\|y_d\|_{\mathbb{H}}^2/2$, to*

$$\min_{p \in D(A)} \left\{ \frac{1}{2} \|Ap - y_d\|_{\mathbb{H}}^2 + \int_{\Gamma_\xi \times D} \psi_\gamma(p)\, d\rho\, dx \right\}. \tag{3.5}$$

Proof. Let

$$p = -Sq.$$

Lemma 3.4 gives $q = -Ap$, while $-Sq = p$. Substitution into (3.3) gives

$$\max_{p \in D(A)} \left\{ -\int \psi_\gamma(p) - \frac{1}{2} \|Ap\|_{\mathbb{H}}^2 + (Ap, y_d)_{\mathbb{H}} \right\}.$$

Changing the sign of the objective and completing the square yields (3.5). ∎

Both duals determine the adjoint, but the quantities of interest—and the residuals used to monitor convergence—are the primal pair (y, u). The final result of this section therefore records how the primal variables are recovered from a dual solution and verifies that the recovered triple solves the classical optimality system, so that nothing is lost in passing to the dual. The projection formula it produces is not merely a postprocessing step: it is the term that reappears in the discrete equation of Section 5 and that the semismooth Newton method differentiates.

THEOREM 3.6 (Optimality system). *Let u^*, y^*, q^*, p^* be primal and dual optimizers, with $y^* = Su^*$ and $p^* = -Sq^*$. Then*

$$q^* = y^* - y_d, \qquad u^* = \mathrm{Proj}_{[a,b]}(p^*/\gamma),$$

and the primal and dual problems are equivalent to

$$\begin{cases} -\nabla \cdot (\kappa \nabla y) = u, & y|_{\partial D} = 0, \\ -\nabla \cdot (\kappa \nabla p) = y_d - y, & p|_{\partial D} = 0, \\ u = \mathrm{Proj}_{[a,b]}(p/\gamma). \end{cases} \tag{3.6}$$

Here p is the negative of the adjoint state in the usual Lagrangian convention and the recovery formula followed as $u = \mathrm{Proj}_{[a,b]}(p/\gamma)$.

Proof. Fenchel equality is equivalent to

$$q^* \in \partial G(Su^*), \qquad -S^*q^* \in \partial F(u^*).$$

The first inclusion gives

$$q^* = Su^* - y_d = y^* - y_d.$$

Since $S^* = S$ and $p^* = -Sq^*$, the second inclusion is

$$p^* \in \gamma u^* + N_{\mathcal{C}}(u^*),$$

where $N_{\mathcal{C}}(u) = \{v \in \mathbb{H} : (v, w-u)_{\mathbb{H}} \le 0 \text{ for all } w \in \mathcal{C}\}$ is the normal cone of $\mathcal{C}$ at $u \in \mathcal{C}$, which is equivalent to

$$u^* = \mathrm{Proj}_{\mathcal{C}}(p^*/\gamma) = \mathrm{Proj}_{[a,b]}(p^*/\gamma) \quad \text{a.e.}$$

Finally,

$$Ap^* = -q^* = y_d - y^*.$$

Together with primal feasibility, these identities give (3.6). ∎

4. Stochastic Galerkin discretization and its dual.

4.1. SG–FE state and mass matrices. Let $\mathcal{P}_P \subset L^2_\rho(\Gamma_\xi)$ be the space of polynomials in ξ of total degree at most P, and let $\{\psi_i\}_{i=1}^{n_P}$, $n_P = \binom{K+P}{P}$, be a basis of $\mathcal{P}_P$ that is orthonormal in $L^2_\rho(\Gamma_\xi)$, i.e., $\mathbb{E}[\psi_i\psi_j] = \delta_{ij}$, which is the generalized polynomial chaos (gPC) basis [40]. Let $\{\phi_r\}_{r=1}^{n_h}$ be the nodal basis of a conforming finite-element space $V_h \subset H^1_0(D)$. The SG method approximates the state in the tensor-product space $\mathcal{P}_P \otimes V_h$,

$$y_{hP}(x,\xi) = \sum_{i=1}^{n_P}\sum_{r=1}^{n_h} y_{(i,r)}\,\psi_i(\xi)\phi_r(x),$$

and requires the weak state equation (2.4) to hold for all test functions $v = \psi_j\phi_s$, $j = 1,\dots,n_P$, $s = 1,\dots,n_h$. The coefficient vector $\mathbf{y} = (y_{(i,r)}) \in \mathbb{R}^N$, $N \coloneqq n_P n_h$, is ordered stochastic-major, i.e., the index (i,r) is stored at position $(i-1)n_h + r$. The adjoint $\mathbf{p}$ is expanded in the same basis.

Unlike the state, the control is not expanded in the gPC basis but represented by its values $\mathbf{u} = (u_{(k,r)}) \in \mathbb{R}^m$, $m \coloneqq n_q n_h$, at the points (x_r, ξ_k). This keeps the box constraint $a \le u \le b$ pointwise and componentwise. The SG load $\mathbb{E}[\int_D u\,\psi_j\phi_s\,dx]$ is approximated by the quadrature rule in ξ and the lumped mass in x, which gives $(E^\top W\mathbf{u})_{(j,s)} = \sum_k w_k\psi_j(\xi_k)(M_L)_{ss}u_{(k,s)}$.

Suppose that the diffusion coefficient κ used in the SG form admits the separable representation

$$\kappa(x,\xi) = \sum_{\ell=0}^{N_K-1} \kappa_\ell(x)\zeta_\ell(\xi), \qquad \zeta_0 \equiv 1,$$

where $\kappa_\ell \in L^\infty(D)$ are deterministic spatial functions, $\zeta_\ell\colon \Gamma_\xi \to \mathbb{R}$ are given scalar functions of the parameter, and N_K is the number of terms. For the affine coefficients used in Section 6, $N_K = K+1$ and $\zeta_\ell(\xi) = \xi_\ell$ for $\ell \ge 1$. Followed by [8], with the definition of

$$\begin{aligned}(G_\ell)_{ij} &= \mathbb{E}[\zeta_\ell\psi_i\psi_j],\\ (M_D)_{ij} &= \int_D \phi_i\phi_j\,dx,\\ (K_\ell)_{ij} &= \int_D \kappa_\ell(x)\nabla\phi_i\cdot\nabla\phi_j\,dx,\end{aligned}$$

the SG mass matrix and stiffness matrix are

$$\mathcal{M} = G_0 \otimes M_D, \qquad \mathcal{K} = \sum_{\ell=0}^{N_K-1} G_\ell \otimes K_\ell. \tag{4.1}$$

For an exactly integrated orthonormal basis, $G_0 = I$. The control is represented by its values at the points (x_r, ξ_k), where $x_1, \dots, x_{n_h}$ are the interior nodes of V_h and $\{\xi_k, w_k\}_{k=1}^{n_q}$ is a positive quadrature rule on Γ_ξ. Let $V_\xi \in \mathbb{R}^{n_q \times n_P}$ with $(V_\xi)_{kj} = \psi_j(\xi_k)$, $W_\xi = \operatorname{diag}(w_1, \dots, w_{n_q})$, and $M_L = \operatorname{diag}(M_D \mathbf{1})$, the row-sum lumped spatial mass. Define

$$E = V_\xi \otimes I_{n_h} \in \mathbb{R}^{n_q n_h \times n_P n_h}, \qquad W = W_\xi \otimes M_L \in \mathbb{R}^{n_q n_h \times n_q n_h}.$$

For a coefficient vector $\mathbf{p}$, $(E\mathbf{p})_{(k,r)} = \sum_j \psi_j(\xi_k) p_{(j,r)}$ is the value at (x_r, ξ_k) of the SG function $\sum_{j,r} p_{(j,r)} \psi_j(\xi) \phi_r(x)$, because the nodal basis satisfies $\phi_s(x_r) = \delta_{rs}$. The entries of W are positive, so W is positive definite and diagonal. Boundary nodes carry no control values, since the Dirichlet degrees of freedom are eliminated. Thus E and W belong to the control discretization, whereas M_D, K_ℓ, $\mathcal{M}$, and $\mathcal{K}$ belong to the state discretization.

We assume that $\mathcal{M}$ and $\mathcal{K}$ are symmetric positive definite; this holds, for instance, when the matrices G_ℓ are computed exactly and κ satisfies (2.3).

4.2. Lagrange dual and KKT system. Let $\mathbf{u} \in \mathbb{R}^m$ be independent control values at the positive quadrature points. Observe that the discrete state $\mathbf{y} \in \mathbb{R}^N$ satisfies the SG weak equation

$$\mathcal{K}\mathbf{y} = E^\top W \mathbf{u}.$$

The exact finite-dimensional constrained primal is

$$\begin{aligned} \min_{\mathbf{y}\in\mathbb{R}^N,\ \mathbf{u}\in\mathbb{R}^m} \quad & \frac{1}{2}(\mathbf{y}-\mathbf{y}_d)^\top \mathcal{M}(\mathbf{y}-\mathbf{y}_d) + \frac{\gamma}{2}\mathbf{u}^\top W \mathbf{u}, \\ \text{subject to} \quad & \mathcal{K}\mathbf{y} - E^\top W \mathbf{u} = 0, \\ & a \le \mathbf{u} \le b. \end{aligned} \tag{4.2}$$

Theorem 4.1 below refers to the dual of (4.2).

THEOREM 4.1. *Problem (4.2) has the concave dual*

$$\max_{\mathbf{p}\in\mathbb{R}^N} \left\{ -\frac{1}{2}\mathbf{p}^\top \mathcal{K}\mathcal{M}^{-1}\mathcal{K}^\top \mathbf{p} + \mathbf{p}^\top \mathcal{K}\mathbf{y}_d - \sum_{\ell=1}^{m} w_\ell \psi_\gamma((E\mathbf{p})_\ell) \right\}. \tag{4.3}$$

There is no duality gap. Moreover, the primal and dual optimizers are unique.

Proof. Let

$$\bar{\mathbf{u}} = \frac{a+b}{2}\mathbf{1}, \qquad \bar{\mathbf{y}} = \mathcal{K}^{-1}E^\top W \bar{\mathbf{u}}.$$

This is an affine-feasible point in the relative interior of the box, so the finite-dimensional relative-interior qualification holds.

Now, consider the Lagrangian

$$\mathcal{L}_h(\mathbf{y}, \mathbf{u}, \mathbf{p}) = \frac{1}{2}(\mathbf{y}-\mathbf{y}_d)^\top \mathcal{M}(\mathbf{y}-\mathbf{y}_d) + \frac{\gamma}{2}\mathbf{u}^\top W \mathbf{u} + \mathbf{p}^\top(\mathcal{K}\mathbf{y} - E^\top W \mathbf{u}) + \mathcal{I}_{[a,b]^m}(\mathbf{u}). \tag{4.4}$$

Minimization in $\mathbf{y}$ of (4.4) gives

$$\mathbf{y} = \mathbf{y}_d - \mathcal{M}^{-1}\mathcal{K}^\top \mathbf{p}$$

and the minimum objective

$$-\frac{1}{2}\mathbf{p}^\top \mathcal{K}\mathcal{M}^{-1}\mathcal{K}^\top \mathbf{p} + \mathbf{p}^\top \mathcal{K}\mathbf{y}_d.$$

For $s_\ell = (E\mathbf{p})_\ell$,

$$\inf_{r\in[a,b]} w_\ell \left(\frac{\gamma}{2}r^2 - s_\ell r\right) = -w_\ell \psi_\gamma(s_\ell).$$

Since $W = \mathrm{diag}(w_1, \ldots, w_m)$ is diagonal and the box is a product of intervals, the $\mathbf{u}$-part of $\mathcal{L}_h$ separates into the m scalar problems above. Summing them over $\ell = 1, \ldots, m$ and adding the minimum in $\mathbf{y}$ gives (4.3). Since $\mathcal{K}$ is invertible, $\mathcal{K}\mathcal{M}^{-1}\mathcal{K}^\top \succ 0$, so the dual objective is strongly concave and tends to $-\infty$ as $\|\mathbf{p}\| \to \infty$. Hence it has a unique maximizer. The primal objective is strictly convex on the affine feasible set, giving primal uniqueness. ∎

Corollary 4.2. *After eliminating* $\mathbf{u}$*, the KKT conditions are*

$$\mathcal{M}\mathbf{y} + \mathcal{K}^\top \mathbf{p} = \mathcal{M}\mathbf{y}_d, \tag{4.5a}$$

$$\mathcal{K}\mathbf{y} = R(\mathbf{p}), \tag{4.5b}$$

where $R(\mathbf{p}) = E^\top W \,\mathrm{Proj}_{[a,b]}(E\mathbf{p}/\gamma)$.

Proof. Stationarity in $\mathbf{y}$, $\nabla_{\mathbf{y}}\mathcal{L}_h = \mathcal{M}(\mathbf{y}-\mathbf{y}_d)+\mathcal{K}^\top \mathbf{p} = 0$, gives (4.5a). Stationarity in each control value is

$$0 \in \gamma\mathbf{u} - E\mathbf{p} + N_{[a,b]^m}(\mathbf{u}),$$

where $N_{[a,b]^m}$ is the normal cone of the box $[a,b]^m \subset \mathbb{R}^m$, because positive diagonal rescaling leaves the componentwise box normal cone unchanged. Therefore

$$\mathbf{u} = \mathrm{Proj}_{[a,b]}(E\mathbf{p}/\gamma).$$

Substitution into the primal feasibility condition $\mathcal{K}\mathbf{y}-E^\top W\mathbf{u} = 0$ gives (4.5b). Equivalently, differentiating (4.3) and using (3.2) gives

$$-\mathcal{K}\mathcal{M}^{-1}\mathcal{K}^\top \mathbf{p} + \mathcal{K}\mathbf{y}_d - R(\mathbf{p}) = 0,$$

which is the same equation after substituting $\mathbf{y} = \mathbf{y}_d - \mathcal{M}^{-1}\mathcal{K}^\top \mathbf{p}$. ∎

Lemma 4.3. *If* $W \succ 0$*, then* $R : \mathbb{R}^N \to \mathbb{R}^N$ *in* (4.5) *is monotone and Lipschitz continuous.*

Proof. For any $\mathbf{v}, \mathbf{z} \in \mathbb{R}^N$,

$$(\mathbf{v}-\mathbf{z})^\top (R(\mathbf{v}) - R(\mathbf{z})) = (E\mathbf{v} - E\mathbf{z})^\top W \left[\mathrm{Proj}_{[a,b]}(E\mathbf{v}/\gamma) - \mathrm{Proj}_{[a,b]}(E\mathbf{z}/\gamma)\right] \geq 0.$$

Each summand is nonnegative because the scalar projection is monotone and each quadrature weight is positive. The scalar projection is 1-Lipschitz, so R is Lipschitz. ∎

5. A semismooth Newton–PCG method. Eliminating $\mathbf{y}$ in (4.5) gives

$$F(\mathbf{p}) = \mathcal{K}\mathcal{M}^{-1}\mathcal{K}^\top \mathbf{p} - \mathcal{K}\mathbf{y}_d + E^\top W \operatorname{Proj}_{[a,b]}(E\mathbf{p}/\gamma) = 0. \tag{5.1}$$

For the diagonal free-set mask $D(\mathbf{p}) = \operatorname{diag}(d_q(\mathbf{p})) \in \mathbb{R}^{m\times m}$, where $d_q(\mathbf{p}) = 1$ on a free component and $d_q(\mathbf{p}) = 0$ on a saturated component (and zero is selected at a breakpoint), the Newton derivative of F in the sense of [18, 36] is

$$J_D = \mathcal{K}\mathcal{M}^{-1}\mathcal{K}^\top + \gamma^{-1}E^\top W D(\mathbf{p})E. \tag{5.2}$$

For a vector $\mathbf{v}$, $J_D\mathbf{v}$ is evaluated matrix-free using one application each of $\mathcal{K}$, $\mathcal{K}^\top$, $\mathcal{M}^{-1}$, E, and $E^\top$. Each Newton system is solved by PCG to the forcing tolerance

$$\eta_k = \min\{\eta_{\max}, \max(10^{-12}, 0.1\sqrt{r_k})\}, \qquad r_k = \|F(\mathbf{p}_k)\| \,/\, \max\{1, \|\mathcal{K}\mathbf{y}_d\|\}, \tag{5.3}$$

and the step is globalized by Armijo backtracking with parameters 10^{-4} and $1/2$, which preserves descent of the convex dual merit function. The iteration stops when the normalized KKT residual is at most 10^{-6} or after $k_{\max}$ Newton steps, and the values of $\eta_{\max}$ and $k_{\max}$ are given in Section 6. The complete procedure is specified by Algorithm 6.1 in Section 6, after the preconditioner has been specified.

5.1. Current-set preconditioner and spectral bounds. Here M_D, K_0, and K_ℓ retain the spatial finite-element meanings defined in Section 4. The calligraphic symbols $\mathcal{M}$, $\mathcal{K}$, and $\overline{\mathcal{K}}$ denote the full stochastic–Galerkin matrices. We use stochastic-major ordering first and apply an orthogonal permutation to node-major ordering for the local assembly estimates. In stochastic-major ordering, since $G_0 = I$ for the exactly integrated orthonormal basis of Section 4, the matrices in (4.1), with the mean term $\ell = 0$ written separately, become

$$\mathcal{M} = I_{n_P} \otimes M_D, \qquad \mathcal{K} = I_{n_P} \otimes K_0 + \sum_{\ell=1}^{N_K-1} G_\ell \otimes K_\ell, \qquad \overline{\mathcal{K}} = I_{n_P} \otimes K_0. \tag{5.4}$$

Here $\overline{\mathcal{K}}$ is the part of $\mathcal{K}$ generated by the mean coefficient κ_0, $N_K - 1$ is the number of non-mean terms of κ, and M_D, K_0 are symmetric positive definite, G_ℓ, K_ℓ are symmetric, and $\mathcal{K}$ is positive definite. The stochastic basis is orthonormal for the mass inner product represented in $\mathcal{M}$. All coefficient amplitudes and normalization factors are already included in G_ℓ or K_ℓ.

The control weights W are positive and diagonal, as above. Written in terms of the reduced equation (5.1), the Jacobian matrix (5.2) associated with a selected projection derivative D has the form

$$H = \mathcal{K}\mathcal{M}^{-1}\mathcal{K}^\top, \quad H_0 = \overline{\mathcal{K}}\mathcal{M}^{-1}\overline{\mathcal{K}}^\top, \qquad J = H + \gamma^{-1}C_D, \quad P_W = H_0 + \gamma^{-1}C_D, \tag{5.5}$$

where $\gamma > 0$ and $C_D = E^\top W D E \succeq 0$. The algorithm uses the binary mask $D = D(\mathbf{p})$ of (5.2) and the spectral results below hold for any positive semidefinite C_D shared by J and P_W.

Let $\|\mathbf{v}\|_Q = (\mathbf{v}^\top Q\mathbf{v})^{1/2}$ for an SPD matrix Q. All unweighted matrix norms below are Euclidean operator norms. The notation $\operatorname{cond}_{\mathrm{sp}}(\mathrm{P_W^{-1}J})$ means the ratio of extreme eigenvalues, equivalently the condition number of the SPD matrix $P_W^{-1/2} J P_W^{-1/2}$.

LEMMA 5.1. *Let*

$$\mathcal{T} = \mathcal{M}^{-1/2}\mathcal{K}\overline{\mathcal{K}}^{-1}\mathcal{M}^{1/2}, \qquad \sigma_- = \sigma_{\min}(\mathcal{T}) > 0, \quad \sigma_+ = \sigma_{\max}(\mathcal{T}),$$

where $\sigma_{\min}(\mathcal{T})$ and $\sigma_{\max}(\mathcal{T})$ denote the smallest and largest singular values of $\mathcal{T}$, and the extreme generalized eigenvalues of (H, H_0) are exactly σ_-^2, σ_+^2. In particular,

$$\sigma_-^2 H_0 \preceq H \preceq \sigma_+^2 H_0. \tag{5.6}$$

Proof. Let $L_0 = \mathcal{M}^{-1/2}\overline{\mathcal{K}}$, an invertible matrix, since $\mathcal{M} \succ 0$ and $\overline{\mathcal{K}} = I_{n_P} \otimes K_0$ is invertible. With the fact that $\mathcal{K} = \mathcal{K}^\top$ and $\overline{\mathcal{K}} = \overline{\mathcal{K}}^\top$ from (5.4), direct multiplication gives $H_0 = L_0^\top L_0$ and $H = L_0^\top \mathcal{T}^\top \mathcal{T} L_0$, indeed $\mathcal{T} L_0 = \mathcal{M}^{-1/2}\mathcal{K}$. Therefore

$$\frac{\mathbf{v}^\top H \mathbf{v}}{\mathbf{v}^\top H_0 \mathbf{v}} = \frac{\|\mathcal{T}\mathbf{z}\|_2^2}{\|\mathbf{z}\|_2^2}, \qquad \mathbf{z} = L_0 \mathbf{v}.$$

Since L_0 is onto, extrema over $\mathbf{v} \neq 0$ equal extrema over $\mathbf{z} \neq 0$. They are the extreme eigenvalues of $\mathcal{T}^\top \mathcal{T}$, namely σ_-^2 and σ_+^2. Note that $\sigma_- > 0$ because $\mathcal{T}$ is invertible. ∎

THEOREM 5.2. *If $cH_0 \preceq H \preceq CH_0$ for constants $0 < c \leq C$, then for every $\gamma > 0$ and $C_D \succeq 0$,*

$$\operatorname{spec}(P_W^{-1} J) \subset [\theta_-, \theta_+], \qquad \theta_- = \min\{c, 1\}, \quad \theta_+ = \max\{C, 1\}. \tag{5.7}$$

Consequently $\operatorname{cond}_{\mathrm{sp}}(\mathrm{P_W^{-1}J}) \leq \theta_+/\theta_-$. *In particular, one can always take $c = \sigma_-^2, C = \sigma_+^2$ at a fixed discretization.*

Proof. First, observe that both

$$J - \theta_- P_W = (H - \theta_- H_0) + (1 - \theta_-)\gamma^{-1} C_D$$

and

$$\theta_+ P_W - J = (\theta_+ H_0 - H) + (\theta_+ - 1)\gamma^{-1} C_D$$

are positive semidefinite: in the first terms because $\theta_- \leq c$ and $\theta_+ \geq C$, and in the second terms because $\theta_- \leq 1 \leq \theta_+$ and $C_D \succeq 0$. Since $P_W \succ 0$, congruence with $P_W^{-1/2}$ preserves semidefiniteness and gives $\theta_- I \preceq P_W^{-1/2} J P_W^{-1/2} \preceq \theta_+ I$. Hence the eigenvalues of the symmetric matrix $P_W^{-1/2} J P_W^{-1/2}$ lie in $[\theta_-, \theta_+]$, and so do those of $P_W^{-1} J = P_W^{-1/2}(P_W^{-1/2} J P_W^{-1/2}) P_W^{1/2}$. ∎

5.2. Uniform bounds from local assembly. We next provide sufficient local structure that yields a uniform bound. Consider a square of fixed side length $L_D > 0$ with $n_e \geq 2$ equal elements per direction, $h = L_D/n_e$, and conforming bilinear elements. All boundary nodal values are zero and are eliminated. On each element T, let q index the four tensor Gauss points x_{Tq}, let $w_{Tq} > 0$ include the physical Jacobian, let $\mathbf{g}_{Tiq} \in \mathbb{R}^2$ be the gradient of the local nodal basis function i at x_{Tq}, and let $a_{\ell,Tq} = \kappa_\ell(x_{Tq})$.

ASSUMPTION 5.3. *Let the scalar reference weights and symmetric coefficient blocks satisfy*

$$0 < \nu_- \leq a_{0,Tq} \leq \nu_+, \qquad \alpha I_{n_P} \preceq A_{Tq} \preceq \beta I_{n_P}, \qquad 0 < \alpha \leq \beta. \tag{5.8}$$

The reference stiffness has entries

$$(K_0)_{ij} = \sum_{T \ni i,j} \sum_q w_{Tq} a_{0,Tq}\, \mathbf{g}_{Tiq}^\top \mathbf{g}_{Tjq}.$$

After an orthogonal permutation into node-major ordering, the full stiffness has $n_P \times n_P$ blocks

$$K_{ij} = \sum_{T \ni i,j} \sum_{q} w_{Tq} a_{0,Tq} (\mathbf{g}_{Tiq}^\top \mathbf{g}_{Tjq}) A_{Tq}. \tag{5.9}$$

The spatial mass is the consistent bilinear mass, integrated exactly by the same Gauss rule.

Orthogonal reordering does not change any spectral or norm assertion. For a Kronecker assembly with scalar coefficient arrays $a_{\ell,Tq}$, (5.9) is exactly (5.4) when

$$A_{Tq} = I_{n_P} + \sum_{\ell=1}^{N_K-1} \frac{a_{\ell,Tq}}{a_{0,Tq}} G_\ell. \tag{5.10}$$

LEMMA 5.4. *Under Assumption 5.3,*

$$\frac{h^2}{9} I \preceq M_D \preceq h^2 I, \qquad (K_0)_{ii} \le \frac{8\nu_+}{3}, \qquad K_0 \succeq \lambda_* M_D, \quad \lambda_* := \frac{8\nu_-}{L_D^2}. \tag{5.11}$$

Moreover $\alpha\overline{\mathcal{K}} \preceq \mathcal{K} \preceq \beta\overline{\mathcal{K}}$.

Proof. Let one-dimensional interior matrices be

$$M_1 = \frac{h}{6} \operatorname{tridiag}(1, 4, 1), \qquad S_1 = \frac{1}{h} \operatorname{tridiag}(-1, 2, -1).$$

Inserting the vector with entries $\sin(ir\pi/n_e)$ into the three-point stencils shows that the eigenvalues are

$$m_r = \frac{h}{3}(2 + \cos\theta_r), \qquad s_r = \frac{2}{h}(1 - \cos\theta_r), \qquad \theta_r = \frac{r\pi}{n_e}.$$

These orthogonal sine vectors form a basis. Since $h/3 \le m_r \le h$, the tensor identity $M_D = M_1 \otimes M_1$ gives the mass bounds. The unit-reference stiffness is $K_{\text{unit}} = M_1 \otimes S_1 + S_1 \otimes M_1$. Its generalized eigenvalues relative to M_D are $s_r/m_r + s_t/m_t$. The ratio is increasing with θ_r and

$$\frac{s_1}{m_1} = \frac{12}{h^2} \frac{\sin^2(\pi/(2n_e))}{2 + \cos(\pi/n_e)} \ge \frac{4}{h^2 n_e^2} = \frac{4}{L_D^2}.$$

Here $\sin x \ge 2x/\pi$ for $0 \le x \le \pi/2$ follows from concavity of sine. Positive quadrature gives $K_0 \succeq \nu_- K_{\text{unit}}$, proving the generalized lower bound.

The diagonal of the unit local 4×4 bilinear stiffness is $2/3$. Indeed, for $N = (1-r)(1-s)/4$ on $[-1,1]^2$, the physical gradient and Jacobian give the integral $\frac{1}{16}\int_{-1}^{1}\int_{-1}^{1}((1-s)^2 + (1-r)^2)\,dr\,ds = 2/3$. Gauss quadrature is exact here. An interior node belongs to four elements, yielding $(K_0)_{ii} \le 4\nu_+(2/3)$.

Finally let $\mathbf{v}_i \in \mathbb{R}^{n_P}$ be nodal values, extended by zero at boundary nodes, and let

$$U_{Tq}(\mathbf{v}) = \sum_{i \in T} \mathbf{v}_i \mathbf{g}_{Tiq}^\top \in \mathbb{R}^{n_P \times 2},$$

then

$$\mathbf{v}^\top \overline{\mathcal{K}} \mathbf{v} = \sum_{T,q} w_{Tq} a_{0,Tq} \|U_{Tq}(\mathbf{v})\|_F^2,$$

while $\mathbf{v}^\top \mathcal{K} \mathbf{v}$ replaces the squared Frobenius norm by $\operatorname{tr}(U_{Tq}(\mathbf{v})^\top A_{Tq} U_{Tq}(\mathbf{v}))$. The eigenvalue bounds on A_{Tq} prove the final assertion. ∎

For each interior node i, let $\mathcal{Q}_i$ consist of all incident element–quadrature pairs. We take A_i to be the arithmetic average

$$A_i = \frac{1}{|\mathcal{Q}_i|} \sum_{(T,q)\in\mathcal{Q}_i} A_{Tq}, \qquad \Omega_h = \max_i \max_{(T,q)\in\mathcal{Q}_i} \frac{\|A_{Tq} - A_i\|_2}{h}. \tag{5.12}$$

Each A_i is a convex combination of the symmetric blocks A_{Tq}, and the set $\{X = X^\top : \alpha I \preceq X \preceq \beta I\}$ is convex, then (5.8) gives $\alpha I \preceq A_i \preceq \beta I$. Consequently the matrix D defined next is symmetric positive definite, with $\|\mathsf{D}\|_2 \le \beta$ and $\|\mathsf{D}^{-1}\|_2 \le \alpha^{-1}$. Define the nodal block-diagonal multiplication matrix $\mathsf{D} = \mathrm{diag}_i(A_i)$ and set

$$\mathcal{K} = \mathsf{D}\overline{\mathcal{K}} + R. \tag{5.13}$$

Lemma 5.5. *Under Assumption 5.3, the remainder in (5.13) satisfies*

$$\|R\mathbf{v}\|_{\mathcal{M}^{-1}} \le C_R \Omega_h \|\mathbf{v}\|_{\overline{\mathcal{K}}}, \qquad C_R = \sqrt{96\nu_+}. \tag{5.14}$$

Proof. Subtracting the two assembled rows before the boundary values are eliminated, we find that the ith block row is exactly

$$(R\mathbf{v})_i = \sum_{(T,q)\in\mathcal{Q}_i} w_{Tq} a_{0,Tq} (A_{Tq} - A_i) U_{Tq}(\mathbf{v}) \mathbf{g}_{Tiq}.$$

Since $\|Ug\|_2 \le \|U\|_F \|g\|_2$, weighted Cauchy–Schwarz gives

$$\|(R\mathbf{v})_i\|_2^2 \le h^2 \Omega_h^2 \left(\sum_{\mathcal{Q}_i} w_{Tq} a_{0,Tq} \|\mathbf{g}_{Tiq}\|_2^2 \right) \left(\sum_{\mathcal{Q}_i} w_{Tq} a_{0,Tq} \|U_{Tq}(\mathbf{v})\|_F^2 \right).$$

The first parenthesis equals $(K_0)_{ii} \le 8\nu_+/3$. Upon summing over interior nodes, any element energy is counted at most four times. Thus

$$\|R\mathbf{v}\|_2^2 \le \frac{32\nu_+}{3} h^2 \Omega_h^2 \mathbf{v}^\top \overline{\mathcal{K}} \mathbf{v}.$$

∎

By (5.11), $M_D^{-1} \preceq 9h^{-2} I$, hence $\mathcal{M}^{-1} = I_{n_P} \otimes M_D^{-1} \preceq 9h^{-2} I$ and

$$\|R\mathbf{v}\|_{\mathcal{M}^{-1}}^2 = (R\mathbf{v})^\top \mathcal{M}^{-1} (R\mathbf{v}) \le (R\mathbf{v})^\top (9h^{-2} I)(R\mathbf{v}) = 9h^{-2} \|R\mathbf{v}\|_2^2 \le 96\, \nu_+ \Omega_h^2\, \|\mathbf{v}\|_{\overline{\mathcal{K}}}^2.$$

Taking square roots gives (5.14) with $C_R = \sqrt{96\nu_+}$. All vector bounds use operator norms of the coefficient blocks. The Frobenius norm above sums the physical gradient energies, not the entries of A_{Tq}.

Lemma 5.6 (Multiplication and discrete energy estimates). *For all $\mathbf{f} \in \mathbb{R}^N$,*

$$\|\mathsf{D}\mathbf{f}\|_{\mathcal{M}^{-1}} \le 3\beta \|\mathbf{f}\|_{\mathcal{M}^{-1}}, \qquad \|\mathsf{D}^{-1}\mathbf{f}\|_{\mathcal{M}^{-1}} \le \frac{3}{\alpha} \|\mathbf{f}\|_{\mathcal{M}^{-1}}. \tag{5.15}$$

With $c_* := \lambda_*^{-1/2}$, one also has

$$\|\mathbf{v}\|_{\overline{\mathcal{K}}} \le c_* \|\overline{\mathcal{K}}\mathbf{v}\|_{\mathcal{M}^{-1}}, \qquad \|\mathbf{v}\|_{\overline{\mathcal{K}}} \le \frac{c_*}{\alpha} \|\mathcal{K}\mathbf{v}\|_{\mathcal{M}^{-1}}. \tag{5.16}$$

Proof. Since the spectral condition number $\mathrm{cond}_2(\mathcal{M}) = \lambda_{\max}(\mathcal{M})/\lambda_{\min}(\mathcal{M})$, the operator norm of the nodal multiplication, which multiplies the gPC coefficients at each node i by A_i, satisfies

$$\|\mathcal{M}^{-1/2}\mathsf{D}\mathcal{M}^{1/2}\|_2 \le \sqrt{\mathrm{cond}_2(\mathcal{M})}\|\mathsf{D}\|_2 \le 3\beta.$$

Applying the same argument to D^{-1} yields $3/\alpha$.

Since $\overline{\mathcal{K}} = I_{n_P} \otimes K_0 \succeq \lambda_*\mathcal{M}$ by (5.11), we have $\|\mathbf{v}\|_{\mathcal{M}} \le c_*\|\mathbf{v}\|_{\overline{\mathcal{K}}}$. The weighted Cauchy–Schwarz inequality $|\mathbf{a}^\top\mathbf{b}| \le \|\mathbf{a}\|_{\mathcal{M}}\|\mathbf{b}\|_{\mathcal{M}^{-1}}$ with $\mathbf{a} = \mathbf{v}$, $\mathbf{b} = \overline{\mathcal{K}}\mathbf{v}$ then gives

$$\|\mathbf{v}\|_{\overline{\mathcal{K}}}^2 = \mathbf{v}^\top\overline{\mathcal{K}}\mathbf{v} \le \|\mathbf{v}\|_{\mathcal{M}}\|\overline{\mathcal{K}}\mathbf{v}\|_{\mathcal{M}^{-1}} \le c_*\|\mathbf{v}\|_{\overline{\mathcal{K}}}\|\overline{\mathcal{K}}\mathbf{v}\|_{\mathcal{M}^{-1}},$$

Similarly,

$$\alpha\|\mathbf{v}\|_{\overline{\mathcal{K}}}^2 \le \mathbf{v}^\top\mathcal{K}\mathbf{v} \le \|\mathbf{v}\|_{\mathcal{M}}\|\mathcal{K}\mathbf{v}\|_{\mathcal{M}^{-1}} \le c_*\|\mathbf{v}\|_{\overline{\mathcal{K}}}\|\mathcal{K}\mathbf{v}\|_{\mathcal{M}^{-1}}.$$

Dividing by $\|\mathbf{v}\|_{\overline{\mathcal{K}}}$ (the case $\mathbf{v} = 0$ is trivial) yields (5.16). ∎

THEOREM 5.7. *Under Assumption 5.3, let*

$$q_h = C_R c_* \Omega_h, \qquad U_h = 3\beta + q_h, \qquad V_h = \frac{3}{\alpha}\left(1 + \frac{q_h}{\alpha}\right). \tag{5.17}$$

Then

$$V_h^{-1}\|\overline{\mathcal{K}}\mathbf{v}\|_{\mathcal{M}^{-1}} \le \|\mathcal{K}\mathbf{v}\|_{\mathcal{M}^{-1}} \le U_h\|\overline{\mathcal{K}}\mathbf{v}\|_{\mathcal{M}^{-1}}, \qquad V_h^{-2}H_0 \preceq H \preceq U_h^2 H_0. \tag{5.18}$$

No smallness of q_h, Ω_h, or $\mathcal{K} - \overline{\mathcal{K}}$ is required.

Proof. The upper bound follows from (5.13) and Lemmas 5.5 and 5.6:

$$\|\mathcal{K}\mathbf{v}\|_{\mathcal{M}^{-1}} \le 3\beta\|\overline{\mathcal{K}}\mathbf{v}\|_{\mathcal{M}^{-1}} + C_R\Omega_h\|\mathbf{v}\|_{\overline{\mathcal{K}}} \le U_h\|\overline{\mathcal{K}}\mathbf{v}\|_{\mathcal{M}^{-1}}.$$

For the other one, let $\mathbf{f} = \mathcal{K}\mathbf{v}$, then the decomposition gives $\overline{\mathcal{K}}\mathbf{v} = \mathsf{D}^{-1}(\mathbf{f} - R\mathbf{v})$, and therefore

$$\|\overline{\mathcal{K}}\mathbf{v}\|_{\mathcal{M}^{-1}} \le \frac{3}{\alpha}\big(\|\mathbf{f}\|_{\mathcal{M}^{-1}} + C_R\Omega_h\|\mathbf{v}\|_{\overline{\mathcal{K}}}\big) \le \frac{3}{\alpha}\left(1 + \frac{C_R c_*\Omega_h}{\alpha}\right)\|\mathbf{f}\|_{\mathcal{M}^{-1}}.$$

Here the second inequality uses energy coercivity of $\mathcal{K}$. Finally, $\|\mathcal{K}\mathbf{v}\|_{\mathcal{M}^{-1}}^2 = \mathbf{v}^\top H\mathbf{v}$ and $\|\overline{\mathcal{K}}\mathbf{v}\|_{\mathcal{M}^{-1}}^2 = \mathbf{v}^\top H_0\mathbf{v}$, so squaring gives the matrix inequalities. ∎

COROLLARY 5.8. *Let the side length L_D be fixed, and suppose the local constants $\nu_-, \nu_+, \alpha, \beta$ can be chosen uniformly over a family of meshes and stochastic spaces, with $\nu_-, \alpha > 0$, and suppose $\Omega_h \le \Omega$ uniformly. Define*

$$q = C_R c_* \Omega, \qquad U = 3\beta + q, \qquad V = \frac{3}{\alpha}(1 + q/\alpha).$$

Then for every member of the family, every $\gamma > 0$, and every current mask,

$$\mathrm{spec}(P_W^{-1}J) \subset [\min\{V^{-2}, 1\}, \max\{U^2, 1\}]. \tag{5.19}$$

The interval is independent of h, n_P, N_K, γ, and the mask $D(\mathbf{p})$.

Proof. The constants in (5.11) and (5.14) depend only on d, ν_-, ν_+. Hence Theorem 5.7 gives uniform constants U, V, and Theorem 5.2 applies with $c = V^{-2}$ and $C = U^2$. ∎

5.3. Perturbation bounds and PCG convergence. Back to stochastic-major ordering, multiplication in (5.4) gives the exact identity

$$\mathcal{T} = I + \mathcal{E}, \qquad \mathcal{E} = \sum_{\ell=1}^{N_K-1} G_\ell \otimes B_\ell, \qquad B_\ell = M_D^{-1/2} K_\ell K_0^{-1} M_D^{1/2}. \tag{5.20}$$

PROPOSITION 5.9 (Small-perturbation enclosure). *If $\|\mathcal{E}\|_2 \le \delta < 1$, then*

$$\operatorname{spec}(P_W^{-1} J) \subset [(1-\delta)^2, (1+\delta)^2], \qquad \operatorname{cond}_{\mathrm{sp}}(\mathrm{P_W^{-1}J}) \le \left(\frac{1+\delta}{1-\delta}\right)^2. \tag{5.21}$$

Moreover,

$$\|\mathcal{E}\|_2 \le \sum_{\ell=1}^{N_K-1} \|G_\ell\|_2 \|B_\ell\|_2. \tag{5.22}$$

Proof. The triangle and reverse triangle inequalities give $(1-\delta)\|\mathbf{z}\|_2 \le \|\mathcal{T}\mathbf{z}\|_2 \le (1+\delta)\|\mathbf{z}\|_2$. Hence $\sigma_- \ge 1-\delta$ and $\sigma_+ \le 1+\delta$, so Lemma 5.1 gives $(1-\delta)^2 H_0 \preceq H \preceq (1+\delta)^2 H_0$. Theorem 5.2 with $c = (1-\delta)^2 \le 1 \le C = (1+\delta)^2$ then yields (5.21). For (5.22), the triangle inequality gives $\|\mathcal{E}\|_2 \le \sum_\ell \|G_\ell \otimes B_\ell\|_2$, and $\|G_\ell \otimes B_\ell\|_2 = \|G_\ell\|_2 \|B_\ell\|_2$ because the singular values of a Kronecker product are the products of the singular values of its factors (tensor the two singular value decompositions). ∎

The following proposition gives computable bounds on $\|B_\ell\|_2$ and $\|\mathcal{E}\|_2$ from local coefficient values, rather than assuming the desired graph-matrix bound. Let K_ℓ be assembled with scalar weights $a_{\ell,Tq}$, and let $\mu_{\ell,Tq} = a_{\ell,Tq}/a_{0,Tq}$. Define

$$d_{\ell i} = \frac{1}{|\mathcal{Q}_i|} \sum_{\mathcal{Q}_i} \mu_{\ell,Tq}, \quad b_\ell = \max_{T,q} |\mu_{\ell,Tq}|, \quad \omega_{\ell,h} = \max_{i,(T,q)\in\mathcal{Q}_i} \frac{|\mu_{\ell,Tq} - d_{\ell i}|}{h}.$$

PROPOSITION 5.10. *With the matrix assumptions above, signed fluctuation arrays satisfy*

$$\|B_\ell\|_2 \le 3b_\ell + C_R c_* \omega_{\ell,h}. \tag{5.23}$$

Alternatively, for the jointly assembled blocks, let $\tau_h = \max_i \|A_i - I\|_2$. Then

$$\|\mathcal{E}\|_2 \le 3\tau_h + C_R c_* \Omega_h. \tag{5.24}$$

Proof. Let $D_\ell = \operatorname{diag}(d_{\ell i})$ and $R_\ell = K_\ell - D_\ell K_0$. The proof of Lemma 5.5 carries over to the scalar case $n_P = 1$, with A_{Tq} and A_i replaced by $\mu_{\ell,Tq}$ and $d_{\ell i}$. Positivity is needed only for $a_{0,Tq}$, not for $\mu_{\ell,Tq}$. Since $|d_{\ell i}| \le b_\ell$, the argument for (5.15) also applies to D_ℓ. Hence

$$\|R_\ell \mathbf{v}\|_{M_D^{-1}} \le C_R \omega_{\ell,h} \|\mathbf{v}\|_{K_0}, \qquad \|D_\ell \mathbf{f}\|_{M_D^{-1}} \le 3b_\ell \|\mathbf{f}\|_{M_D^{-1}}.$$

Substituting $\mathbf{v} = K_0^{-1} M_D^{1/2} \mathbf{z}$ in the definition of B_ℓ gives

$$\|B_\ell\|_2 = \sup_{\mathbf{z}\neq 0} \frac{\|M_D^{-1/2} K_\ell K_0^{-1} M_D^{1/2} \mathbf{z}\|_2}{\|\mathbf{z}\|_2} = \sup_{\mathbf{v}\neq 0} \frac{\|K_\ell \mathbf{v}\|_{M_D^{-1}}}{\|K_0 \mathbf{v}\|_{M_D^{-1}}}.$$

Since $K_\ell = D_\ell K_0 + R_\ell$, the two estimates above with $\mathbf{f} = K_0\mathbf{v}$, together with the first inequality in (5.16) for $n_P = 1$, i.e., $\|\mathbf{v}\|_{K_0} \le c_*\|K_0\mathbf{v}\|_{M_D^{-1}}$, give

$$\|K_\ell\mathbf{v}\|_{M_D^{-1}} \le 3b_\ell\|K_0\mathbf{v}\|_{M_D^{-1}} + C_R\omega_{\ell,h}\|\mathbf{v}\|_{K_0} \le \big(3b_\ell + C_R c_*\omega_{\ell,h}\big)\|K_0\mathbf{v}\|_{M_D^{-1}}.$$

Taking the supremum over $\mathbf{v} \neq 0$ proves (5.23). For (5.24), we find $\mathcal{K} - \overline{\mathcal{K}} = (\mathsf{D} - I)\overline{\mathcal{K}} + R$ and apply the same estimates with the joint block norm τ_h. Then, taking the supremum over $\mathbf{v}$ gives the claim. ∎

COROLLARY 5.11. *Suppose, uniformly over the considered meshes and stochastic spaces,*

$$\sum_{\ell=1}^{N_K-1} \|G_\ell\|_2 b_\ell \le \eta < 1, \qquad \sum_{\ell=1}^{N_K-1} \|G_\ell\|_2\omega_{\ell,h} \le \Omega_0 < \infty.$$

Then (5.10) satisfies the uniform local hypotheses with $\alpha = 1-\eta$, $\beta = 1+\eta$, and $\Omega = \Omega_0$. Consequently Corollary 5.8 applies, whether or not $3\eta + C_R c_*\Omega_0 < 1$.

Proof. By (5.10) and the triangle inequality, $\|A_{Tq} - I\|_2 \le \sum_\ell |\mu_{\ell,Tq}|\,\|G_\ell\|_2 \le \sum_\ell b_\ell\|G_\ell\|_2 \le \eta$. Since every G_ℓ is symmetric, so is $A_{Tq} - I$. Hence its eigenvalues lie in $[-\eta,\eta]$, i.e., $(1-\eta)I \preceq A_{Tq} \preceq (1+\eta)I$. Linearity of the same nodal averaging gives $A_i = I + \sum_\ell d_{\ell i}G_\ell$ and

$$\|A_{Tq} - A_i\|_2 \le \sum_\ell \|G_\ell\|_2|\mu_{\ell,Tq} - d_{\ell i}| \le h\Omega_0.$$

∎

It remains to show how the preconditioner is applied.

PROPOSITION 5.12 (Exact current-set inverse). *Let $C_D = BB^\top$ with $B \in \mathbb{R}^{N\times r}$; the columns need not be independent. From the SMW identity*

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U\big(C^{-1} + VA^{-1}U\big)^{-1}VA^{-1}.$$

Here we use $A = H_0, U = B, V = B^\top, C = \gamma^{-1}I_r$, gives

$$P_W^{-1} = (H_0 + B\,(\gamma^{-1}I_r)\,B^\top)^{-1} = H_0^{-1} - H_0^{-1}B\big(\gamma I_r + B^\top H_0^{-1}B\big)^{-1}B^\top H_0^{-1}. \tag{5.25}$$

For a binary mask, one may take the columns of $E^\top W^{1/2}$ selected by D. Reusing unchanged columns of $B, H_0^{-1}B$, computing added columns, and discarding removed columns preserves this inverse if the correction matrix is formed and solved for the current columns.

Proof. The matrix $Z_\gamma = \gamma I_r + B^\top H_0^{-1}B$ is SPD since $\mathbf{q}^\top Z_\gamma\mathbf{q} \ge \gamma\|\mathbf{q}\|_2^2$ for all $\mathbf{q} \in \mathbb{R}^r$. Multiplying the right-hand side of (5.25) by P_W on the left gives

$$I + B\left[\gamma^{-1}I_r - Z_\gamma^{-1} - \gamma^{-1}B^\top H_0^{-1}BZ_\gamma^{-1}\right](H_0^{-1}B)^\top = I.$$

For $r = 0$, the correction is absent. Because H_0 is fixed, an unchanged column has the same inverse image. Reordering the columns simply permutes Z_γ. Finally, we have

$$H_0^{-1} = I_{n_P} \otimes (K_0^{-1}M_D K_0^{-1}),$$

so its action uses reference stiffness solves and a mass multiplication, without assembling the double-stiffness matrix. ∎

COROLLARY 5.13 (PCG consequence). *Consider one Newton system $J\mathbf{x} = \mathbf{b}$ with current-set preconditioner P_W. If (5.7) holds, let $\chi = \theta_+/\theta_-$. With exact arithmetic and exact preconditioner applications, PCG satisfies*

$$\|\mathbf{x}_k - \mathbf{x}_*\|_J \le 2\left(\frac{\sqrt{\chi}-1}{\sqrt{\chi}+1}\right)^k \|\mathbf{x}_0 - \mathbf{x}_*\|_J. \tag{5.26}$$

Under Proposition 5.9, the right-hand side is bounded by $2\delta^k\|\mathbf{x}_0 - \mathbf{x}_*\|_J$.

Proof. PCG is CG for $P_W^{-1/2} J P_W^{-1/2}$ after the substitution $\widehat{\mathbf{x}} = P_W^{1/2}\mathbf{x}$. Its energy-error norm equals $\|\mathbf{x}-\mathbf{x}_*\|_J$. CG minimizes the SPD quadratic over its affine Krylov space. The Galerkin orthogonality of that minimizer bounds its relative energy error by the maximum modulus of any degree-k error polynomial with value one at zero on the spectral interval. The normalized Chebyshev polynomial gives the factor in (5.26); see, e.g., [14]. In the perturbation case, $\sqrt{\chi} \le (1+\delta)/(1-\delta)$, and substitution gives the factor δ. ∎

6. Numerical Experiments. This section reports two examples. Example 1 has a manufactured solution and measures the discretization error together with the solver cost. Example 2 has no known solution and measures how the solver cost depends on the noise amplitude, the stochastic space, the mesh and γ. The stochastic approximation uses total-degree Legendre polynomials in random variables. We use continuous bilinear finite elements on uniform Cartesian meshes with 2^{nc} elements in each coordinate direction, that is, $n_e = 2^{nc}$ and $h = 2^{1-nc}$ in the notation of Assumption 5.3 , where $nc = 4, \dots, 8$. All runs start from the zero dual variable and use a normalized KKT stopping tolerance of 10^{-6}. The inexact Newton forcing tolerance is bounded below by 10^{-12} and above by $\eta_{\max} = 10^{-2}$, and at most $k_{\max} = 100$ Newton steps are allowed in both examples. Successful runs additionally satisfy a box-constraint violation not exceeding 10^{-12}. In Example 1, they also satisfy a normalized primal–dual gap not exceeding 10^{-9}. The computations use MATLAB R2023b on the 26-core Intel Xeon Gold 6230R machine.

We use the notation of (5.4): $\mathcal{M} = I_{n_P} \otimes M_D$, $\overline{\mathcal{K}} = I_{n_P} \otimes K_0$ and $H_0 = \overline{\mathcal{K}}\mathcal{M}^{-1}\overline{\mathcal{K}}$. The stochastic quadrature integrates $\psi_i\psi_j$ exactly, so $V_\xi^\top W_\xi V_\xi = I_{n_P}$ and $E^\top W E = I_{n_P} \otimes M_L \approx \mathcal{M}$.

If every control value is free ($D = I_m$) and $\mathcal{K}$ is replaced by $\overline{\mathcal{K}}$, the Jacobian matrix (5.2) becomes approximately $H_0 + \gamma^{-1}\mathcal{M}$. This sum is not factorized directly. Following the matching strategy of [29], we use

$$\begin{aligned} P_{\text{mean}} &= (\overline{\mathcal{K}} + \gamma^{-1/2}\mathcal{M})\mathcal{M}^{-1}(\overline{\mathcal{K}} + \gamma^{-1/2}\mathcal{M}) \\ &= I_{n_P} \otimes \big[(K_0 + \gamma^{-1/2}M_D)M_D^{-1}(K_0 + \gamma^{-1/2}M_D)\big]. \end{aligned} \tag{6.1}$$

Expanding (6.1) gives

$$P_{\text{mean}} = H_0 + 2\gamma^{-1/2}\overline{\mathcal{K}} + \gamma^{-1}\mathcal{M} = I_{n_P} \otimes \big[K_0 M_D^{-1} K_0 + 2\gamma^{-1/2}K_0 + \gamma^{-1}M_D\big].$$

The two outer terms reproduce $H_0 + \gamma^{-1}\mathcal{M}$ exactly. The cross term is controlled by $(\overline{\mathcal{K}} - \gamma^{-1/2}\mathcal{M})\mathcal{M}^{-1}(\overline{\mathcal{K}} - \gamma^{-1/2}\mathcal{M}) \succeq 0$, which gives

$$H_0 + \gamma^{-1}\mathcal{M} \preceq P_{\text{mean}} \preceq 2\,(H_0 + \gamma^{-1}\mathcal{M}).$$

Only the sparse spatial matrix $\widehat{K}_\gamma = K_0 + \gamma^{-1/2}M_D \in \mathbb{R}^{n_h \times n_h}$ is factorized by Cholesky decomposition. Using the identity $(B^\top \otimes A)\,\mathrm{vec}(X) = \mathrm{vec}(AXB)$, we observe from (6.1) that

$$P_{\text{mean}}^{-1}\mathbf{r} = \mathrm{vec}\big(\widehat{K}_\gamma^{-1} M_D \widehat{K}_\gamma^{-1} R\big),$$

where $\mathbf{r} = \mathrm{vec}(R)$ with $R \in \mathbb{R}^{n_h \times n_P}$.

In practice, part of the control is saturated, and the term $\gamma^{-1}\mathcal{M}$ overestimates $\gamma^{-1}E^\top W D(\mathbf{p}_k)E$. At Newton step k, we therefore replace γ^{-1} by $\widehat{\vartheta}_k/\gamma$, where

$$\vartheta_k = \frac{1}{m}\,\mathrm{Tr}\, D(\mathbf{p}_k) \in [0,1],$$

and Tr stands for the trace operator. Next, let $c_k = \sqrt{\widehat{\vartheta}_k/\gamma}$ and $\widehat{K}_k = K_0 + c_k M_D$. We then define an approximation P_k of P_{mean} as follows:

$$P_k = (\overline{\mathcal{K}} + c_k\mathcal{M})\mathcal{M}^{-1}(\overline{\mathcal{K}} + c_k\mathcal{M}), \qquad P_k^{-1}\mathbf{r} = \mathrm{vec}\big(\widehat{K}_k^{-1} M_D \widehat{K}_k^{-1} R\big).$$

Observe then that $P_k = P_{\mathrm{mean}}$ when $\widehat{\vartheta}_k = 1$ and $P_k = H_0$ when $\widehat{\vartheta}_k = 0$. Note that $\widehat{K}_k$ is not recomputed at every Newton step. Let $t_k = \vartheta_k/\gamma$, and let t_{last} be its value at the last refactorization. We refactorize only at the first Newton step or when

$$|t_k - t_{\mathrm{last}}| > 0.15\max\{|t_k|, |t_{\mathrm{last}}|, \sqrt{\varepsilon_{\mathrm{mach}}}\}.$$

In that case $\widehat{\vartheta}_k = \vartheta_k$; otherwise $\widehat{\vartheta}_k$ is the fraction stored at the last refactorization and the old factor is reused. The free set enters only through the scalar ϑ_k, so no information on the location of the free set is needed.

Algorithm 6.1 Direct semismooth Newton–PCG for the reduced dual equation (5.1).

Set $\mathbf{p}_0 = 0$ and evaluate $F(\mathbf{p}_0)$ and $D(\mathbf{p}_0)$.
for $k = 0, 1, 2, \ldots$ **do**
  Compute the free fraction ϑ_k. Refactorize $\widehat{K}_k$ if the test above is met, otherwise reuse the stored factor. This defines P_k.
  Solve $J_D\,\delta\mathbf{p}_k = -F(\mathbf{p}_k)$ by PCG with preconditioner P_k to the tolerance η_k of (5.3).
  Choose α_k by Armijo backtracking and set $\mathbf{p}_{k+1} = \mathbf{p}_k + \alpha_k\,\delta\mathbf{p}_k$.
  Recompute the projection, the active set, the residual r_{k+1} and the stopping indicators.
  if r_{k+1} meets the stopping tolerance or $k+1 = k_{\max}$ **then**
    break
  end if
end for
return $\mathbf{y} = \mathbf{y}_d - \mathcal{M}^{-1}\mathcal{K}^\top\mathbf{p}$ and $\mathbf{u} = \mathrm{Proj}_{[a,b]}(E\mathbf{p}/\gamma)$

Example 1 We consider a control box-constrained stochastic elliptic optimal control problem on $D = (-1,1)^2$. Let ξ_1, ξ_2, ξ_3 be independent random variables uniformly distributed on $[-1,1]$, and set

$$\Gamma = [-1,1]^3, \qquad \rho(\xi) = \frac{1}{8}, \qquad \kappa(\xi) = 10 + \sum_{\ell=1}^{3}\xi_\ell.$$

Thus $7 \le \kappa(\xi) \le 13$ throughout the parameter domain. The optimization problem is

$$\begin{aligned} \min_{y,u} \quad & \frac{1}{2}\mathbb{E}\Big[\|y - y_d\|^2_{L^2(D)}\Big] + \frac{\gamma}{2}\mathbb{E}\Big[\|u\|^2_{L^2(D)}\Big], \\ \text{subject to} \quad & -\nabla\cdot\big(\kappa(\xi)\nabla y(x,\xi)\big) = f(x,\xi) + u(x,\xi), \\ & y(x,\xi) = 0 \quad \text{on } \partial D, \\ & -1 \le u(x,\xi) \le 1 \quad \text{a.e. in } D\times\Gamma, \end{aligned} \tag{6.2}$$

where $\gamma = 10^{-3}$.

To obtain a known optimal solution, define

(6.3)
$$y^*(x,\xi) = \sin(\pi x_1)\sin(\pi x_2), p^*(x,\xi) = \frac{\beta}{\kappa(\xi)} s(x), u^*(x,\xi) = \mathrm{Proj}_{[-1,1]}\left(\frac{p^*(x,\xi)}{\gamma}\right).$$

Here $s(x) = \sin(\pi x_1)\sin(\pi x_2)$, so that $y^* = s$, and $\beta = 0.02$. The corresponding target and source are

$$y_d(x) = (1 + 2\pi^2\beta)s(x), \qquad f(x,\xi) = 2\pi^2\kappa(\xi)s(x) - u^*(x,\xi). \tag{6.4}$$

These choices satisfy the state equation and the adjoint relation

$$-\nabla \cdot \big(\kappa(\xi)\nabla p^*(x,\xi)\big) = y_d(x) - y^*(x,\xi),$$

together with the box-projection optimality condition. The sign convention is therefore consistent with $u^* = \mathrm{Proj}_{[-1,1]}(p^*/\gamma)$.

For the accuracy assessment, define

$$\|v\|_{0,\rho}^2 = \int_\Gamma \int_D |v(x,\xi)|^2 \, dx \, \rho(\xi) \, d\xi,$$

and report the absolute errors

$$e_y = \|y_h - y^*\|_{0,\rho}, \qquad e_p = \|p_h - p^*\|_{0,\rho}, \qquad e_u = \|u_h - u^*\|_{0,\rho}.$$

At independent evaluation points, the control is reconstructed as

$$u_h(x,\xi) = \mathrm{Proj}_{[-1,1]}\left(\frac{p_h(x,\xi)}{\gamma}\right),$$

The empirical state convergence rate is defined by

$$\mathrm{EOC}_y(nc) = \log_2\left(\frac{e_y(nc-1)}{e_y(nc)}\right),$$

where $e_y(nc)$ is the error on the mesh with 2^{nc} elements per direction.

TABLE 1
SSN–PCG performance and absolute $L^2_\rho(\Gamma; L^2(D))$ errors for the manufactured problem. The stochastic degree is fixed at $P = 3$ with $\beta = 0.02$. PCG mean/max refers to the iterations per Newton step. Here $\gamma = 10^{-3}$, and PCG uses the preconditioner P_k of Section 6.

nc	N	Newton	PCG total	PCG mean/max	e_y	e_p	e_u	EOC_y
4	4,500	3	16	5.33/7	1.5116×10^{-2}	4.1864×10^{-5}	1.6610×10^{-2}	2.004
5	19,220	3	16	5.33/7	3.7607×10^{-3}	1.0550×10^{-5}	4.1985×10^{-3}	2.007
6	79,380	3	16	5.33/7	9.3835×10^{-4}	2.6878×10^{-6}	1.0740×10^{-3}	2.003
7	322,580	3	16	5.33/7	2.3437×10^{-4}	8.5077×10^{-7}	3.3461×10^{-4}	2.001
8	1,300,500	3	16	5.33/7	5.8571×10^{-5}	5.6443×10^{-7}	2.1656×10^{-4}	2.001

Observe from Table 1 that the value of EOC_y for all the mesh sizes is roughly 2, which suggests that the method achieves approximately second-order convergence in the state. Moreover, we notice small errors corresponding to the state, the adjoint, and the control on fine meshes. On every mesh the solver takes three Newton steps and 16 PCG iterations in total, at most seven per step, while N grows by a factor of 289 from 4,500 to 1,300,500.

Example 2 is designed for stress test for parameters to verify the preconditioner robustness.

Example 2 The optimization problem is

$$\begin{aligned} \min_{y,u} \quad & \frac{1}{2}\mathbb{E}\Big[\|y-y_d\|^2_{L^2(D)}\Big] + \frac{\gamma}{2}\mathbb{E}\Big[\|u\|^2_{L^2(D)}\Big], \\ \text{subject to} \quad & -\nabla\cdot\big(\kappa(x,\xi)\nabla y(x,\xi)\big) = u(x,\xi), \\ & y(x,\xi) = 0 \quad \text{on } \partial D, \\ & 0.3 \le u(x,\xi) \le 1 \quad \text{a.e. in } D\times\Gamma, \end{aligned} \tag{6.5}$$

and the random coefficient $\kappa(x,\xi)$ is the truncated Karhunen–Loève (KL) expansion $\kappa(x,\xi) = 1 + \sqrt{3}\,\sigma\sum_{j=1}^{K}\sqrt{\lambda_j}\,g_j(x)\xi_j$ with $\xi_j \sim U[-1,1]$, where (λ_j, g_j) are the eigenpairs of the covariance function

$$C_g(x,y) = \sigma^2 \exp\left(-\frac{|x_1-y_1|}{\ell_1} - \frac{|x_2-y_2|}{\ell_2}\right) \quad \forall (x,y)\in[-1,1]^2,$$

where σ denotes the standard deviation and $\ell_1 = \ell_2 = 1$ are correlation lengths. The discontinuous target is defined on $D = (-1,1)^2$ as

$$y_d(x,\xi) = \begin{cases} 1, & x_1 \le 0, \quad x_2 \le 0, \\ 0, & \text{otherwise.} \end{cases}$$

We fix $\gamma = 10^{-3}$ and vary the amplitude $\sigma \in \{0.01, 0.02, 0.05, 0.1, 0.2\}$, the pair $(K,P) \in \{(3,3),(4,4),(6,3)\}$ of the number of retained expansion terms and the total polynomial degree, and the mesh level $nc = 4,\dots,8$. The total-degree Legendre space has $n_P = \binom{K+P}{P} \in \{20, 70, 84\}$ modes, and with $n_h = (2^{nc}-1)^2$ interior nodes the SG–FE system has $N = n_h n_P$ unknowns, from 4,500 for $(K,P,nc) = (3,3,4)$ to 5,462,100 for $(K,P,nc) = (6,3,8)$. Example 2 tests the solver on a problem without a known solution. Since $u \ge 0.3$ while $y_d = 0$ on three quarters of D, the lower bound is expected to be active on a large part of the domain, and the active set changes during the iteration.

TABLE 2

Each entry is Newton / mean PCG (maximum PCG) / solve time in seconds with $\gamma = 10^{-3}$. Example 2 with the preconditioner P_k. Here σ is the amplitude, K is the number of random variables, and P is the degree of the polynomials.

σ	(K,P)	$nc=4$	$nc=5$	$nc=6$	$nc=7$	$nc=8$
0.01	(3,3)	8 / 5.50 (9) / 0.162	7 / 5.71 (8) / 0.31	7 / 5.86 (9) / 1.49	7 / 6.43 (10) / 6.78	6 / 6.33 (10) / 25.6
	(4,4)	9 / 6.22 (11) / 0.444	7 / 5.71 (8) / 1.17	7 / 6.00 (9) / 5.33	7 / 6.29 (10) / 21.7	6 / 6.33 (10) / 97.9
	(6,3)	8 / 5.38 (8) / 0.845	7 / 5.71 (8) / 2.69	7 / 6.00 (10) / 10.8	6 / 5.83 (9) / 37.5	6 / 6.33 (10) / 168
0.02	(3,3)	8 / 5.75 (9) / 0.175	7 / 5.86 (8) / 0.295	7 / 6.29 (10) / 1.53	7 / 6.86 (10) / 7.11	6 / 6.83 (11) / 23.3
	(4,4)	8 / 5.62 (9) / 0.38	8 / 6.25 (10) / 1.42	7 / 6.43 (10) / 5.48	6 / 6.33 (10) / 21.7	6 / 6.83 (11) / 102
	(6,3)	8 / 5.62 (9) / 0.872	7 / 5.86 (8) / 2.75	7 / 6.43 (10) / 11.1	6 / 6.33 (10) / 38.9	6 / 6.83 (11) / 155
0.05	(3,3)	8 / 6.25 (9) / 0.177	8 / 6.88 (10) / 0.367	7 / 6.86 (10) / 1.61	6 / 6.67 (10) / 5.93	6 / 7.17 (10) / 27.6
	(4,4)	8 / 6.50 (10) / 0.429	7 / 6.57 (9) / 1.24	7 / 7.00 (11) / 5.67	6 / 6.83 (10) / 22.5	6 / 7.17 (10) / 104
	(6,3)	8 / 6.38 (10) / 0.914	7 / 6.43 (10) / 2.83	7 / 7.00 (11) / 12.2	6 / 6.67 (10) / 39.5	6 / 7.17 (10) / 177
0.1	(3,3)	7 / 6.29 (9) / 0.146	7 / 7.29 (11) / 0.343	7 / 8.29 (13) / 1.76	7 / 8.71 (13) / 8.61	6 / 8.50 (13) / 29.6
	(4,4)	7 / 6.29 (9) / 0.379	7 / 7.29 (11) / 1.32	7 / 8.43 (13) / 6.48	6 / 8.00 (13) / 24.4	6 / 8.67 (13) / 115
	(6,3)	8 / 7.00 (12) / 0.984	7 / 7.43 (11) / 2.92	7 / 8.43 (13) / 13.6	6 / 8.00 (13) / 42.9	6 / 8.50 (13) / 191
0.2	(3,3)	7 / 8.57 (12) / 0.18	7 / 9.14 (12) / 0.381	7 / 10.43 (16) / 2.04	6 / 10.00 (14) / 7.41	6 / 10.50 (14) / 34.1
	(4,4)	8 / 9.25 (15) / 0.506	7 / 9.86 (15) / 1.38	7 / 10.71 (16) / 7.33	6 / 10.17 (14) / 27.7	6 / 11.00 (14) / 115
	(6,3)	8 / 9.50 (16) / 1.08	7 / 9.43 (14) / 4.96	6 / 9.67 (14) / 11.6	6 / 10.00 (14) / 47.7	6 / 10.50 (14) / 212

TABLE 3
Dependence on γ of Direct SSN–PCG with the preconditioner P_k at $(K, P, nc, \sigma) = (3, 3, 4, 0.1)$. Entries report Newton steps, total PCG, mean/max PCG, and solve time in seconds.

γ	Newton	PCG	mean/max	time
10^{-1}	4	20	5.00/7	0.074
10^{-2}	6	36	6.00/9	0.138
10^{-3}	7	44	6.29/9	0.146
10^{-4}	16	168	10.50/16	0.364
10^{-5}	48	937	19.52/33	2.066
10^{-6}	47	2,106	44.81/127	2.980

Table 2 shows three trends. First, the Newton count is insensitive to the discretization. Over all 75 runs it lies between 6 and 9, and it does not grow under refinement: it decreases from 7–9 at $nc = 4$ to exactly 6 at $nc = 8$ for every σ and every (K, P). This is in line with the mesh-independent behaviour of semismooth Newton methods [19].

Second, the number of PCG iterations per Newton step is essentially independent of the mesh and of the stochastic discretization. For fixed σ and (K, P), the mean count rises by at most about two iterations from $nc = 4$ to $nc = 6$ and then stays nearly constant. For fixed σ and nc, the three choices of (K, P) differ by about one iteration or less, although n_P grows from 20 to 84.

Third, the amplitude σ is the only parameter with an effect on PCG. At $nc = 8$, as σ grows from 0.01 to 0.2, the mean count rises from 6.33 to 10.5–11.0 and the maximum from 10 to 14. This is consistent with the theory: the reference H_0 uses only the mean coefficient, and the constants U and V of Corollary 5.8 grow with the deviation of the coefficient from its mean through α, β and Ω. Even for $\sigma = 0.2$ the growth is moderate, and the Newton count is unaffected. The solution time grows reasonably with the scale of the problem, yet it remains within an acceptable range, even the most difficult one can still be solved in 212 seconds.

Table 3 separates the two levels of the method. Its row $\gamma = 10^{-3}$ is the entry $\sigma = 0.1$, $(K, P) = (3, 3)$, $nc = 4$ of Table 2. For $\gamma \geq 10^{-3}$, the Newton count lies between 4 and 7, and the mean PCG count per Newton step lies between 5.00 and 6.29. Below $\gamma = 10^{-3}$ both counts grow. At $\gamma = 10^{-6}$ the solver takes 47 Newton steps with 44.81 PCG iterations per step on average and 127 at most, that is, 2,106 PCG iterations in total against 44 at $\gamma = 10^{-3}$. This is the regime in which P_k stops being robust. The growth of the Newton count is a property of the nonlinear iteration: the projection term in (5.1) has slope γ^{-1}, so for small γ the reduced equation approaches a bang-bang system, and more steps are needed before the fast local convergence of the semismooth Newton method sets in. While, the solve time is acceptable and it grows from 0.074 s at $\gamma = 10^{-1}$ to 2.980 s at $\gamma = 10^{-6}$.

7. Conclusions. We have studied the optimal control problem of elliptic PDEs with random coefficients subject to a box-constrained control variable via Fenchel duality. In both the function space and the discrete setting via stochastic Galerkin–finite element methods, the dual formulation yields an unconstrained problem in the adjoint variable and reduces to a single semismooth equation without a duality gap. Our matrix-free semismooth Newton–PCG algorithm solves this equation directly without requiring outer iterations, penalties, or smoothing parameters. The current-set preconditioner P_W adds the active-set correction to a mean-based reference operator and is inverted exactly by the Sherman–Morrison–Woodbury identity. For P_W we prove

that the preconditioned spectrum is bounded uniformly in the mesh size, the stochastic dimension, γ and the active set. In the numerical experiments, the solver uses the cheap approximation P_k of the preconditioner P_W. The state solution converges at second order, and the Newton and PCG counts stay bounded under mesh refinement and enlargement of the stochastic space.

Future work includes developing continuation strategies for small regularization parameters, extending the duality framework to lognormal diffusion coefficients using weighted spaces, and exploring risk-averse objectives alongside alternative stochastic discretizations.

REFERENCES

[1] H. Antil, S. Dolgov, and A. Onwunta, *TTRISK: Tensor train decomposition algorithm for risk averse optimization*, Numer. Linear Algebra Appl., 30 (2023), p. e2481.

[2] H. Antil, S. Dolgov, and A. Onwunta, *Smoothed Moreau–Yosida tensor-train approximation of state-constrained optimization problems under uncertainty*, Numer. Linear Algebra Appl., 32 (2025), p. e70028.

[3] H. Antil, S. Dolgov, and A. Onwunta, *Smoothed Moreau-Yosida tensor-train approximation of state-constrained optimization problems under uncertainty*, Numer. Linear Algebra Appl., (2025), p. 32:e70028, https://doi.org/10.1002/nla.70028.

[4] I. Babuška, R. Tempone, and G. E. Zouraris, *Galerkin finite element approximations of stochastic elliptic partial differential equations*, SIAM J. Numer. Anal., 42 (2004), pp. 800–825.

[5] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, CMS Books in Mathematics, Springer, Cham, second ed., 2017.

[6] P. Benner, S. Dolgov, A. Onwunta, and M. Stoll, *Low-rank solvers for unsteady Stokes–Brinkman optimal control problem with random data*, Comput. Methods Appl. Mech. Engrg., 304 (2016), pp. 26–54.

[7] P. Benner, S. Dolgov, A. Onwunta, and M. Stoll, *Low-rank solution of an optimal control problem constrained by random Navier-Stokes equations*, International Journal on Numerical Methods for Fluids, 92 (2020), pp. 1653–1678.

[8] P. Benner, A. Onwunta, and M. Stoll, *Block-diagonal preconditioning for optimal control problems constrained by PDEs with uncertain inputs*, SIAM J. Matrix Anal. Appl., 37 (2016), pp. 491–518.

[9] M. Bergounioux, K. Ito, and K. Kunisch, *Primal-dual strategy for constrained optimal control problems*, SIAM J. Control Optim., 37 (1999), pp. 1176–1194.

[10] P. Chen and A. Quarteroni, *Weighted reduced basis method for stochastic optimal control problems with elliptic PDE constraint*, SIAM/ASA J. Uncertain. Quantif., 2 (2014), pp. 364–396.

[11] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Canadian Mathematical Society Series of Monographs and Advanced Texts, John Wiley & Sons, Inc., New York, 1983.

[12] C. Clason and K. Kunisch, *A duality-based approach to elliptic control problems in non-reflexive Banach spaces*, ESAIM Control Optim. Calc. Var., 17 (2011), pp. 243–266.

[13] I. Ekeland and R. Témam, *Convex analysis and variational problems*, vol. 28 of Classics in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, english ed., 1999, https://doi.org/10.1137/1.9781611971088, https://doi.org/10.1137/1.9781611971088. Translated from the French.

[14] H. Elman, D. Silvester, and A. Wathen, *Finite Elements and Fast Iterative Solvers: with Applications in Incompressible Fluid Dynamics*, Oxford University Press, 2014.

[15] F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Springer Series in Operations Research, Springer-Verlag, New York, 2003.

[16] R. G. Ghanem and P. D. Spanos, *Stochastic Finite Elements: A Spectral Approach*, Springer-Verlag New York, Inc., New York, NY, USA, 1991.

[17] W. W. Hager, *Updating the inverse of a matrix*, SIAM Rev., 31 (1989), pp. 221–239.

[18] M. Hintermüller, K. Ito, and K. Kunisch, *The primal-dual active set strategy as a semismooth Newton method*, SIAM J. Optim., 13 (2002), pp. 865–888.

[19] M. Hintermüller and M. Ulbrich, *A mesh-independence result for semismooth Newton methods*, Math. Program., 101 (2004), pp. 151–184.

[20] K. Ito and K. Kunisch, *Lagrange Multiplier Approach to Variational Problems and Appli-*

cations, vol. 15 of Advances in Design and Control, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008.

[21] D. P. Kouri, *A multilevel stochastic collocation algorithm for optimization of PDEs with uncertain coefficients*, SIAM/ASA J. Uncertain. Quantif., 2 (2014), pp. 55–81.

[22] D. P. Kouri, M. Heinkenschloss, D. Ridzal, and B. van Bloemen Waanders, *A trust-region algorithm with adaptive stochastic collocation for PDE optimization under uncertainty*, SIAM J. Sci. Comput., 35 (2013), pp. A1847–A1879.

[23] O. Le Maître and O. M. Knio, *Spectral Methods for Uncertainty Quantification: With Applications to Computational Fluid Dynamics*, Scientific Computation, Springer, 2010.

[24] H.-C. Lee and J. Lee, *A stochastic Galerkin method for stochastic control problems*, Commun. Comput. Phys., 14 (2013), pp. 77–106.

[25] Z. Li, A. Onwunta, and B. Sousedík, *Stochastic galerkin method and hierarchical preconditioning for pde-constrained optimization*, arXiv preprint arXiv:2512.23804, (2025).

[26] G. J. Lord, C. E. Powell, and T. Shardlow, *An Introduction to Computational Stochastic PDEs*, Cambridge Texts in Applied Mathematics, Cambridge University Press, 2014.

[27] J. Martínez-Frutos and F. Periago Esparza, *Optimal Control of PDEs under Uncertainty: An Introduction with Application to Optimal Shape Design of Structures*, Springer Briefs in Mathematics, Springer Cham, 2018.

[28] R. Mifflin, *Semismooth and semiconvex functions in constrained optimization*, SIAM J. Control Optim., 15 (1977), pp. 959–972.

[29] J. W. Pearson and A. J. Wathen, *A new approximation of the Schur complement in preconditioners for PDE-constrained optimization*, Numer. Linear Algebra Appl., 19 (2012), pp. 816–829.

[30] L. Qi and J. Sun, *A nonsmooth version of Newton's method*, Math. Program., 58 (1993), pp. 353–367.

[31] R. T. Rockafellar, *Convex analysis*, vol. No. 28 of Princeton Mathematical Series, Princeton University Press, Princeton, NJ, 1970.

[32] E. Rosseel and G. N. Wells, *Optimal control with stochastic PDE constraints and uncertain controls*, Comput. Methods Appl. Mech. Engrg., 213–216 (2012), pp. 152–167.

[33] W. Shen, T. Sun, B. Gong, and W. Liu, *Stochastic Galerkin method for constrained optimal control problem governed by an elliptic integro-differential PDE with stochastic coefficients*, Int. J. Numer. Anal. Model., 12 (2015), pp. 593–616.

[34] H. Tiesler, R. M. Kirby, D. Xiu, and T. Preusser, *Stochastic collocation for optimal control problems with stochastic PDE constraints*, SIAM J. Control Optim., 50 (2012), pp. 2659–2682.

[35] F. Tröltzsch, *Optimal Control of Partial Differential Equations: Theory, Methods and Applications*, American Mathematical Society, Providence, Rhode Island, 2010.

[36] M. Ulbrich, *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*, vol. 11 of MOS-SIAM Series on Optimization, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2011.

[37] H. Wang, D. Wu, C. Yu, and K. L. Teo, *A duality-based approach for linear parabolic optimal control problems*, Optimal Control Appl. Methods, 45 (2024), pp. 1140–1165, https://doi.org/10.1002/oca.3094, https://doi.org/10.1002/oca.3094.

[38] M. Winkler and H. Yücel, *A stochastic Galerkin method for optimal Dirichlet boundary control problems with uncertain data*, 2025, https://arxiv.org/abs/2506.11479.

[39] D. Xiu, *Numerical Methods for Stochastic Computations: A Spectral Method Approach*, Princeton University Press, 2010.

[40] D. Xiu and G. E. Karniadakis, *Modeling uncertainty in flow simulations via generalized polynomial chaos*, Journal of computational physics, 187 (2003), pp. 137–167.