Talk 3: Concerning the Monte Carlo finite element method

Seminar: Data based mathematical modelling
Johannes Gutenberg-University Mainz
Institute of Mathematics

Speaker: Colin Schwanke
Matriculation-number: 2731242
Supervisor: Prof. Dr. Mária Lukáčová

winter semester 22/23

1 Introduction

The MCFEM is a procedure, wich consists out of two processes. The finite element method discretizes a continous problem. The Monte Carlo simulation observes the sample-means of the outputs of observed experiments with diffrent inputs. The law of large numbers assures, that those sample-means converge to the value of desire.

We want to use these methods to approximate the expected value and the variance of the solution of the predetermined boundary value problem (BVP):

$$-\nabla(a(\boldsymbol{x})\nabla u(\boldsymbol{x})) = -\sum_{j=1}^{2} \frac{\partial}{\partial x_j}\left(a(\boldsymbol{x})\frac{\partial u(\boldsymbol{x})}{\partial x_j}\right) = f(\boldsymbol{x}), \quad \boldsymbol{x} \in D \tag{1.1}$$

$$u(\boldsymbol{x}) = g(\boldsymbol{x}), \quad \boldsymbol{x} \in \partial D. \tag{1.2}$$

Expressed informally, $\mathbb{E}$ is the true value of the arithmetic mean and therefore fixes the location of the distribution of our solution. The variance than gives us the quadratic deviation, so that we have an idea about how far the solution can differ from its expected value. That is the reason for our interest in approximating $\mathbb{E}[u]$ and $Var[u]$.

The PDE (1.1) provides a model for diffusion in a steady state. Therefore the problem we look at is time independent and imposes the following additional conditions. D needs to be a bounded subset in $\mathbb{R}^2$ with a piecewise smooth boundary ∂D. The diffucion coefficient a only takes positive values, $g : \partial D \to \mathbb{R}$ provides the the boundary data and for simplicity we assume, that the source-term satisfies $f = 1$.

The initial problem assumes, that a, f and u are deterministic functions wich assigne values in D to values in $\mathbb{R}$. To give meaning to our intention, approximating values that are originally defined for random variables, we have to make some adjustments first. a, f and u now become real valued second order random fields.

Definition 1.1. ***(second-order random field)***
A real valued second-order random field, where $D \subset \mathbb{R}^d$, is a set of real valued random variables $\{u(\boldsymbol{x}), \boldsymbol{x} \in D\}$ on a probability space $(\Omega, \Sigma, \mathbb{P})$ with the additional condition, that $u(\boldsymbol{x}) : \Omega \to \mathbb{R} \in L^2(\Omega)$ for every $\boldsymbol{x} \in D$ (notice $u : D \times \Omega \to \mathbb{R}$).

The notion of the realisation denotes one possible outcome of such a random field.

Definition 1.2. ***(realisation)***
For a fixed $\omega \in \Omega$, the realisation of a random field is a function $\zeta : D \to \mathbb{R}$, $\zeta(\boldsymbol{x}) := u(\boldsymbol{x}, \omega)$.

Demanding that the random fields lie in $L^2(\Omega)$ is necessary, so that the first two moments of u are finite and therefore $\mathbb{E}$ and Var are well-defined. With these adjustements done, the initial problem transforms into

$$-\nabla(a(\boldsymbol{x}, \omega)\nabla u(\boldsymbol{x}, \omega)) = f(\boldsymbol{x}, \omega), \quad \boldsymbol{x} \in D \tag{1.3}$$

$$u(\boldsymbol{x}, \omega) = g(\boldsymbol{x}), \quad \boldsymbol{x} \in \partial D. \tag{1.4}$$

We know that under certain conditions the associated deterministic variational problem is well-posed (see [1], p.372) for realisations $a(\cdot,\omega)$ and $f(\cdot,\omega)$. Therefore we know, that $u(\cdot,\omega) \in H^1_g(D) := \{w \in H^1(D) : \gamma w = g\}$ is well-defined with the trace operator $\gamma : H^1(D) \to L^2(\partial D)$ wich maps functions from the sobolev space $H^1(D)$ themself with the restriction on ∂D. The FEM now helps us to approximate individual realisations of the solution.

2 The finite element method in $\mathbb{R}^2$

For the discretization we subdivide the domain into n_e triangles, wich we set to $D = (0,1) \times (0,1)$. Triangles are ideal, because there is a variety of arrangement-options if the whole domain should be covered. $\mathcal{T}_h := \{\Delta_1, \ldots, \Delta_{n_e}\}$ is the set of all these triangles, where h_k is the longest edge of Δ_k, $\bar{D} = \bigcup_{k=1}^{n_e} \bar{\Delta}_k$ and h as the longest edge of all triangles $h := \max_k h_k$ determines the meshsize. J is the number of inner nodes and J_b the number of boundary nodes.

To observe if the mistake of our method is getting smaller while the meshsize h converges to zero, we need a sequence of increasingly smaller grids. Also we want to avoid difficulties while computing, which arise by having triangles with very sharp angles. Therefore the term of shape-regularity is introduced.

Definition 2.1. ***(shape-regularity)***
A sequence of grids stays regular in shape, if for every element $\mathcal{T}_h$ there is a constant $\eta > 0$, which is independent of h and satisfies

$$\frac{\rho_k}{h_k} \geq \eta \quad ,\forall \Delta_k \in \mathcal{T}_h$$

where ρ_k is the radius of the circle inside of Δ_k with maximum surface.

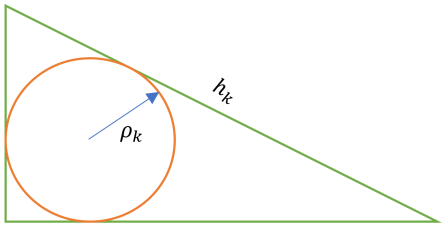

Figure 1: rectangular triangles always do the trick

Of course a greater mesh-density goes along with more computing effort but also with a higher accuracy. Hence comparing different subdivisions of the domain and their influence of the computing time while h approaches zero is an interesting field that surely is worth delving into.
Now we choose our grid-refinement-sequence, for the subdivision of the domain D. One step of refinement is shown underneath in Figure 2.

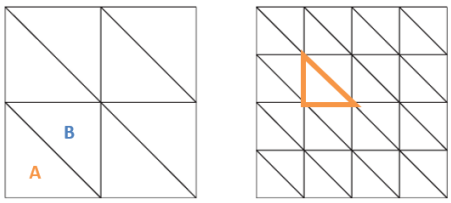

Figure 2: two elements from the meshsequence of our choice

In the implementation, all elements and their nodes have to be numbered, from left to right and from bottom to top. For the triangles we first number those of type A and after that, those of type B (see Figure 2). In the figure above the tenth triangle is highlighted with the global nodes 12, 13 and 17. Locally the nodes are numbered as well, starting with the one that is sitting on the right angle and then anticlockwise. Of course all the real coordinates have to be stored as well.

After subdividing the domain, we now contruct an approximation of u over our gridpattern with a set $V^h \subset H_0^1(D)$ of polynomials which are piecewise defined on the elements Δ_k, while $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_J\}$ are the inner nodes, wich should guarantee that $V^h \subset \mathcal{C}(\bar{D})$, so that $v|_{\partial D}$ is well-defined for every $v \in V^h$.

$$V^h := \left\{ v \in \mathcal{C}(\bar{D}) : v|_{\partial D} = 0 \text{ and } v|_{\Delta_k} \in \mathbb{P}(\Delta_k) \ , \forall \Delta_k \in \mathcal{T}_h \right\} = span\{phi_1(\boldsymbol{x}), \ldots, \phi_J(\boldsymbol{x})big\}$$

.
The variable r determindes the maximum degree of the basisfunctions, wich have to satisfy $\phi_j(x_i) = \delta_{ij}$, where δ_{ij} is the Kronecker-Delta function. The number of nodes on a Element and therefore J depends on r and n_e. J gets bigger, when when r gets bigger and h gets smaller. That means, that possibly the number of nodes have to be adjusted so that there are $r+1$ nodes on every edge to ensure getting a globally continuous approximation.

As said before, in the case of inhomogeneous boundary conditions an adjustement has to be made in regard to the room V. Because we just have to handle homogenous boundary data in the following examples, we refer to [1] p. 70 ff. here for the curious reader. The FE-approximation for the solution then has the form

$$u_h(\boldsymbol{x}) = \sum_{i=1}^{J} u_i \phi_i(\boldsymbol{x})$$

and it has to satisfy the following equations

$$\sum_{i=1}^{J} u_i B(\phi_i, \phi_j) = \ell(\phi_j) \quad , \ j = 1, \ldots, J.$$

which can be written as a matrix vector equation:

$$Mu = b$$

where

$$m_{ij} := B(\phi_i, \phi_j) = \int_D \tilde{a}_r(\boldsymbol{x}) \nabla \phi_i(\boldsymbol{x}) \cdot \nabla \phi_j(\boldsymbol{x}) d\boldsymbol{x} \in \mathbb{R} \quad , \ i, j = 1, \ldots, J$$

and

$$b_i := \ell(\phi_i) = \int_D f(\boldsymbol{x})\phi_i(\boldsymbol{x})d\boldsymbol{x} \quad , \ i = 1, \ldots, J.$$

The entiries of the so called Galerkin matrix M differ from zero, if the supports of ϕ_i and ϕ_j intersect. Is $r = 1$, this happens only if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are on the same edge of a triangle. Therefor M is thinly occupied or sparse. The integrals of m_{ij} and b_i correspond to the sum of the integrals above the individual triangles.

Before we come to the fun part, there is one thing, that has to be considered. Often in practice the diffusion coefficient a is not known, and we have to content with samples of an approximation of it. Therefore it is mportant for the whole procedure, that we know at least some exact samples at the gridpoints of $\mathcal{T}_h$, so that we can take means on every element

$$\tilde{a}(\boldsymbol{x}, \omega)|_{\Delta_k} := \frac{1}{3}\sum_{j=1}^{3} a(v_j^k, \omega) \quad , \forall \ \Delta_k \in \mathbb{T}_h$$

.
With the knowledge of a procedure that enables us to approximate u we can go on and make a Monte Carlo Simulation by generating multiple solutions for different diffusion coefficients which can be seen as the inputs of our experiments. As a result the mean of those solutions then provides an approximation for the expected value. Analogously one obtains the variance afterwards.

Let $\tilde{a}_r := \tilde{a}(\cdot, \omega_r), \ r = 1, \ldots, Q$ be *iid* samples of an approximation of the diffusion coefficient. The associated *iid* samples $\tilde{u}_h^r := \tilde{u}_h^r(\boldsymbol{x}, \omega_r)$ of the FE-solution are obtained then by solving the associated variational problem

$$\int_D \tilde{a}(\boldsymbol{x}, \omega_r) \cdot \nabla \tilde{u}_h(\boldsymbol{x}, \omega_r) \cdot \nabla v(\boldsymbol{x}) d\boldsymbol{x} = \int_D f(\boldsymbol{x}) v(\boldsymbol{x}) d\boldsymbol{x} \quad , \forall \ v \in V^h.$$

Particularly important is, that

$$\mu_{Q,h}(\boldsymbol{x}) := \frac{1}{Q}\sum_{r=1}^{Q} \tilde{u}_h^r(\boldsymbol{x})$$

and

$$\sigma_{Q,h}(\boldsymbol{x}) := \frac{1}{Q-1}\left(\sum_{r=1}^{Q} \tilde{u}_h^r(\boldsymbol{x})^2 - Q \cdot \mu_{Q,h}(\boldsymbol{x})^2\right)$$

converge $\mathbb{P} - a.s.$ to $\mathbb{E}[\tilde{u}(\boldsymbol{x})]$ respectively $Var(\tilde{u}(\boldsymbol{x}))$. Now we are ready to take a look at two different esamples:

3 Example for a onedimensional application

Let $D = (0, 1), g = 0$ and $f = 1$. Therefore the PDE is given by

$$\frac{d}{dx}\left(a(x)\frac{d}{dx}u(x)\right) = 1 \quad , \; x \in D$$

with the diffusion coefficient

$$a(x, \omega) = \mu + \sum_{k=1}^{P} \frac{\sigma}{k^2\pi^2} \cos(\pi k x)\xi_k(\omega) \quad , \; \xi_k \sim U(-1, 1)\boldsymbol{iid}.$$

$\mu = 1$ and $\sigma = 4$ are the expected value and the variance of the diffusion coefficient, which is a random variable for a fixed $x \in D$. For a mesh with 512 elements and $P = 10$ one gets the following results with the code in the appendix:

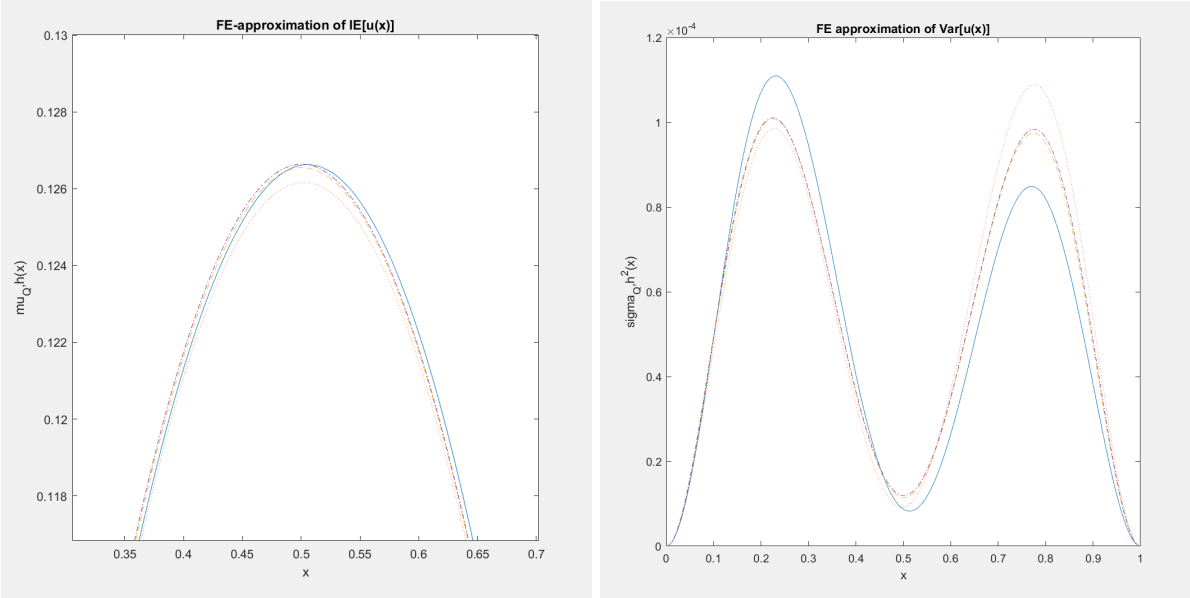

Figure 3: $Q = 10$ (solid & blue), $Q = 10^2$ (dotted & orange), $Q = 10^3$ (dashed & yellow), and $Q = 10^4$ (dash-dotted & purple).

4 Error analysis

After just applying the method, one might ask whether it is very accurate at all and therefor take a look at the arising error. The resulting error, consists of the statistical error E_{MC} which depends on Q and the discretization error $E_{h,a}$ and for both of them, the asymptotic behavior can be determined (see [1], p.385-386).

$$||\mathbb{E}[u] - \mu_{Q,h}||_{H_0^1(D)} \leq E_{h,a} + E_{MC} = ||\mathbb{E}[u] - \mathbb{E}[\tilde{u}_h]||_{H_0^1(D)} + ||\mathbb{E}[\tilde{u}_h] - \mu_{Q,h}||_{H_0^1(D)} = \mathcal{O}(h) + \mathcal{O}(Q^{-1/2})$$

We now know, that technically the error can be kept small but in practice this is expensive. Getting a sample of the solution means solving a PDE and so the rate of convergence concerning Q is already not very attractive. More so, to implement the MCFEM for the twodimensional case costs $\mathcal{O}(\epsilon^{-4})$ if we presuppose that we want to approximate $\mathbb{E}[u]$ in the $H_0^1(D)$ norm to an accuracy of ϵ.

5 The variational formulation on $D \times \Omega$

Finally we derive another variational formulation of (1.3)-(1.4) on $D \times \Omega$ and seek for weak solutions $u : D \times \Omega \to \mathbb{R}$. In general we do not search for a solution in the classical sense, because such a function only exists, if the diffusion coefficient is continuously differentiable over $\bar{D}$, wich is often not the case.
Therefore we want to convice ourselfs, that there is at least a well-defined so called "weak" solution for the adjusted initial BVP. In other words we want to find a function $u : \bar{D} \times \Omega \to \mathbb{R}$ in $L^2(\Omega, H_0^1(D))$ of lesser smoothness that satisfies (1.3)-(1.4) $\mathbb{P} - a.s.$, so that the individual realisations lie in the solution space of the determenistic BVP.
That is why we define the weak solution by formulating a variational problem by multiplying the PDE with a testfunction $v \in V := L^2(\Omega, \mathcal{F}, H_0^1(D))$, integrating both sides over D and taking the expected value. Notice that

$$||v||_V := \mathbb{E}\big[|v|^2_{H^1(D)}\big]^{1/2} = \mathbb{E}\bigg[\sum_{|\alpha|=1} ||\mathcal{D}^\alpha v||^2_{L^2(D)}\bigg]^{1/2} = \mathbb{E}\bigg[\sum_{|\alpha|=1} \int_D |\mathcal{D}^\alpha v(\boldsymbol{x})|^2 d\boldsymbol{x}\bigg]^{1/2}$$

Definition 5.1. ***(weak solution)***
A weak solution of the BVP (1.3)-(1.4) *is a function* $u \in V$ *that satisfies*

$$B(u, v) = \ell(v) \quad , \forall v \in V \tag{5.1}$$

where

$$B(u, v) := \mathbb{E}\bigg[\int_D a(\boldsymbol{x}, \cdot)\nabla u(\boldsymbol{x}, \cdot) \cdot \nabla v(\boldsymbol{x}, \cdot) d\boldsymbol{x}\bigg]$$

is a bilinear form and

$$\ell(v) := \mathbb{E}\big[\langle f, v\rangle_{L^2(D)}\big] := \mathbb{E}\bigg[\int_D f(\boldsymbol{x}, \cdot)v(\boldsymbol{x}, \cdot) d\boldsymbol{x}\bigg]$$

If we have homogenous boundary conditions, g=0 and we do not have to distiguish between the solution-space for u and the testspace for v. Now we can prove the well-posedness of the variational problem.

Theorem 5.2. *If $f \in L^2(\Omega, L^2(D)), g = 0$ and*

$$0 < a_{min} \leq a(\boldsymbol{x}, \omega) \leq a_{min} < \infty \quad , a.e.\ in\ D \times \Omega$$

with $a \in L^\infty(\Omega, L^\infty(D))$ and $a_{min}, a_{max} \in \mathbb{R}$, then (5.1) possesses a unique solution $u \in V$ whose behavior changes continuously with the initial conditions. Furthermore, $u(\boldsymbol{x}, \cdot)$ is $\mathcal{G}$-measurable for some sub σ-algebra $\mathcal{G} \subset \mathcal{F}$, if $a(\boldsymbol{x}, \cdot)$ and $f(\boldsymbol{x}, \cdot)$ are $\mathcal{G}$-measurable.

Proof:
If all requirements of the Lax-Milgram Lemma (see [1], p.16) are valid, we know, that u in (5.1) is well-defined, so we check those. $B : V \times V \to \mathbb{R}$ needs to be bounded and therefore we use the Cauchy Schwarz inequality three times. First we use it for the normal scalar product, after a substitution for the product of integrals and last for the product of expected values:

$$\begin{aligned}
|B(u,v)| &= \mathbb{E}\Big[\int_D a(\boldsymbol{x}, \cdot)\nabla u(\boldsymbol{x}, \cdot) \cdot \nabla v(\boldsymbol{x}, \cdot) d\boldsymbol{x}\Big] \leq a_{max}\mathbb{E}\Big[\int_D ||\nabla u(\boldsymbol{x}, \cdot)||_2 \cdot ||\nabla v(\boldsymbol{x}, \cdot)||_2 d\boldsymbol{x}\Big] \\
&= a_{max}\mathbb{E}\Big[\int_D \gamma(\boldsymbol{x})\delta(\boldsymbol{x}) d\boldsymbol{x}\Big] \leq a_{max}\mathbb{E}\Big[\int_D \gamma(\boldsymbol{x}) d\boldsymbol{x} \int_D \delta(\boldsymbol{x}) d\boldsymbol{x}\Big] \\
&= a_{max}\mathbb{E}\big[|u|_{H^1(D)}|v|_{H^1(D)}\big] \leq a_{max}\mathbb{E}\big[|u|^2_{H^1(D)}\big]^{1/2}\mathbb{E}\big[|v|^2_{H^1(D)}\big]^{1/2} \\
&= a_{max}||u||_V||v||_V
\end{aligned}$$

for all $u, v \in V$.

With the same arguments but now for a_{min} we get the coercivity of B, i.e. $B(v, v) \geq a_{min}||v||^2_V$ and we just need the boundedness of ℓ to close the proof. Again we use Cauchy Schwarz:

$$\begin{aligned}
|\ell(v)| &= \mathbb{E}\Big[\int_D f(\boldsymbol{x}, \cdot)v(\boldsymbol{x}, \cdot) d\boldsymbol{x}\Big] \leq \mathbb{E}\Big[\int_D v(\boldsymbol{x}, \cdot) d\boldsymbol{x}\Big]\mathbb{E}\Big[\int_D f(\boldsymbol{x}, \cdot) d\boldsymbol{x}\Big] \\
&= ||f||_{L^2(\Omega,L^2(D))}||v||_{L^2(\Omega,L^2(D))}
\end{aligned}$$

Additionally the Poincaré inequality (see [1], p.14) provides an upper bound for every realisation

$$||v(\cdot, \omega)||_{L^2(D)} \leq C|v(\cdot, \omega)|_{H^1(D)} \quad , \forall\ \omega \in \Omega$$

Taking second moments gives us the boundedness of the linear functional ℓ

$$|\ell(v)| \leq C||f||_{L^2(\Omega,L^2(D))}||v||_V =: \beta||v||_V$$

with $\beta > 0$.

q.e.d.

Remark 5.3. (inhomogeneous boundary conditions)
If $g \neq 0$, the testspace V differs from the solutionspace $W := L^2(\Omega, H^1_g(D))$ with $|v|_W := \mathbb{E}\big[|v|^2_{H^1(D)}\big]^{1/2}$ that fits to the adjustment. The variational problem can be modified to an equivalent problem with homogeneous boundary conditions and so it is also well-posed under the same conditions when $B : W \times V \to \mathbb{R}$ and $g \in H^{1/2}(\partial D) := \gamma(H^1(D)) = \{\gamma w : w \in H^1(D)\}$ with the norm $||g||_{H^{1/2}(\partial D)} := inf\big\{||w||_{H^1(D)} : \gamma w = g, w \in H^1(D)\big\}$.

Appendix

```
%the parameters:


%number of samples
Q=[10 10^2 10^3 10^4];
%expected value and variance of the diffusion coefficient
mu=1;
sigma=4;
% number of summands to compute samples of a(x)
P=10;
%number of the finite elements respectively triangles
ne=512;
%source term
f=1;
%------------------------------------------------------------------------




%vectors that store the values of mu_Q,h(x) und sigma_Q,h^2(x) at the nodes
%of the grid

[mean_u, var_u]=oned_MC_FEM(512,sigma,mu,P,Q(1));
[mean_u1, var_u1]=oned_MC_FEM(512,sigma,mu,P,Q(2));
[mean_u2, var_u2]=oned_MC_FEM(512,sigma,mu,P,Q(3));
[mean_u3, var_u3]=oned_MC_FEM(512,sigma,mu,P,Q(4));

D=0:1/512:1;
%------------------------------------------------------------------------



%implementation of the finite element method
function [uh,A,b,K,M] = oned_linear_FEM(ne,p,q,f)

% constructing the onedimensional grid T_h
h=(1/ne); D=0:h:1; nvtx=length(D);
J=ne-1; elt2vert=[1:J+1;2:(J+2)]';


% declaration of the global matrix
K = sparse(nvtx,nvtx); M = sparse(nvtx,nvtx); b=zeros(nvtx,1);
```

```
% Computation of the element matrices
[Kks,Mks,bks]=get_elt_arrays(h,p,q,f,ne);

% arrange elemente arrays into global arrays
for row_no=1:2
    nrow=elt2vert(:,row_no);
    for col_no=1:2
        ncol=elt2vert(:,col_no);
        K=K+sparse(nrow,ncol,Kks(:,row_no,col_no),nvtx,nvtx);
        M=M+sparse(nrow,ncol,Mks(:,row_no,col_no),nvtx,nvtx);
    end
    b = b+sparse(nrow,1,bks(:,row_no),nvtx,1);
end

% homogeneous boundary conditions
K([1,end],:)=[]; K(:,[1,end])=[]; M([1,end],:)=[]; M(:,[1,end])=[];
A=K+M; b(1)=[]; b(end)=[];
%solving the linear system for the inner nodes
u_int=A\b; uh=[0;u_int;0];
plot(D,uh,'-'); title('FE N herungen von u(x)');
xlabel('x');ylabel('u_h(x)')
end
%-------------------------------------------------------------------------
```

```
%local computation of the integrals of the entries of A and b above the elements.
function [Kks,Mks,bks] = get_elt_arrays(h,p,q,f,ne)
Kks = zeros(ne,2,2); Mks=zeros(ne,2,2);
Kks(:,1,1)=(p./h); Kks(:,1,2)=-(p./h);
Kks(:,2,1)=-(p./h); Kks(:,2,2)=(p./h);
Mks(:,1,1)=(q.*h./3); Mks(:,1,2)=(q.*h./6);
Mks(:,2,1)=(q.*h./6); Mks(:,2,2)=(q.*h./3);
bks=zeros(ne,2); bks(:,1)= f.*(h./2); bks(:,2) = f.*(h./2);
end

%-------------------------------------------------------------------------
```

```
%piecewise constant approximation of the realisations of a(x) by P iid samplesxi_k
function [mean_u,var_u] = oned_MC_FEM(ne,sigma,mu,P,Q)
h=1/ne; x=[(h/2):h:(1-h/2)]';
sum_us=zeros(ne+1,1); sum_sq=zeros(ne+1,1);
for j=1:Q
    xi=-1+2.*rand(P,1); a=mu.*ones(ne,1);
```

```
    for i=1:P
        a=a+sigma.*((i.*pi).^(-2)).*cos(pi.*i.*x).*xi(i);
    end
    [u,A,b]=oned_linear_FEM(ne,a,zeros(ne,1),ones(ne,1)); hold on;
    sum_us=sum_us+u; sum_sq=sum_sq+(u.^2);
end
mean_u=sum_us./Q;
var_u=(1/(Q-1)).*(sum_sq-(sum_us.^2./Q));
end
```

References

[1]

J. Lord/C. E. Powell/T. Shardlow (2014), An Introduction to Computational Stochastic PDEs, Cambridge Texts in Applied Mathematics, Cambridge: Cambridge University Press.