Numerical methods in mathematical finance
Part 2

Tobias Jahnke

Karlsruher Institut für Technologie
Fakultät für Mathematik
Institut für Angewandte und Numerische Mathematik

tobias.jahnke@kit.edu

Version: July 27, 2017

Preface

These notes are the basis of my lecture *Numerical methods in mathematical finance 2* given at Karlsruhe Institute of Technology in the summer terms 2011, 2013, 2015, and 2017. The purpose of this notes is to help students who have missed parts of the course to fill these gaps, and to provide a service for those students who can concentrate better if they do not have to copy what I write on the blackboard.

It is *not* the purpose of these notes, however, to replace the lecture itself, or to write a text which could compete with the many excellent books about the subject. This is why the style of presentation is rather sketchy. As a rule of thumb, one could say that these notes only cover what I *write* during the lecture, but not everything I *say*.

There are probably still many typos and possibly also other mistakes. Of course, I will try to correct any mistake I find as soon as possible, but the reader should be aware of the fact that he or she cannot trust these notes.

I thank Andreas Arnold, Axel Heuser, and Lisa Lungershausen for typing preliminary versions of parts of these lecture notes, and Michael Baumann, Fabian Castelli, Johannes Eilinghoff, Matthias Henze, and Christoph Tiemann for finding a lot of typos. All remaining misprints and errors are, of course, my fault.

Karlsruhe, summer term 2017,
Tobias Jahnke

Contents

Chapter 1

Multilevel Monte Carlo methods

1.1 Standard Monte Carlo approach

Consider a European option on a single underlying. Let $S(t)$ be the value of the underlying, $V(t,S)$ value of the option with maturity $T>0$.
Assume that under the risk-neutral measure $\mathbb{Q}$, $S(t)$ is the solution of the SDE

$$\begin{aligned} dS(t) &= f(t,S)dt + g(t,S)dW(t) \qquad\qquad t\in[0,T] \\ S(0) &= S_0 \end{aligned} \tag{1.1}$$

Example: For geometric Brownian motion we choose $f(t,S)=rS$ and $g(t,S)=\sigma S$ with risk-free interest rate $r>0$ and volatility $\sigma>0$.

Let $\psi(S)$ be the discounted payoff function of the option: $\psi(S)=e^{-rT}(S-K)^+$ for a call, $\psi(S)=e^{-rT}(K-S)^+$ for a put. Notation: $(S-K)^+ := \max\{S-K,0\}$.

Goal: Compute the price of the option at time $t=0$. In the first part of the lecture, we have seen that

$$V(0,S_0)=\mathbb{E}_{\mathbb{Q}}\Big[\psi\big(S(T)\big)\Big].$$

Standard Monte Carlo method:

- Choose $N\in\mathbb{N}$, let $\tau=T/N$ and $t_n=n\tau$ for $n=0,\dots,N$.
  Generate $m\in\mathbb{N}$ paths $t\mapsto W(t,\omega_j)$ of the Wiener process ($j=1,\dots,m$).
  For each path, compute approximations $S_n(w_j)\approx S(t_n,\omega_j)$ by solving (1.1) with a numerical method of weak order γ.
- Approximate the risk-neutral expectation:

$$\mathbb{E}_{\mathbb{Q}}\Big[\psi\big(S(T)\big)\Big] \approx \frac{1}{m}\sum_{j=1}^{m}\psi\big(S(T,\omega_j)\big) \approx \frac{1}{m}\sum_{j=1}^{m}\psi\big(S_N(\omega_j)\big)$$

Two sources of error:

- Estimate the expectation from finitely many samples.
- Approximate the exact $S(T, \omega_j)$ by a numerical method.

Both errors are measured by the mean-square-error.

Definition 1.1.1 (Mean-square-error) *Let $\hat{\theta}$ be an estimator for an unknown (deterministic) quantity θ. Then, the mean-square-error of $\hat{\theta}$ is*

$$MSE(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right] = \mathbb{V}(\hat{\theta}) + \mathbb{E}(\hat{\theta} - \theta)^2.$$

Notation: $\mathbb{E}(X)^2 := \big(\mathbb{E}(X)\big)^2 \neq \mathbb{E}(X^2)$.
Applying this with $\mathbb{E} = \mathbb{E}_{\mathbb{Q}}$ and

$$\theta = \mathbb{E}\Big(\psi\big(S(T)\big)\Big), \qquad \hat{\theta} = \frac{1}{m}\sum_{j=1}^{m} \psi\big(S_N(\omega_j)\big)$$

yields

$$\mathrm{MSE}(\hat{\theta}) \leq \frac{C}{m} + C\tau^{2\gamma} \qquad \text{and} \qquad \sqrt{\mathrm{MSE}(\hat{\theta})} \sim C\sqrt{m^{-1} + \tau^{2\gamma}}.$$

Slow convergence with respect to m!

Example: Euler-Maruyama method. If $\varepsilon > 0$ is a given error tolerance, then

$$\mathrm{MSE}(\hat{\theta}) = \varepsilon^2 \leq \frac{C}{m} + C\tau^2 \qquad \Longleftrightarrow \qquad m = \mathcal{O}\big(\varepsilon^{-2}\big) \text{ and } \tau = \mathcal{O}(\varepsilon)\,.$$

Since $\tau = T/N$, we have to compute $m = \mathcal{O}(\varepsilon^{-2})$ simulations with $N = \mathcal{O}(\varepsilon^{-1})$ time-steps. Hence, the total numerical work (= total number of time-steps) is $\mathcal{O}(\varepsilon^{-3})$.

Is there a more efficient method?

- Using the Milstein method instead of the Euler-Maruyama method does not give any improvement.
- If a method with weak order $\gamma = 2$ is used, then we need $m = \mathcal{O}(\varepsilon^{-2})$, $\tau = \mathcal{O}(\sqrt{\varepsilon})$ and $N = \mathcal{O}\big(\varepsilon^{-1/2}\big)$. In this case, the total number of time-steps is $\mathcal{O}(\varepsilon^{-2.5})$, but each time-step is more costly.
- Variance reduction may give an improvement
- The multilevel Monte Carlo method (see below) only needs $\mathcal{O}(\varepsilon^{-2}(\ln \varepsilon)^2)$ time-steps of the Euler-Maruyama method. This method was proposed by Giles in [Gil08b].

1.2 Multilevel Monte Carlo: The simplest case

(a) Ansatz and notation

Idea: Compute approximations with Euler-Maruyama and **different step-sizes** $\tau_\ell = M^{-\ell}T$ where $M \in \mathbb{N}$ and $\ell = 0, 1, \dots, L$ for some $L \in \mathbb{N}$.

Notation:

- Let $P = \psi\big(S(T)\big)$ be the (discounted) payoff of a European option. Hence, $\mathbb{E}(P)$ is the value of the option at time $t = 0$.
- Let $\widehat{S}_{\ell,M^\ell} \approx S(T)$ be the approximation of the stock price at maturity computed with M^ℓ Euler-Maruyama steps and step-size τ_ℓ.
- Let $\widehat{P}_\ell = \psi(\widehat{S}_{\ell,M^\ell})$ be the corresponding (discounted) payoff.
- The superscript "$\dots^{(\ell,i)}$" means that the i-th path $t \mapsto W(t,\omega_{\ell,i})$ of the Wiener process is considered. For example, $\widehat{P}_\ell^{(\ell,i)} = \psi(\widehat{S}_{\ell,M^\ell}^{(\ell,i)})$ where $\widehat{S}_{\ell,M^\ell}^{(\ell,i)}$ is the approximation corresponding to the Wiener path $t \mapsto W(t,\omega_{\ell,i})$. For every ℓ, a new set of paths is generated, i.e. for $\ell \neq \tilde{\ell}$ the paths $t \mapsto W(t,\omega_{\ell,i})$ and $t \mapsto W(t,\omega_{\tilde{\ell},i})$ are independent.

Idea: Represent $\mathbb{E}[\widehat{P}_L]$ by a telescoping sum

$$\mathbb{E}[\widehat{P}_L] = \mathbb{E}[\widehat{P}_0] + \sum_{\ell=1}^{L} \mathbb{E}[\widehat{P}_\ell - \widehat{P}_{\ell-1}]$$

and simulate each term on the right-hand side in a near-optimal way.

Define the following estimators:

$$\widehat{Y}_0 := \frac{1}{m_0}\sum_{i=1}^{m_0} \widehat{P}_0^{(0,i)} \qquad \approx \mathbb{E}[\widehat{P}_0] \qquad (m_0 \text{ samples, step-size } \tau_0)$$

$$\widehat{Y}_\ell := \frac{1}{m_\ell}\sum_{i=1}^{m_\ell} \Big(\widehat{P}_\ell^{(\ell,i)} - \widehat{P}_{\ell-1}^{(\ell,i)}\Big) \approx \mathbb{E}[\widehat{P}_\ell - \widehat{P}_{\ell-1}] \quad (m_\ell \text{ samples, step-size } \tau_\ell, \tau_{\ell-1})$$

$$\widehat{Y} := \sum_{\ell=0}^{L} \widehat{Y}_\ell \qquad \approx \mathbb{E}[\widehat{P}_L] \approx \mathbb{E}[P] \quad \text{(multilevel estimator)}$$

Important: $\widehat{P}_\ell^{(\ell,i)}$ and $\widehat{P}_{\ell-1}^{(\ell,i)}$ are computed with two different step-sizes, but with **the same path of the Wiener process**.

We will show that L and m_ℓ can be chosen in such a way that only $\mathcal{O}(\varepsilon^{-2}(\ln\varepsilon)^2)$ time-steps are required for $\mathrm{MSE}(\widehat{Y}) = \mathcal{O}(\varepsilon^2)$, i.e. for an accuracy of $\mathcal{O}(\varepsilon)$.

(b) Variance of the estimator

For $\ell = 1, \dots, L$, the variance of $\widehat{Y}_\ell$ is

$$\mathbb{V}[\widehat{Y}_\ell] = \mathbb{V}\left[\frac{1}{m_\ell}\sum_{i=1}^{m_\ell}\left(\widehat{P}_\ell^{(\ell,i)} - \widehat{P}_{\ell-1}^{(\ell,i)}\right)\right] = \frac{1}{m_\ell^2}\sum_{i=1}^{m_\ell}\mathbb{V}\left[\widehat{P}_\ell^{(\ell,i)} - \widehat{P}_{\ell-1}^{(\ell,i)}\right] = \frac{\mathbb{V}_\ell}{m_\ell}$$

where $\mathbb{V}_\ell := \mathbb{V}\left[\widehat{P}_\ell^{(\ell,i)} - \widehat{P}_{\ell-1}^{(\ell,i)}\right]$ is the variance of a single path (same variance for every path). For $\widehat{Y}_0$ we obtain $\mathbb{V}[\widehat{Y}_0] = \mathbb{V}_0/m_0$ with $\mathbb{V}_0 = \mathbb{V}\left[\widehat{P}_0^{(0,i)}\right]$. Since paths for different values of ℓ are independent, the variance of $\widehat{Y}$ is

$$\mathbb{V}[\widehat{Y}] = \sum_{\ell=0}^{L}\mathbb{V}[\widehat{Y}_\ell] = \sum_{\ell=0}^{L}\frac{\mathbb{V}_\ell}{m_\ell} \tag{1.2}$$

The total numerical work (number of time-steps) is now

$$\mathcal{O}\left(\sum_{\ell=0}^{L}\frac{m_\ell}{\tau_\ell}\right)$$

because on the ℓ-th level, we compute m_ℓ samples with step-size τ_ℓ.

Question: How to choose m_ℓ?

Goal: Minimize the variance $\mathbb{V}[\widehat{Y}]$ under the condition that

$$\sum_{\ell=0}^{L}\frac{m_\ell}{\tau_\ell} = w$$

for a given work $w > 0$. This leads to the constrained minimization problem

$$F(x) := \sum_{\ell=0}^{L}\frac{\mathbb{V}_\ell}{x_\ell} = \min_x! \qquad\qquad G(x) := \sum_{\ell=0}^{L}\frac{x_\ell}{\tau_\ell} - w = 0.$$

If $x = (x_0, \dots, x_L)^T \in \mathbb{R}^{L+1}$ is real-valued, then every solution is also a solution of the unconstrained minimization problem

$$\widetilde{F}(x, \lambda) := F(x) + \lambda G(x) = \min_{x,\lambda}!$$

with Lagrange multiplier $\lambda \in \mathbb{R}$. At the minimum, we have $\nabla\widetilde{F}(x, \lambda) = 0$ and in particular

$$0 = \partial_{x_k}\widetilde{F}(x, \lambda) = -\frac{\mathbb{V}_k}{x_k^2} + \frac{\lambda}{\tau_k}$$

which yields

$$x_k = \sqrt{\frac{\mathbb{V}_k \tau_k}{\lambda}} = \mathcal{O}\Big(\sqrt{\mathbb{V}_k \tau_k}\Big)\,.$$

Although $m_\ell \in \mathbb{N}$ is *not* real-valued, we can conclude that for a fixed numerical work, the variance (1.2) is minimized if

$$m_\ell = \mathcal{O}\Big(\sqrt{\mathbb{V}_\ell \tau_\ell}\Big)\,. \tag{1.3}$$

(c) Variance of $\widehat{P}_\ell - P$ and $\widehat{P}_\ell - \widehat{P}_{\ell-1}$

Now suppose that $L = \infty$ and consider the limit $\ell \to \infty$. We know that

$$\mathbb{E}[\widehat{P}_\ell - P] = \mathcal{O}(\tau_\ell) \qquad \text{(weak convergence)} \tag{1.4}$$

$$\mathbb{E}\big[|\widehat{S}_{\ell,M^\ell} - S(T)|^2\big] = \mathcal{O}(\tau_\ell) \qquad \text{(strong convergence)} \tag{1.5}$$

(cf. chapter 5 in part I).

Estimate $\mathbb{V}[\widehat{P}_\ell - P]$: The (discounted) payoff functions of European puts and calls are Lipschitz-continuous:

$$|\widehat{P}_\ell - P| = |\psi(\widehat{S}_{\ell,M^\ell}) - \psi(S(T))| \le c|\widehat{S}_{\ell,M^\ell} - S(T)|.$$

Hence, it follows that

$$\begin{aligned}\mathbb{V}[\widehat{P}_\ell - P] &= \mathbb{E}[(\widehat{P}_\ell - P)^2] - \mathbb{E}[\widehat{P}_\ell - P]^2 \le \mathbb{E}\big[(\widehat{P}_\ell - P)^2\big] \\ &\le c^2 \mathbb{E}\big[|\widehat{S}_{\ell,M^\ell} - S(T)|^2\big] \overset{(1.5)}{=} \mathcal{O}(\tau_\ell)\,.\end{aligned} \tag{1.6}$$

Estimate $\mathbb{V}[\widehat{P}_\ell - \widehat{P}_{\ell-1}]$: For all random variables $\mathcal{X}$ and $\mathcal{Y}$ with expectation and finite variance, the Cauchy-Schwartz inequality yields

$$\begin{aligned}\mathbb{V}(\mathcal{X} \pm \mathcal{Y}) &= \mathbb{V}(\mathcal{X}) \pm 2\mathbb{E}\left[\big(\mathcal{X} - \mathbb{E}(\mathcal{X})\big)\big(\mathcal{Y} - \mathbb{E}(\mathcal{Y})\big)\right] + \mathbb{V}(\mathcal{Y}) \\ &\le \mathbb{V}(\mathcal{X}) + 2\sqrt{\mathbb{V}(\mathcal{X})\mathbb{V}(\mathcal{Y})} + \mathbb{V}(\mathcal{Y}) \\ &= \Big(\sqrt{\mathbb{V}(\mathcal{X})} + \sqrt{\mathbb{V}(\mathcal{Y})}\Big)^2\end{aligned}$$

(Minkowski's inequality). Hence, it follows from (1.6) that

$$\begin{aligned}\mathbb{V}[\widehat{P}_\ell - \widehat{P}_{\ell-1}] &= \mathbb{V}[(\widehat{P}_\ell - P) - (\widehat{P}_{\ell-1} - P)] \\ &\le \Big(\sqrt{\mathbb{V}[\widehat{P}_\ell - P]} + \sqrt{\mathbb{V}[\widehat{P}_{\ell-1} - P]}\Big)^2 \\ &= \mathcal{O}(\tau_\ell)\end{aligned} \tag{1.7}$$

because $\mathcal{O}(\tau_{\ell-1}) = \mathcal{O}(M\tau_\ell) = \mathcal{O}(\tau_\ell)$

(d) MSE and total work

Goal: Choose m_ℓ and L in such a way that

$$\mathrm{MSE}(\hat{Y}) = \mathbb{V}\big(\hat{Y}\big) + \mathbb{E}\big(\hat{Y} - \mathbb{E}(P)\big)^2 = \mathcal{O}\big(\varepsilon^2\big)$$

and that the numerical work is as small as possible.
Since $\mathbb{V}_\ell = \mathbb{V}\left[\widehat{P}_\ell^{(\ell,i)} - \widehat{P}_{\ell-1}^{(\ell,i)}\right]$ by definition, (1.7) implies $\mathbb{V}_\ell = \mathcal{O}(\tau_\ell)$ and, by (1.3), $m_\ell = \mathcal{O}\big(\sqrt{\mathbb{V}_\ell \tau_\ell}\big) = \mathcal{O}(\tau_\ell)$ for the optimal m_ℓ. If we choose

$$m_\ell = \mathcal{O}\big(\varepsilon^{-2} L \tau_\ell\big)\,, \tag{1.8}$$

then it follows from (1.2) that

$$\mathbb{V}[\widehat{Y}] = \sum_{\ell=0}^{L} \frac{\mathbb{V}_\ell}{m_\ell} = \sum_{\ell=0}^{L} \mathcal{O}\big((\varepsilon^{-2} L \tau_\ell)^{-1} \tau_\ell\big) = \frac{1}{L} \sum_{\ell=0}^{L} \mathcal{O}\big(\varepsilon^2\big) = \mathcal{O}\big(\varepsilon^2\big)\,. \tag{1.9}$$

If we choose

$$L = \frac{\ln \varepsilon^{-1}}{\ln M} + \mathcal{O}(1) \tag{1.10}$$

in the limit $\varepsilon \to 0$, then

$$\tau_L = M^{-L} T = \mathcal{O}(\varepsilon)\,, \tag{1.11}$$

because

$$\begin{aligned} L = \frac{\ln \varepsilon^{-1}}{\ln M} + C \quad &\Longrightarrow \quad L \ln M = \ln \varepsilon^{-1} + C \ln M \\ &\Longrightarrow \quad \ln(M^L) = \ln(\varepsilon^{-1} M^C) \\ &\Longrightarrow \quad M^L = \varepsilon^{-1} M^C \\ &\Longrightarrow \quad M^{-L} = M^{-C} \varepsilon = O(\varepsilon) \quad \Longrightarrow \quad M^{-L} T = O(\varepsilon). \end{aligned}$$

As a consequence, we obtain that the bias of $\widehat{Y}$ is

$$\begin{aligned} \mathbb{E}\left[\widehat{Y} - \mathbb{E}(P)\right] &= \sum_{\ell=0}^{L} \mathbb{E}[\widehat{Y}_\ell] - \mathbb{E}(P) \\ &= \underbrace{\frac{1}{m_0} \sum_{i=1}^{m_0} \mathbb{E}\left[\widehat{P}_0^{(0,i)}\right]}_{\ell=0} + \sum_{\ell=1}^{L} \frac{1}{m_\ell} \sum_{i=1}^{m_\ell} \mathbb{E}\left[\widehat{P}_\ell^{(\ell,i)} - \widehat{P}_{\ell-1}^{(\ell,i)}\right] - \mathbb{E}(P) \\ &= \mathbb{E}\left[\widehat{P}_0\right] + \sum_{\ell=1}^{L} \mathbb{E}\left[\widehat{P}_\ell - \widehat{P}_{\ell-1}\right] - \mathbb{E}(P) \\ &= \mathbb{E}[\widehat{P}_L] - \mathbb{E}(P) \\ &\overset{(1.4)}{=} \mathcal{O}(\tau_L) \overset{(1.11)}{=} \mathcal{O}(\varepsilon)\,. \end{aligned} \tag{1.12}$$

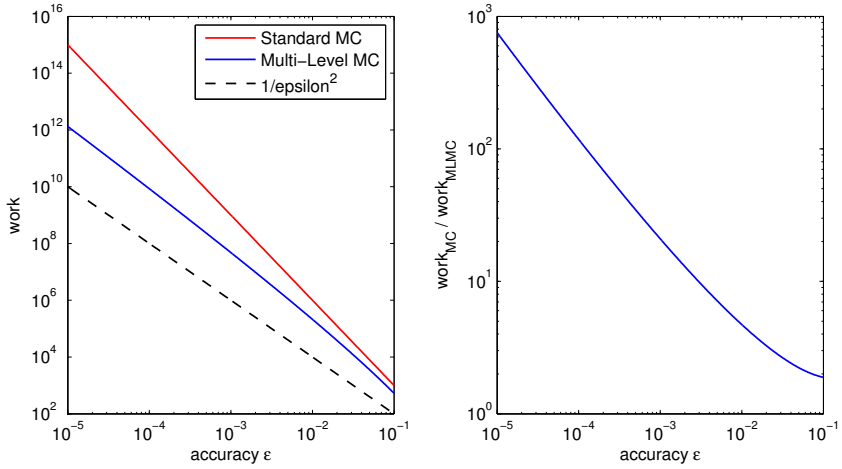


Figure 1.1: Numerical costs for standard and multilevel Monte Carlo.

Final result: The combined estimator has a mean-square-error

$$\mathrm{MSE}(\widehat{Y}) = \mathbb{E}\left[(\widehat{Y} - \mathbb{E}(P))^2\right] = \underbrace{\mathbb{V}[\widehat{Y}]}_{(1.9)} + \underbrace{\mathbb{E}\left[\widehat{Y} - \mathbb{E}(P)\right]^2}_{(1.12)} = \mathcal{O}\big(\varepsilon^2\big)\,,$$

but according to (1.8) and (1.10) the computational costs are only

$$\begin{aligned}\mathcal{O}\left(\sum_{\ell=0}^{L}\frac{m_\ell}{\tau_\ell}\right) &= \mathcal{O}\left(\sum_{\ell=0}^{L}(\varepsilon^{-2}L\tau_\ell)\tau_\ell^{-1}\right) = (L+1)\mathcal{O}\big(\varepsilon^{-2}L\big) = \mathcal{O}\big(\varepsilon^{-2}L^2\big)\\ &= \mathcal{O}\big(\varepsilon^{-2}(\ln\varepsilon)^2\big)\end{aligned}$$

instead of $\mathcal{O}(\varepsilon^{-3})$ for standard Monte Carlo. The efficiency gain is illustrated in Figure 1.2.

Remark. 1. After all, our goal is to approximate $\mathbb{E}(P) = \mathbb{E}\left(\psi\big(S(T)\big)\right)$. For European options, this is a "weak quantity", and if the standard Monte Carlo method is used, only the weak order of the Euler-Maruyama method matters. For the multilevel Monte Carlo method, this is different: both the weak and the strong order are important! The strong convergence (1.5) and the Lipschitz continuity of the payoff are required to prove the bound (1.6), which is then used to prove $\mathbb{V}[\widehat{P}_\ell - \widehat{P}_{\ell-1}] = \mathcal{O}(\tau_\ell)$ in (1.7).

2. The ansatz to represent $\mathbb{E}[\widehat{P}_L]$ by a telescoping sum is similar to (but different from) applying variance reduction by decomposition (method 1 in 6.2 in part I of the lecture).

1.3 The complexity theorem

So far, we have only considered European put or call options, and the price process has been approximated with the Euler-Maruyama method. Now the above result will be generalized.

Notation:

- Let the price process $S(t)$ be the solution of the SDE

$$dS(t) = f(t,S)dt + g(t,S)dW(t), \qquad t \in [0,T], \qquad S(0) = S_0. \tag{1.13}$$

- Let P be the payoff of an option. $P = P(t \mapsto S(t))$ is a functional of the solution and can depend on the entire path $t \mapsto S(t)$. This allows to consider, e.g., Asian, barrier or lookback options. No Lipschitz assumption is made for P.
- Let $\widehat{P}_\ell \approx P$ be the approximation obtained by approximating $S(t)$ with a numerical method with step-size $\tau_\ell = M^{-\ell}T$. For some types of options (e.g. Asian options), the payoff function has to be discretized, too. It is assumed that $M \in \mathbb{N}$ and $M \geq 2$.

The following result is due to Giles, cf. [Gil08b].

Theorem 1.3.1 (Complexity theorem) *Assume that there are independent estimators $\widehat{Y}_\ell$ based on $m_\ell \in \mathbb{N}$ Monte Carlo samples and constants $\alpha \geq \frac{1}{2}, \beta, c_1, c_2, c_3 > 0$ such that the following holds:*

(A1) $|\mathbb{E}[\widehat{P}_\ell - P]| \leq c_1 \tau_\ell^\alpha$ *(weak error)*

(A2) $\mathbb{E}[\widehat{Y}_\ell] = \begin{cases} \mathbb{E}[\widehat{P}_0] & \text{if } \ell = 0 \\ \mathbb{E}[\widehat{P}_\ell - \widehat{P}_{\ell-1}] & \text{if } \ell > 0 \end{cases}$

(A3) $\mathbb{V}[\widehat{Y}_\ell] \leq c_2 m_\ell^{-1} \tau_\ell^\beta$

(A4) *The computational work R_ℓ to compute $\widehat{Y}_\ell$ is bounded by $R_\ell \leq c_3 m_\ell \tau_\ell^{-1}$.*

If $2 \leq M \in \mathbb{N}$, then there is a constant $c_4 > 0$ such that for every $\varepsilon < e^{-1}$, there are $L \in \mathbb{N}$ and $m_\ell \in \mathbb{N}$ such that the multilevel estimator

$$\widehat{Y} = \sum_{\ell=0}^{L} \widehat{Y}_\ell$$

achieves the accuracy

$$\mathrm{MSE}(\widehat{Y}) = \mathbb{E}\left[(\widehat{Y} - \mathbb{E}[P])^2\right] = \mathbb{V}[\widehat{Y}] + \mathbb{E}\left[\widehat{Y} - P\right]^2 \leq \varepsilon^2 \tag{1.14}$$

with a total computational work of

$$R_{ML} \leq \begin{cases} c_4 \varepsilon^{-2} & \text{if } \beta > 1, \\ c_4 \varepsilon^{-2} (\ln \varepsilon)^2 & \text{if } \beta = 1, \\ c_4 \varepsilon^{-2-(1-\beta)/\alpha} & \text{if } 0 < \beta < 1. \end{cases} \tag{1.15}$$

Proof: We define $\lceil x \rceil := \min\{k \in \mathbb{Z} : k \geq x\}$, i.e. $x \leq \lceil x \rceil < x + 1$. Choose

$$L = \left\lceil \frac{\ln(\sqrt{2} c_1 T^\alpha \varepsilon^{-1})}{\alpha \ln M} \right\rceil. \tag{1.16}$$

The reader can check that this implies that

$$\frac{1}{\sqrt{2}} M^{-\alpha} \varepsilon < c_1 \tau_L^\alpha \leq \frac{1}{\sqrt{2}} \varepsilon. \tag{1.17}$$

With the assumptions (A1) and (A2), the bias term in (1.14) can be bounded as follows:

$$\begin{aligned}
\mathbb{E}\left[\widehat{Y} - P\right]^2 &= \left(\sum_{\ell=0}^{L} \mathbb{E}[\widehat{Y}_\ell] - \mathbb{E}[P] \right)^2 \\
&\overset{(A2)}{=} \left(\mathbb{E}[\widehat{P}_0] + \mathbb{E}[\widehat{P}_1 - \widehat{P}_0] + \cdots + \mathbb{E}[\widehat{P}_L - \widehat{P}_{L-1}] - \mathbb{E}[P] \right)^2 \\
&= \left(\mathbb{E}[\widehat{P}_L] - \mathbb{E}[P] \right)^2 \\
&\overset{(A1)}{\leq} (c_1 \tau_L^\alpha)^2 \overset{(1.17)}{\leq} \frac{\varepsilon^2}{2}.
\end{aligned} \tag{1.18}$$

We will later show that $\mathbb{V}[\widehat{Y}] \leq \varepsilon^2/2$, which then yields the accuracy (1.14) for the mean-square error.

As a preparatory step, we prove a bound for $\sum_{\ell=0}^{L} \tau_\ell^{-1}$. Since $M \geq 2$ by assumption and $\tau_\ell^{-1} = M^{\ell - L} \tau_L^{-1}$, the geometric sum yields

$$\sum_{\ell=0}^{L} \tau_\ell^{-1} = \tau_L^{-1} \sum_{\ell=0}^{L} M^{\ell - L} = \tau_L^{-1} \frac{1 - (M^{-1})^{L+1}}{1 - M^{-1}} = \tau_L^{-1} \frac{M - (M^{-1})^L}{M - 1} < \frac{M}{M-1} \tau_L^{-1}. \tag{1.19}$$

It follows from (1.17) that

$$\tau_L^{-1} < M \left(\frac{\sqrt{2} c_1}{\varepsilon} \right)^{1/\alpha} \leq M \left(\sqrt{2} c_1 \right)^{1/\alpha} \varepsilon^{-2}$$

because the assumptions $\alpha \geq 1/2$ and $\varepsilon < e^{-1} < 1$ yield $\varepsilon^{-1/\alpha} \leq \varepsilon^{-2}$. Substituting into (1.19) gives

$$\sum_{\ell=0}^{L} \tau_\ell^{-1} < \frac{M^2}{M-1} \left(\sqrt{2} c_1 \right)^{1/\alpha} \varepsilon^{-2}, \tag{1.20}$$

Now the three cases $\beta = 1$, $\beta > 1$ and $0 < \beta < 1$ have to be investigated.

Case $\beta = 1$. Choose $m_\ell = \lceil 2\varepsilon^{-2}(L+1)c_2\tau_\ell \rceil$. This yields

$$\begin{aligned}\mathbb{V}[\widehat{Y}] = \sum_{\ell=0}^{L} \mathbb{V}[\widehat{Y}_\ell] &\overset{(A3)}{\leq} \sum_{\ell=0}^{L} c_2 m_\ell^{-1} \tau_\ell \\ &\leq \sum_{\ell=0}^{L} c_2\tau_\ell \big(\underbrace{2\varepsilon^{-2}(L+1)c_2\tau_\ell}_{\leq m_\ell} \big)^{-1} = \sum_{\ell=0}^{L} \frac{1}{L+1} \cdot \frac{1}{2}\varepsilon^2 = \frac{1}{2}\varepsilon^2\end{aligned}$$

Together with (1.18), this proves the bound (1.14) for the mean-square error.

In order to prove the bound (1.15) for the computational work, we note that (1.16) provides the upper bound

$$L < \frac{\ln \varepsilon^{-1}}{\alpha \ln M} + \frac{\ln(\sqrt{2}c_1 T^\alpha)}{\alpha \ln M} + 1.$$

Since $\ln \varepsilon^{-1} > \ln e = 1$ by assumption, it follows that

$$L + 1 < c_5 \ln \varepsilon^{-1} \quad \text{with} \quad c_5 = \frac{1}{\alpha \ln M} + \max\left(0, \frac{\ln(\sqrt{2}c_1 T^\alpha)}{\alpha \ln M}\right) + 2. \tag{1.21}$$

Our particular choice of m_ℓ implies

$$m_\ell < 2\varepsilon^{-2}(L+1)c_2\tau_\ell + 1.$$

Hence, the computational costs of the multilevel estimator $\widehat{Y} = \sum_{\ell=0}^{L} \widehat{Y}_\ell$ are bounded by

$$\begin{aligned}R_{ML} \overset{(A4)}{\leq} c_3 \sum_{\ell=0}^{L} m_\ell \tau_\ell^{-1} &\leq c_3 \sum_{\ell=0}^{L} \big(2\varepsilon^{-2}(L+1)c_2\tau_\ell + 1\big)\tau_\ell^{-1} \\ &\leq c_3 \left(2c_2\varepsilon^{-2}(L+1)^2 + \sum_{\ell=0}^{L} \tau_\ell^{-1} \right).\end{aligned}$$

Substituting (1.21) and (1.20) yields

$$R_{ML} \leq c_3 \left(2c_2\varepsilon^{-2} \underbrace{(c_5 \ln \varepsilon^{-1})^2}_{=(c_5 \ln \varepsilon)^2} + \left(\frac{M^2}{M-1} \left(\sqrt{2}c_1\right)^{1/\alpha} \varepsilon^{-2} \right) \right)$$

and since $1 < (\ln \varepsilon)^2$ we obtain

$$R_{ML} \leq c_4 \varepsilon^{-2} (\ln \varepsilon)^2 \quad \text{with} \quad c_4 = 2c_2c_3c_5^2 + c_3 \frac{M^2}{M-1} \left(\sqrt{2}c_1\right)^{1/\alpha}$$

which proves (1.15) in the case $\beta = 1$.

Case $\beta > 1$. Choosing

$$m_\ell = \left\lceil 2\varepsilon^{-2} c_2 \frac{T^{(\beta-1)/2}}{1 - M^{-(\beta-1)/2}} \tau_\ell^{(\beta+1)/2} \right\rceil.$$

yields

$$\mathbb{V}[\widehat{Y}] = \sum_{\ell=0}^{L} \mathbb{V}[\widehat{Y}_\ell] \overset{(A3)}{\leq} \sum_{\ell=0}^{L} c_2 m_\ell^{-1} \tau_\ell^\beta \leq \frac{1}{2}\varepsilon^2 T^{-(\beta-1)/2} \left(1 - M^{-(\beta-1)/2}\right) \sum_{\ell=0}^{L} \tau_\ell^{(\beta-1)/2}. \quad (1.22)$$

Since $\tau_\ell = M^{-\ell} T$ by definition, we obtain

$$\begin{aligned} \sum_{\ell=0}^{L} \tau_\ell^{(\beta-1)/2} &= T^{(\beta-1)/2} \sum_{\ell=0}^{L} \left(M^{-(\beta-1)/2}\right)^\ell = T^{(\beta-1)/2} \frac{1 - \left(M^{-(\beta-1)/2}\right)^{L+1}}{1 - M^{-(\beta-1)/2}} \\ &< \frac{T^{(\beta-1)/2}}{1 - M^{-(\beta-1)/2}}. \end{aligned} \quad (1.23)$$

Substituting (1.23) into (1.22) yields the bound

$$\mathbb{V}[\widehat{Y}] \leq \frac{1}{2}\varepsilon^2$$

which, together with (1.18), proves (1.14).

In order to prove the bound (1.15) for the computational work, we note that our choice of m_ℓ implies

$$m_\ell \leq 2\varepsilon^{-2} c_2 \frac{T^{(\beta-1)/2}}{1 - M^{-(\beta-1)/2}} \tau_\ell^{(\beta+1)/2} + 1$$

and hence

$$R_{ML} \overset{(A4)}{\leq} c_3 \sum_{\ell=0}^{L} m_\ell \tau_\ell^{-1} \leq c_3 \left(2\varepsilon^{-2} c_2 \frac{T^{(\beta-1)/2}}{1 - M^{-(\beta-1)/2}} \sum_{\ell=0}^{L} \tau_\ell^{(\beta-1)/2} + \sum_{\ell=0}^{L} \tau_\ell^{-1} \right).$$

Hence, it follows from (1.20) and (1.23) that

$$R_{ML} \leq \underbrace{c_3 \left(2c_2 \left(\frac{T^{(\beta-1)/2}}{1 - M^{-(\beta-1)/2}} \right)^2 + \frac{M^2}{M-1} \left(\sqrt{2} c_1 \right)^{1/\alpha} \right)}_{=:c_4} \varepsilon^{-2} = c_4 \varepsilon^{-2}$$

which proves (1.15) for $\beta > 1$.

Case $0 < \beta < 1$: Similar arguments. See Theorem 3.1 in [Gil08b] for details. ∎

Discussion.

1. The proof shows that the parameter β from (A3), i.e.

$$\mathbb{V}[\widehat{Y}_\ell] \overset{(A3)}{\leq} c_2 m_\ell^{-1} \tau_\ell^\beta,$$

plays an important role. For $\beta = 1$ and $\beta > 1$, we choose

$$m_\ell = \mathcal{O}\Big(\tau_\ell^{(\beta+1)/2}\Big).$$

This implies that the numerical costs on level l is

$$\mathcal{O}\big(m_\ell \tau_\ell^{-1}\big) = \mathcal{O}\Big(\tau_\ell^{(\beta-1)/2}\Big).$$

If $\beta = 1$, then the numerical work is approximately the same on all levels, but for $\beta > 1$, the computational work on the coarser levels is larger than on the finer levels.

2. For European options and the Euler-Maruyama approximation (cf. 1.2) we have $\alpha = 1$ and $\beta = 1$.

3. If the Milstein method is used and the payoff is Lipschitz continuous, then the computational work is reduced. For the Milstein method, we have

$$\mathbb{E}[\widehat{P}_\ell - P] = \mathcal{O}(\tau_\ell) \qquad \text{(weak convergence)} \tag{1.24}$$

$$\mathbb{E}\big[|\widehat{S}_{\ell,M^\ell} - S(T)|^2\big] = \mathcal{O}\big(\tau_\ell^2\big) \qquad \text{(strong convergence)} \tag{1.25}$$

and hence (1.6) can be replaced by

$$\mathbb{V}[\widehat{P}_\ell - P] \leq c^2 \mathbb{E}\big[|\widehat{S}_{\ell,M^\ell} - S(T)|^2\big] \overset{(1.25)}{=} \mathcal{O}\big(\tau_\ell^2\big).$$

As a consequence, (1.7) turns to

$$\mathbb{V}[\widehat{P}_\ell - \widehat{P}_{\ell-1}] \leq \Big((\mathbb{V}[\widehat{P}_\ell - P])^{1/2} + (\mathbb{V}[\widehat{P}_{\ell-1} - P])^{1/2}\Big)^2 = \mathcal{O}\big(\tau_\ell^2\big),$$

and we obtain

$$\mathbb{V}[\widehat{Y}_\ell] = \mathbb{V}\left[m_\ell^{-1} \sum_{i=1}^{m_\ell} \Big(\widehat{P}_\ell^{(\ell,i)} - \widehat{P}_{\ell-1}^{(\ell,i)}\Big)\right] = m_\ell^{-2} \sum_{i=1}^{m_\ell} \mathbb{V}\left[\widehat{P}_\ell^{(\ell,i)} - \widehat{P}_{\ell-1}^{(\ell,i)}\right] = \mathcal{O}\big(m_\ell^{-1}\tau_\ell^2\big).$$

Hence, we are in the situation $\beta = 2 > 1$ in (1.15), and the computational work is only $\mathcal{O}(\varepsilon^{-2})$ instead of $\mathcal{O}(\varepsilon^{-2}(\ln \varepsilon)^2)$.

4. How cheap or expensive is a numerical work of $\mathcal{O}(\varepsilon^{-2})$ for the accuracy $\mathrm{MSE}(\widehat{Y}) = \mathcal{O}(\varepsilon^2)$? For comparison, we consider a very simple situation.
   Consider a European put or call and assume that $S(t)$ is the geometric Brownian motion, i.e. $f(t,S) = rS$ and $g(t,S) = \sigma S$ in (SDE). We do not have to approximate $S(T)$ with many Milstein steps, because we can evaluate the exact solution

   $$S(t) = \exp\left(\left(r - \tfrac{\sigma^2}{2}\right)t + \sigma W(t)\right) S(0)$$

   directly (much cheaper!).

   Now we use this information in the standard Monte Carlo method:

   $$\mathbb{E}_{\mathbb{Q}}\Big[\psi\big(S(T)\big)\Big] \approx \frac{1}{m}\sum_{j=1}^{m} \psi\big(S(T,\omega_j)\big) =: \tilde{\theta}.$$

   As in Section 1.1 it follows that

   $$\mathrm{MSE}(\tilde{\theta}) = \mathbb{V}(\tilde{\theta}) + \underbrace{\mathbb{E}(\tilde{\theta} - \theta)^2}_{=0} \leq \frac{C}{m}$$

   and hence

   $$\mathrm{MSE}(\tilde{\theta}) = \mathcal{O}\big(\varepsilon^2\big) \qquad \Longleftrightarrow \qquad m = \mathcal{O}\big(\varepsilon^{-2}\big) \text{ samples.}$$

   If every evaluation of $S(T,\omega_j)$ is about as expensive as one time-step of the Milstein method, then standard Monte Carlo with the *exact* underlying is *still not cheaper* than multilevel Monte Carlo with Milstein approximation of $S(T,\omega)$. This is surprising.

5. Asian, lookback, barrier and digital options have been studied in [Gil08a]. It was shown how estimators with $\beta > 1$ can be constructed with the Milstein method. The convergence analysis is more complicated, in particular if the payoff function $\psi(S(T))$ is not Lipschitz continuous. This case has been analyzed in [GHM09].

6. The choice of the optimal value M is discussed in section 4.1 in [Gil08b]. $M = 4$ was used for the numerical experiments presented there.

1.4 Algorithm and implementation

The complexity theorem guarantees that $\mathrm{MSE}(\widehat{Y}) \leq \varepsilon^2$ if all assumptions are true. In order to implement the multilevel Monte Carlo method, however, we need to know the parameters L (= number of levels) and m_ℓ (= number of samples used to compute $\widehat{Y}_\ell$). In principle, formulas to determine these parameters are given in the proof, but these formulas depend on the constants c_1 and c_2 from assumption (A1) and (A3), and these

constants are usually not explicitly known. Hence, we have to estimate L and m_ℓ from the data which has already been computed.
In order to have

$$\mathrm{MSE}(\widehat{Y}) = \mathbb{V}[\widehat{Y}] + \mathbb{E}\left[\widehat{Y} - P\right]^2 \leq \varepsilon^2,$$

we impose that

$$\mathbb{V}[\widehat{Y}] \leq \frac{\varepsilon^2}{2} \qquad \text{and} \qquad \mathbb{E}\left[\widehat{Y} - P\right]^2 \leq \frac{\varepsilon^2}{2}.$$

For simplicity, we consider again European puts or calls and use the same notation as in 1.2.

Optimal number of samples. Let again $\mathbb{V}_\ell = \mathbb{V}\left[\widehat{P}_\ell^{(\ell,i)} - \widehat{P}_{\ell-1}^{(\ell,i)}\right]$ be the variance of a single path (same variance for every path). For given $\mathbb{V}_\ell$, ε, τ_ℓ, the optimal number of samples is

$$m_\ell = \left\lceil 2\varepsilon^{-2}\sqrt{\mathbb{V}_\ell \tau_\ell}\left(\sum_{k=0}^{L}\sqrt{\mathbb{V}_k/\tau_k}\right)\right\rceil, \tag{1.26}$$

because with this choice, we obtain

$$\begin{aligned}
\mathbb{V}[\widehat{Y}] &\overset{(1.2)}{=} \sum_{\ell=0}^{L}\frac{\mathbb{V}_\ell}{m_\ell} \\
&\leq \sum_{\ell=0}^{L}\mathbb{V}_\ell\left(2\varepsilon^{-2}\sqrt{\mathbb{V}_\ell\tau_\ell}\left(\sum_{k=0}^{L}\sqrt{\mathbb{V}_k/\tau_k}\right)\right)^{-1} \\
&= \left(\sum_{k=0}^{L}\sqrt{\mathbb{V}_k/\tau_k}\right)^{-1}\frac{\varepsilon^2}{2}\sum_{\ell=0}^{L}\frac{\mathbb{V}_\ell}{\sqrt{\mathbb{V}_\ell\tau_\ell}} = \frac{\varepsilon^2}{2}.
\end{aligned}$$

The variance $\mathbb{V}_\ell$, however, is not known and must be estimated from the available data (see algorithm below)

Bias estimation. For a method with weak order one, we know that

$$\mathbb{E}[P - \widehat{P}_\ell] \approx C\tau_\ell$$

and hence

$$\begin{aligned}
\mathbb{E}[\widehat{P}_\ell - \widehat{P}_{\ell-1}] &= \mathbb{E}[\widehat{P}_\ell - P] + \mathbb{E}[P - \widehat{P}_{\ell-1}] \approx -C\tau_\ell + C\underbrace{\tau_{\ell-1}}_{M\tau_\ell} \\
&= C\tau_\ell(M-1) \approx (M-1)\mathbb{E}[P - \widehat{P}_\ell]
\end{aligned}$$

If $\mathbb{E}[\widehat{Y}_L] \leq \frac{\varepsilon}{\sqrt{2}}(M-1)$, then it follows that

$$\mathbb{E}[P - \widehat{P}_L] \approx \frac{\mathbb{E}[\widehat{P}_L - \widehat{P}_{L-1}]}{M-1} \approx \frac{\mathbb{E}[\widehat{Y}_L]}{M-1} \leq \frac{\varepsilon}{\sqrt{2}}.$$

Since $\mathbb{E}[\widehat{Y}_L]$ is unknown but $\widehat{Y}_L$ is available, this suggests to increase the value of L until $|\widehat{Y}_L| \leq (M-1)\varepsilon/\sqrt{2}$. In practice, however, this is not reliable. The approximation

$$\mathbb{E}[\widehat{P}_L - \widehat{P}_{L-1}] \approx \frac{\mathbb{E}[\widehat{P}_{L-1} - \widehat{P}_{L-2}]}{M} \approx \frac{\mathbb{E}[\widehat{Y}_{L-1}]}{M}$$

suggests to impose the additional condition $|\widehat{Y}_{L-1}|/M \leq (M-1)\varepsilon/\sqrt{2}$, i.e. to use the terminal condition

$$\max\left\{|\widehat{Y}_L|, \frac{|\widehat{Y}_{L-1}|}{M}\right\} \leq \frac{(M-1)\varepsilon}{\sqrt{2}}. \tag{1.27}$$

Now we sketch algorithms for implementing the multilevel Monte Carlo method. For simplicity, we only consider European options.

Computing a single sample $\widehat{P}_0^{(0,i)}$.

- Generate a path $t \mapsto W(t, \omega_{0,i})$ of the Wiener process with step-size $\tau_0 = T$.
- Approximate the solution of the SDE (1.13) with one single time-step and step-size $\tau_0 = T$. Let $\widehat{S}_{0,1}^{(0,i)} \approx S(T)$ be the approximation of the price at maturity.
- Compute the corresponding (discounted) payoff: $\widehat{P}_0^{(0,i)} = \psi\left(\widehat{S}_{0,1}^{(0,i)}\right)$.

Computing a single sample of $\widehat{P}_\ell^{(\ell,i)} - \widehat{P}_{\ell-1}^{(\ell,i)}$ with $\ell > 0$.

- Generate a path $t \mapsto W(t, \omega_{\ell,i})$ of the Wiener process with step-size $\tau_\ell = M^{-\ell}T$.
- For $k \in \{\ell-1, \ell\}$:
  - Approximate the solution of the SDE (1.13) by M^k time-steps with step-size $\tau_k = M^{-k}T$. Let $\widehat{S}_{k,M^k}^{(\ell,i)} \approx S(T)$ be the approximation of the price at maturity.
  - Compute the corresponding (discounted) payoff: $\widehat{P}_k^{(\ell,i)} = \psi\left(\widehat{S}_{k,M^k}^{(\ell,i)}\right)$

  Use the same path of the Wiener process for both approximations!

Main algorithm.

1. Set $L = 0$.
2. Set $m_L = 10^4$.

3. For $\ell = 0, \ldots, L$, compute $\widetilde{\mathbb{V}}_\ell \approx \mathbb{V}_\ell$:
   For $i = 1, \ldots, m_\ell$, compute samples

$$X_\ell^{(\ell,i)} := \begin{cases} \widehat{P}_0^{(\ell,i)} & \text{if } \ell = 0, \\ \widehat{P}_\ell^{(\ell,i)} - \widehat{P}_{\ell-1}^{(\ell,i)} & \text{if } \ell > 0. \end{cases}$$

   Then, set

$$\mu_\ell := \frac{1}{m_\ell} \sum_{i=1}^{m_\ell} X_\ell^{(\ell,i)}, \qquad \widetilde{\mathbb{V}}_\ell := \frac{1}{m_\ell} \sum_{i=1}^{m_\ell} \left(X_\ell^{(\ell,i)} \right)^2 - \mu_\ell^2.$$

4. For $\ell = 0, \ldots, L$ compute the optimal number of samples according to (1.26):

$$m_\ell = \max \left\{ m_\ell, \left\lceil 2\varepsilon^{-2} \sqrt{\widetilde{\mathbb{V}}_\ell \tau_\ell} \left(\sum_{\ell=0}^{L} \sqrt{\widetilde{\mathbb{V}}_\ell / \tau_\ell} \right) \right\rceil \right\}$$

5. For $\ell = 0, \ldots, L$ compute additional samples $\widehat{P}_0^i$ and $\widehat{P}_\ell^{(\ell,i)} - \widehat{P}_{\ell-1}^{(\ell,i)}$ if m_ℓ has increased in step 4.

6. If $L \geq 2$, then test for convergence. Compute $\widehat{Y}_L$ and $\widehat{Y}_{L-1}$ according to

$$\widehat{Y}_\ell := \frac{1}{m_\ell} \sum_{i=1}^{m_\ell} \left(\widehat{P}_\ell^{(\ell,i)} - \widehat{P}_{\ell-1}^{(\ell,i)} \right) \qquad \text{for } \ell = L-1, L. \tag{1.28}$$

   If the terminal condition (1.27) is true, i.e. if

$$\max \left\{ |\widehat{Y}_L|, \frac{|\widehat{Y}_{L-1}|}{M} \right\} \leq \frac{(M-1)\varepsilon}{\sqrt{2}}.$$

   then go to step 8.

7. Set $L := L + 1$ and go to step 2.

8. Compute the multilevel Monte Carlo estimator

$$\widehat{Y} = \sum_{\ell=0}^{L} \widehat{Y}_\ell$$

   with $\widehat{Y}_\ell$ defined by (1.28) for $\ell > 0$ and $\widehat{Y}_0 := \frac{1}{m_0} \sum_{i=1}^{m_0} \widehat{P}_0^{(\ell,i)}$. Then, $\widehat{Y}$ is the approximation of the option price.

Remarks.

1. This algorithm is based on heuristic estimates. There is no guarantee that the mean-square error will be below the error tolerance, but the algorithm performs very well in numerical tests.

2. For the implementation it is suggested to store the variables

$$\sum_{i=1}^{m_\ell} X_\ell^{(\ell,i)} \quad \text{and} \quad \sum_{i=1}^{m_\ell} \left(X_\ell^{(\ell,i)}\right)^2 \quad \text{where} \quad X_\ell^{(\ell,i)} := \begin{cases} \widehat{P}_0^{(\ell,i)} & \text{if } \ell = 0, \\ \widehat{P}_\ell^{(\ell,i)} - \widehat{P}_{\ell-1}^{(\ell,i)} & \text{if } \ell > 0. \end{cases}$$

   These terms can easily be updated when new samples are computed, and with these variables, μ_ℓ, $\widetilde{\mathbb{V}}_\ell$, and $\widehat{Y}_\ell$ can be easily computed.

3. Of course, most of the "for"-loops can (and should) be avoided by vectorization.

Chapter 2

Historical, implied and local volatility

As an alternative to Monte Carlo methods, European options can be priced by solving the Black-Scholes equation

$$\partial_t V(t,S) + \frac{\sigma^2}{2} S^2 \partial_S^2 V(t,S) + rS\partial_S V(t,S) - rV(t,S) = 0 \qquad t \in [0,T], S \geq 0 \quad \text{(BSE)}$$
$$V(T,S) = \psi(S)$$

(cf. chapter 3.2 in part I). $V(t,S)$ is the value of the option, $S \geq 0$ is the price of the underlying, ψ denotes the (undiscounted) payoff function, $T > 0$ is the maturity time, $r > 0$ is the risk-free interest rate, and $\sigma \in \mathbb{R}$ is the volatility.

We consider a European call, i.e. $\psi(S) = (S-K)^+$. Similar results hold for puts.

In the standard Black-Scholes market, the solution of (BSE) is given by the Black-Scholes formula:

$$V(t,S) = S\Phi(d_1) - K\exp(-r(T-t))\Phi(d_2) \tag{2.1}$$
$$\text{with } \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-s^2/2}\, ds \qquad \text{and} \qquad d_{1/2} = \frac{\ln \frac{S}{K} + (r \pm \sigma^2/2)\,(T-t)}{\sigma\sqrt{T-t}}$$

(cf. chapter 3.3 in part I). In practice, the parameters T, K and r are typically given, but the volatility σ is not known.

Goal: Determine σ.

2.1 Historical volatility

Idea: Approximate $\sigma \approx \sigma_{hist}$ from values of the underlying in the past.

Let $S(n)$ be the value of the underlying on day $n = 1, \ldots, N$. Assume that $S(t)$ is the geometric Brownian motion

$$S(t) = S_0 \exp(at + \sigma W_t), \qquad a = r - \frac{\sigma^2}{2}$$

and let

$$y_n := \ln S(n+1) - \ln S(n) = a + \sigma \Delta W_n, \qquad \Delta W_n = W(n+1) - W(n).$$

Since $\mathbb{E}[\Delta W_n] = 0$ and $\mathbb{V}[\Delta W_n] = 1$, it follows that $\mathbb{E}[y_n] = a$ and $\mathbb{V}[y_n] = \sigma^2$. Then, we define

$$\bar{y} := \frac{1}{N-1} \sum_{n=1}^{N-1} y_n \approx a$$

and the historical volatility by

$$\sigma_{hist} := \sqrt{N_{trade}} \cdot \sqrt{\frac{1}{N-2} \sum_{n=1}^{N-1} (y_n - \bar{y})^2}$$

where N_{trade} is the total number of trading days per year, and the term $\frac{1}{N-2} \sum_{n=1}^{N-1} (y_n - \bar{y})^2$ is the sample variance. Note that we divide by $N-2$ instead of $N-1$ (Bessel's correction).

2.2 Implied volatility

Assume that there is a $t_\star \in (0, T)$ and an $S_\star > 0$ such that $V_\star = V(t_\star, S_\star)$ is known. Consider the Black-Scholes formula with $t = t_\star$, $S = S_\star$ as a function in σ, i.e.

$$w(\sigma) = S_\star \Phi(d_1) - K \exp(-r(T - t_\star)) \Phi(d_2)$$
$$d_{1/2}(\sigma) = \frac{\ln \frac{S_\star}{K} + (r \pm \sigma^2/2)\,(T - t_\star)}{\sigma \sqrt{T - t_\star}}.$$

Find σ_{impl} such that $w(\sigma_{impl}) = V_\star$. Question: Is there a unique solution?

It can be shown that $\partial_\sigma V(t, S) > 0$ for all $t \in [0, T]$ and $S > 0$. Consequence: $w'(\sigma) > 0$. In chapter 1.4 in part I, it has been shown that

$$\left(S_\star - K e^{-r(T - t_\star)}\right)^+ \leq V_\star \leq S_\star.$$

For $\sigma \longrightarrow 0$, we have

$$d_{1/2}(\sigma) \longrightarrow \infty \qquad \Longrightarrow \qquad \Phi(d_{1/2}(\sigma)) \longrightarrow 1$$

and hence

$$\lim_{\sigma\to 0} w(\sigma) = S_\star \underbrace{\lim_{\sigma\to 0} \Phi\big(d_1(\sigma)\big)}_{=1} - K\exp(-r(T-t_\star)) \underbrace{\lim_{\sigma\to 0} \Phi\big(d_2(\sigma)\big)}_{=1} \le V_\star.$$

For $\sigma \longrightarrow \infty$, we have $d_1(\sigma) \longrightarrow \infty$, $d_2(\sigma) \longrightarrow -\infty$ and hence

$$\lim_{\sigma\to\infty} w(\sigma) = S_\star - K\cdot 0 \ge V_\star$$

This yields

$$w(\sigma) - V_\star \le 0 \text{ for } \sigma \to 0, \qquad w(\sigma) - V_\star \ge 0 \text{ for } \sigma \to \infty,$$

and since w is continuous, it follows that a unique solution exists.
The solution σ_{impl} can be approximated with Newton's method.

Newton's method

Let $f \in C^1(\mathbb{R}^d)$, i.e. $f : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ be continuously differentiable ($d \in \mathbb{N}$).

Goal: Find $\xi \in \mathbb{R}^d$ such that $f(\xi) = 0$.

Idea: For given $x_n \approx \xi$, consider the linearization

$$f(x) \approx f(x_n) + f'(x_n)(x - x_n) =: g_n(x).$$

Find x_{n+1} such that $g_n(x_{n+1}) = 0$ and use x_{n+1} as new approximation.

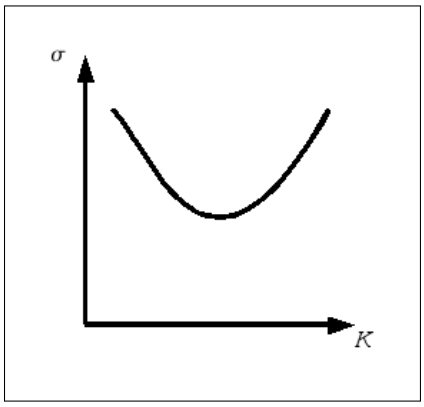
Algorithm:
Choose $x_0 \in \mathbb{R}^d$
For $n = 0, 1, 2, \ldots$
  Solve $f'(x_n)\Delta x_n = -f(x_n)$
  let $x_{n+1} = x_n + \Delta x_n$
end for

Note that $f'(x_n) \in \mathbb{R}^{d\times d}$ is a matrix, and that a linear system has to be solved in order to compute Δx_n.

The theorems by Newton-Mysovskii and Newton-Kantorovich specify conditions for convergence. The domain of convergence can be increased by a damping strategy.
Details: Lectures on numerical mathematics.

If the implied volatility is computed for several data sets (K_n, T_n, V_n), then typically slightly **different** volatilities are obtained. (K_n strike, T_n maturity, V_n option value). If one parameter is varied, one typically obtains a convex curve called the volatility smile. Hence, the modelling assumption that the volatility is constant is questionable. Solutions:

- Stochastic volatility (cf. chapter 5.1 in part I)
- Local volatility: $\sigma = \sigma(t, S)$.



Goals for the rest of this chapter:

- Derive a formula for $\sigma(t, S)$ that is consistent with the observed market data → **Dupire's equation**.
- Numerical approximation of $\sigma(t, S)$ by solving a minimization problem.

Notation: Let $d, D \in \mathbb{N}$, $U \subset \mathbb{R}$, $V \subset \mathbb{R}$, $u : U \to V$, and $k \in \mathbb{N}_0$. For a multi-index $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ we set

$$\partial^\alpha u := \partial_{x_1}^{\alpha_1} \ldots \partial_{x_d}^{\alpha_d} u, \qquad |\alpha|_1 = \alpha_1 + \ldots + \alpha_d$$

and define

$$\begin{aligned} C^k(U,V) &:= \Big\{ u : U \to V, \quad \partial^\alpha u \text{ is continuous for all } \alpha \in \mathbb{N}_0^d \text{ with } |\alpha|_1 \le k \Big\}, \\ C_c^k(U,V) &:= \Big\{ u \in C^k(U,V) \text{ with compact support } \overline{\{x \in U : u(x) \neq 0\}} \Big\}, \\ C^k(U) &:= C^k(U,\mathbb{R}), \qquad C_c^k(U) := C_c^k(U,\mathbb{R}). \end{aligned}$$

2.3 Preparation: The Fokker-Planck equation

Let X_t be the solution of the d-dimensional SDE

$$dX_t = f(t, X_t)dt + g(t, X_t)dW_t$$

with $X_t \in \mathbb{R}^d$, $f(t, X_t) \in \mathbb{R}^d$, $f = (f_1, \ldots, f_d)^T$, $g(t, X_t) \in \mathbb{R}^{d \times m}$, $g = (g_{ij})_{i,j}$ and $W_t \in \mathbb{R}^m$. We assume

- that $x \mapsto f_i(t, x) \in C^1(\mathbb{R}^d)$ and $x \mapsto g_{ij}(t, x) \in C^2(\mathbb{R}^d)$ have continuous partial derivatives on $\mathbb{R}^d$,
- that a unique solution of the SDE exists, and

- that the solution has a smooth density $u(t,x)$, i.e.

$$\mathbb{P}(X_t \in \mathcal{B}) = \int_{\mathcal{B}} u(t,x)dx \qquad \text{for all Borel sets } \mathcal{B} \text{ and } t \geq 0.$$

"Smooth" means that $t \mapsto u(t,x) \in C^1([0,\infty))$ and $x \mapsto u(t,x) \in C^2(\mathbb{R}^d)$.
This assumption implies that X_0 is a *random* vector with density function $u_0(x) := u(0,x)$.

Theorem 2.3.1 (Fokker-Planck equation) *Under these conditions, $u(t,x)$ is the solution of the* ***Fokker-Planck equation (FPE)***

$$\partial_t u(t,x) = \frac{1}{2}\sum_{i,j=1}^{d} \partial_{x_i}\partial_{x_j}\big(G_{ij}(t,x)u(t,x)\big) - \sum_{i=1}^{d}\partial_{x_i}\big(f_i(t,x)u(t,x)\big) \qquad \text{for } t>0, x \in \mathbb{R}^d$$
$$u(0,x) = u_0(x)$$

with $G_{ij}(t,x) = \sum_{k=1}^{m} g_{ik}(t,x)g_{jk}(t,x)$.

If we know $u(0,x)$, we can thus compute $u(t,x)$ for $t>0$ by solving the FPE.

Remark: In the special case $d = m = 1$, the FPE reads

$$\partial_t u(t,x) = \frac{1}{2}\partial_x^2\big(g^2(t,x)u(t,x)\big) - \partial_x\big(f(t,x)u(t,x)\big).$$

Examples for $d = m = 1$:

1. Let $f(t,X_t) \equiv 0$ and $g(t,X_t) \equiv 1$.

$$\begin{aligned} &\Longrightarrow \quad dX_t = f(t,X_t)dt + g(t,X_t)dW_t = dW_t \\ &\Longrightarrow \quad X_t = W_t + X_0 \quad \text{Wiener process plus constant} \end{aligned}$$

Fokker-Planck equation:

$$\partial_t u(t,x) = \frac{1}{2}\partial_x^2 u(t,x) \qquad \longrightarrow \text{heat equation.}$$

2. Let $f(t,X_t) = rX_t$ and $g(t,X_t) = \sigma X_t$ for some $r \in \mathbb{R}, \sigma > 0$.

$$\Longrightarrow dX_t = rX_t dt + \sigma X_t dW_t, \quad X_t \geq 0 \quad \text{(geometric Brownian motion)}$$

Fokker-Planck equation:

$$\partial_t u(t,x) = \frac{1}{2}\sigma^2\partial_x^2\big(x^2 u(t,x)\big) - r\partial_x\big(xu(t,x)\big), \qquad x>0$$

If $u_0(x)$ is a log-normal distribution, the $u(t,x)$ remains a log-normal distribution with time-dependent parameters (cf. Chapter 3 in part I).

3. If $g(t, X_t) \equiv 0$ then $dX_t = f(t, X_t)\, dt$. This is equivalent to the **ordinary** differential equation

$$\frac{dX_t}{dt} = f(t, X_t)$$

with distributed initial data, i.e.

$$\mathbb{P}(X_0 \in \mathcal{B}) = \int_{\mathcal{B}} u(0, x) dx.$$

"Stochastic initial value, deterministic evolution/equation"
Fokker-Planck equation:

$$\partial_t u(t, x) = -\partial_x \big(f(t, x) u(t, x) \big) \qquad \text{(Liouville equation)}$$

For the proof of Theorem 2.3.1 we need the following lemma:

Lemma 2.3.2 (Fundamental lemma of calculus of variations) *Let $\Omega \subset \mathbb{R}$ be connected, open, non-empty, and let $w : \Omega \longrightarrow \mathbb{R}$ be continuous. If*

$$\int_\Omega w(x) v(x)\, dx = 0 \qquad \text{for all } v \in C_c^\infty(\Omega)$$

then $w(x) = 0$ for all $x \in \Omega$.

Sketch of the proof of Lemma 2.3.2: Let $v_r \in C^\infty(\mathbb{R}^d)$ be a function with the following properties:

(a) $v_r(x) > 0$ for $\|x\|_2 < r$ and $v_r(x) = 0$ for $\|x\|_2 > r$

(b) $\displaystyle\int_{\mathbb{R}^d} v_r(x)\, dx = 1$

(The construction of v_r is not presented.) Assume that $w(\xi) > 0$ for a $\xi \in \Omega$. Then, there is a $\varepsilon > 0$ such that

(c) $K_\varepsilon(\xi) := \{x \in \mathbb{R}^d : \|x - \xi\|_2 \leq \varepsilon\} \subset \Omega$, and

(d) $w(x) \geq \frac{1}{2} w(\xi) > 0$ for all $x \in K_\varepsilon(\xi)$.

If we choose $v(x) = v_\varepsilon(x - \xi)$, then

$$\begin{aligned} 0 = \int_\Omega w(x) v_\varepsilon(x - \xi)\, dx &\overset{(a)}{=} \int_{K_\varepsilon(\xi)} w(x) v_\varepsilon(x - \xi)\, dx \\ &\overset{(d)}{\geq} \frac{1}{2} w(\xi) \int_{K_\varepsilon(\xi)} v_\xi(x - \xi)\, dx \overset{(b)}{=} \frac{1}{2} w(\xi) > 0 \end{aligned}$$

which yields a contradiction. ∎

Proof of Theorem 2.3.1. We consider only the one-dimensional case ($d = m = 1$). For $d > 1$ and/or $m > 1$ the assertion can be shown along the same lines (but with more work). For every $v \in C_c^\infty(\mathbb{R})$ the Itô formula yields

$$dv(X_t) = \left(v'(X_t) f(t, X_t) + \frac{1}{2} v''(X_t) g^2(t, X_t) \right) dt + v'(X_t) g(t, X_t) dW_t.$$

Hence, we obtain for the expectation

$$\mathbb{E}(v(X_t)) = \mathbb{E}(v(X_0)) + \int_0^t \mathbb{E}\left(v'(X_s) f(s, X_s) + \frac{1}{2} v''(X_s) g^2(s, X_s) \right) ds$$

and for its time-derivative

$$\begin{aligned} \frac{d}{dt}\mathbb{E}(v(X_t)) &= \mathbb{E}\left(v'(X_t) f(t, X_t) + \frac{1}{2} v''(X_t) g^2(t, X_t) \right) \\ &= \int_{-\infty}^{\infty} \left(v'(x) f(t, x) + \frac{1}{2} v''(x) g^2(t, x) \right) u(t, x)\, dx \end{aligned}$$

since $u(t, \cdot)$ is the density of X_t. We integrate by parts and obtain

$$\begin{aligned} &\int_{-\infty}^{\infty} \left(v'(x) f(t, x) + \frac{1}{2} v''(x) g^2(t, x) \right) u(t, x)\, dx \\ &= \underbrace{\left[\left(v(x) f(t, x) + \frac{1}{2} v'(x) g^2(t, x) \right) u(t, x) \right]_{-\infty}^{\infty}}_{=0 \text{ because } v \in C_c^\infty(\mathbb{R})} \\ &\quad - \int_{-\infty}^{\infty} v(x) \partial_x \big(f(t, x) u(t, x) \big) dx - \int_{-\infty}^{\infty} \frac{1}{2} v'(x) \partial_x \big(g^2(t, x) u(t, x) \big) dx \\ &= - \int_{-\infty}^{\infty} v(x) \partial_x \big(f(t, x) u(t, x) \big) dx + 0 + \int_{-\infty}^{\infty} \frac{1}{2} v(x) \partial_x^2 \big(g^2(t, x) u(t, x) \big) dx. \end{aligned} \tag{2.2}$$

On the other hand, we have

$$\frac{d}{dt}\mathbb{E}(v(X_t)) = \frac{d}{dt} \int_{-\infty}^{\infty} v(x) u(t, x) dx = \int_{-\infty}^{\infty} v(x) \partial_t u(t, x) dx, \tag{2.3}$$

and (2.2) and (2.3) yield

$$\int_{-\infty}^{\infty} v(x) \partial_t u(t, x) dx = \int_{-\infty}^{\infty} v(x) \left(\frac{1}{2} \partial_x^2 \big(g^2(t, x) u(t, x) \big) - \partial_x \big(f(t, x) u(t, x) \big) \right) dx$$

Since this holds for arbitrary test functions $v \in C_c^\infty(\mathbb{R})$, it follows from Lemma 2.3.2 that

$$\partial_t u(t, x) = \frac{1}{2} \partial_x^2 \big(g^2(t, x) u(t, x) \big) - \partial_x \big(f(t, x) u(t, x) \big)$$

which is the one-dimensional Fokker-Planck equation. ∎

2.4 Dupire's equation

Back to option pricing: Suppose that the value $S(t) \geq 0$ of the underlying is described by the SDE

$$\begin{aligned} dS_t &= rS_t dt + \sigma(t, S_t) S_t dW_t \qquad\qquad t \in [0, T] \\ S_0 &= S_* \end{aligned} \tag{2.4}$$

Note that the volatility $\sigma(t, S_t)$ is not constant and can depend on t and S ("local volatility"). Similar to the derivation of the Black-Scholes equation with constant volatility (cf. section 3.2 in part I), we obtain the generalized Black-Scholes equation

$$\begin{aligned} &\partial_t V + \frac{\sigma^2(t,S)}{2} S^2 \partial_S^2 V + rS\partial_S V - rV = 0 \\ &V(T, S) = (S - K)^+ \end{aligned}$$

for calls (similar for puts). Represent the solution as a discounted expectation

$$V(0, S_*) = V(0, S_*, T, K) = e^{-rT} \int_0^\infty \phi(T, x)(x - K)^+ \, dx \tag{2.5}$$

where $\phi(t, x)$ is the density of the SDE (2.4), i.e.

$$\mathbb{P}(S_t \in \mathcal{B} | S_0 = S_*) = \int_{\mathcal{B}} \phi(t, x) dx \qquad \text{for all Borel sets } \mathcal{B}.$$

Note that $\phi(t, x)$ depends on S_*.
Assume that $\sigma(t, S)$ is sufficiently smooth, and let $\phi_\varepsilon(t, x)$ be the solution of the one-dimensional Fokker-Planck equation

$$\begin{aligned} \partial_t \phi_\varepsilon(t, x) &= \frac{1}{2} \partial_x^2 \left(\sigma^2(t, x) x^2 \phi_\varepsilon(t, x) \right) - r \partial_x \big(x \phi_\varepsilon(t, x) \big) \\ \phi_\varepsilon(0, x) &= \frac{1}{\varepsilon \sqrt{2\pi}} \exp\left(-\frac{(x - S_*)^2}{2\varepsilon^2} \right). \end{aligned} \tag{2.6}$$

According to Theorem 2.3.1, $\phi_\varepsilon(t, x)$ is the density associated to the SDE (2.4) with random initial data $S_0 \sim \phi_\varepsilon(0, \cdot)$. Hence, we expect that on a finite time interval $\phi_\varepsilon \approx \phi$ for sufficiently small ε.

On the other hand, deriving (2.5) with respect to K (!!) yields

$$\begin{aligned} e^{rT}\partial_K^2 V(0,S_*,T,K) &= \partial_K^2 \int_K^\infty \phi(T,x)(x-K)dx \\ &= \underbrace{\partial_K^2 \int_{-\infty}^\infty \phi(T,x)(x-K)dx}_{=0} - \partial_K^2 \int_{-\infty}^K \phi(T,x)(x-K)dx \\ &= -\partial_K^2 \int_{-\infty}^K x\phi(T,x)dx + \partial_K^2 \left(K \int_{-\infty}^K \phi(T,x)dx \right) \\ &= -\partial_K(\phi(T,K)K) + \partial_K \int_{-\infty}^K \phi(T,x)dx + \partial_K(K\phi(T,K)) \\ &= \phi(T,K) \end{aligned}$$

According to the Fokker-Planck equation (2.6), we have

$$\partial_T \phi_\varepsilon(T,K) = \frac{1}{2}\partial_K^2 \left(\sigma^2(T,K)K^2\phi_\varepsilon(T,K)\right) - r\partial_K\big(K\phi_\varepsilon(T,K)\big) \tag{2.7}$$

(replace $t \to T$ and $x \to K$). Substituting $\phi_\varepsilon(T,K) \approx \phi(T,K) = e^{rT}\partial_K^2 V(0,S_*,T,K)$ yields for $\varepsilon \to 0$

$$\begin{aligned} \partial_T \left(e^{rT}\partial_K^2 V\right) &= re^{rT}\partial_K^2 V + e^{rT}\partial_K^2\partial_T V \\ &\overset{(2.7)}{=} \frac{1}{2}\partial_K^2 \left(\sigma^2(T,K)K^2 e^{rT}\partial_K^2 V\right) - r\partial_K(Ke^{rT}\partial_K^2 V) \end{aligned}$$

with $V = V(0,S_*,T,K)$. Multiply with e^{-rT}, rearrange terms and use that

$$\partial_K^2 rK\partial_K V = r\partial_K^2 V + r\partial_K(K\partial_K^2 V)$$

(check!). This gives

$$\begin{aligned} 0 &= r\partial_K^2 V + \partial_K^2\partial_T V - \frac{1}{2}\partial_K^2(\sigma^2(T,K)K^2\partial_K^2 V) + r\partial_K(K\partial_K^2 V) \\ &= \partial_K^2 \left(\partial_T V - \frac{1}{2}\sigma^2(T,K)K^2\partial_K^2 V + rK\partial_K V \right). \end{aligned}$$

Integrate twice with respect to K:

$$0 = \partial_T V - \frac{1}{2}\sigma^2(T,K)K^2\partial_K^2 V + rK\partial_K V + A(T)K + B(T)$$

with unknown terms $A(T), B(T)$. Since we consider a call, we have

$$\begin{aligned} \lim_{K\to\infty} V(0,S_*,T,K) &= 0 \\ \lim_{K\to\infty} \partial_T V(0,S_*,T,K) &= 0 \\ \lim_{K\to\infty} \partial_K^i V(0,S_*,T,K) &= 0 \quad i \in \{1,2\} \\ \Rightarrow A(T) = B(T) &= 0 \end{aligned}$$

This yields **Dupire's equation:**

$$\partial_T V - \frac{1}{2}\sigma^2(T,K)K^2\partial_K^2 V + rK\partial_K V = 0, \qquad T > 0, K > 0 \tag{2.8}$$

where $V = V(0, S_*, T, K)$. The initial condition for a call is $V(0, S_*, 0, K) = (S_* - K)^+$. Dupire's equation can be solved for the local volatility:

$$\sigma(T,K) = \frac{\sqrt{2}}{K}\sqrt{\frac{\partial_T V + rK\partial_K V}{\partial_K^2 V}} \tag{2.9}$$

Summary: If $V(0, S_*, T, K)$ is known for fixed S_* and for **all** T, K, then the local volatility is given by (2.9).

2.5 Numerical approximation of the local volatility

Problem: The assumption that $V(0, S_*, T, K)$ were known for **all** T, K is not realistic.

Weaker assumption: The exact value

$$V_\ell^{ex} = V^{ex}(0, S_*, T_\ell, K_\ell), \qquad \ell = 1, \ldots, L$$

is known for fixed S_* and finitely many maturities T_ℓ and strikes K_ℓ.

Let $V_\ell(t, S, \sigma)$ be the solution of the generalized Black-Scholes equation

$$\begin{aligned} &\partial_t V_\ell + \frac{1}{2}\sigma^2(t,S)S^2\partial_S^2 V_\ell + rS\partial_S V_\ell - rV_\ell = 0, \qquad t \in [0, T_\ell] \\ &V_\ell(T_\ell, S) = (S - K_\ell)^+ \end{aligned} \tag{2.10}$$

Goal: Find a function $\sigma(t, S)$ such that

$$V_\ell(0, S_*, \sigma) \approx V_\ell^{ex}$$

with minimal mean-square error, i.e.

$$\sum_{\ell=1}^{L}\left(V_\ell(0, S_*, \sigma) - V_\ell^{ex}\right)^2 = \min_\sigma! \tag{2.11}$$

This is a minimization problem where the argument is a **function**. Since such a function has "infinitely many degrees of freedom" but only L conditions are imposed, this problem has usually infinitely many solutions with zero residuum, but many of these solutions are not plausible. It is natural to assume that the exact volatility function $\sigma(t, S)$ is smooth.

Idea: Consider only functions σ from a finite-dimensional ansatz space $\mathcal{V}$.

Choose a partition $0 \le t_1 < t_2 < \ldots < t_{\widehat{n}}$ and $0 \le S_1 < S_2 < \ldots < S_{\widehat{m}}$ with $\widehat{n} \cdot \widehat{m} < L$. Let $\mathcal{V}$ be the space of bicubic splines $w = w(t, S)$ with nodes (t_n, S_m). This space is uniquely characterized by the following properties:

1. $\mathcal{V} \subset C^2\Big([t_1, t_{\widehat{n}}] \times [S_1, S_{\widehat{m}}], \mathbb{R}\Big)$.

2. Every $w(t,S) \in \mathcal{V}$ is locally a cubic polynomial in both variables: For $t \in [t_n, t_{n+1}]$ and $S \in [S_m, S_{m+1}]$, there is a representation

$$w(t,S) = \sum_{j=0}^{3} \sum_{k=0}^{3} a_{jk}^{(n,m)} \cdot (t - t_n)^j (S - S_m)^k$$

with coefficients $a_{jk}^{(n,m)} \in \mathbb{R}$.

Remark: Note that $w(t,S) \in \mathcal{V}$ is only **piecewise** polynomial, i.e. $w(t,S)$ is represented by *different* polynomials on different subdomains $[t_n, t_{n+1}] \times [S_m, S_{m+1}]$. At the boundary between two subdomains, the corresponding polynomials are "glued together" in such a way that the transition is smooth.

It can be shown that for given values $y_{nm} \in \mathbb{R}$, there is a unique spline $w(t,S) \in \mathcal{V}$ such that

$$\begin{aligned}
w(t_n, S_m) &= y_{nm} && \text{for all } n \in \{1, \dots, \widehat{n}\} \text{ and } m \in \{1, \dots, \widehat{m}\} \\
\partial_t^2 w(t_1, S_m) &= \partial_t^2 w(t_{\widehat{n}}, S_m) = 0 && \text{for all } m \in \{1, \dots, \widehat{m}\} \\
\partial_S^2 w(t_n, S_1) &= \partial_S^2 w(t_n, S_{\widehat{m}}) = 0 && \text{for all } n \in \{1, \dots, \widehat{n}\} \\
\partial_t^2 \partial_S^2 w(t_n, S_m) &= 0 && \text{if } n \in \{1, \widehat{n}\} \text{ and } m \in \{1, \widehat{m}\}
\end{aligned}$$

Hence, the spline is an interpolation of the y_{nm} in the nodes (t_n, S_m), and the spline $w(t,S)$ is uniquely determined by the values y_{nm}. In order to express this, we use the notation

$$w(t,S) = sp(t,S,Y), \qquad Y := \begin{pmatrix} y_{11} & \cdots & y_{1\widehat{m}} \\ \vdots & & \vdots \\ y_{\widehat{n}1} & \cdots & y_{\widehat{n}\widehat{m}} \end{pmatrix}.$$

For $u \in C^2\Big([t_1, t_{\widehat{n}}] \times [S_1, S_{\widehat{m}}], \mathbb{R}\Big)$ we define the functional

$$\mathcal{F}(u) = \left(\int_{t_1}^{t_{\widehat{n}}} \int_{S_1}^{S_{\widehat{m}}} (\partial_S^2 u)^2 + (\partial_S \partial_t u)^2 + (\partial_t^2 u)^2 \, dS dt \right)^{\frac{1}{2}}.$$

$\mathcal{F}(u)$ measures (approximately) the curvature of a function $u(t,S)$. It can be shown that

$$\mathcal{F}\big(sp(t,S,Y)\big) \leq \mathcal{F}(u)$$

for all $u \in C^2\Big([t_1, t_{\widehat{n}}] \times [S_1, S_{\widehat{m}}], \mathbb{R}\Big)$ which interpolate the values y_{nm} in the nodes (t_n, S_m). Hence, the spline is the smoothest (in terms of curvature) interpolant.

The construction of the bicubic interpolating spline for given data y_{nm} will not be explained. The one-dimensional case is presented in the lecture *Numerik 1/2*.
Now we approximate the unknown volatility function by the spline which interpolates the values $\{y_{nm}\}_{n,m}$ in the nodes (t_n, S_m):

$$\sigma(t,S) \approx sp(t,S,Y), \qquad Y := \begin{pmatrix} y_{11} & \cdots & y_{1\widehat{m}} \\ \vdots & & \vdots \\ y_{\widehat{n}1} & \cdots & y_{\widehat{n}\widehat{m}} \end{pmatrix}, \qquad \widehat{n}\cdot\widehat{m} < L.$$

The unknown matrix Y is supposed to be determined in such a way that the corresponding spline $sp(t,S,Y)$ solves the nonlinear least-squares problem

$$\|f(Y)\|_2^2 = \sum_{\ell=1}^{L} f_\ell^2(Y) = \min_Y ! \tag{2.12a}$$

$$f_\ell(Y) = V_\ell\big(0, S_*, sp(t,S,Y)\big) - V_\ell^{ex}. \tag{2.12b}$$

$V_\ell\big(0, S_*, sp(t,S,Y)\big)$ could in principle be computed by solving the generalized Black-Scholes equation (2.10), but then L PDE initial value problems have to be solved for every evaluation of f. Instead, one can solve Dupire's equation (2.8) **once** and evaluate the solution at (T_ℓ, K_ℓ).
Approximating the local volatility is an **inverse problem**:

	given	unknown
forward problem	$T, K, \sigma(t,S)$	$V(t,S)$
inverse problem	$V_\ell^{ex} \approx V_\ell(t,S,\sigma)$	$\sigma(t,S)$

Inverse problems are notoriously ill-conditioned: Small perturbations of the data $V_\ell^{ex} = V^{ex}(0, S_*, T_\ell, K_\ell)$ can change the result significantly. This effect can be reduced by regularization. The approximation via splines can be considered as an implicit regularization. Typically, the solution Y of (2.12) is not sensitive to the data V_ℓ^{ex} (but sensitive to the choice of the nodes (t_n, S_m)).

Nonlinear least-squares problems and the Gauss-Newton method

Consider the nonlinear least-squares problem

$$\|f(y)\|_2^2 = \sum_{\ell=1}^{L} f_\ell^2(y) = \min_y !$$

with a given function $f : \mathbb{R}^m \to \mathbb{R}^L$ with $L > m$. If ξ is the solution, then

$$0 = \nabla\Big(\|f(\xi)\|_2^2\Big) = 2\big(f'(\xi)\big)^T f(\xi)$$

where $f'(y) \in \mathbb{R}^{L\times m}$ is the Jacobian of f. Hence, one could try to approximate ξ by applying Newton's method to the function $y \mapsto \big(f'(y)\big)^T f(y)$, but this would require the second derivative of f. Derivatives of f must be approximated by difference quotients, which requires many function evaluations (one evaluation of $f \longleftrightarrow$ solving Dupire once). Better alternative: Gauss-Newton method. Replace one nonlinear least-squares problem by many linear least-squares problems. (Similar to Newton's method, where solving one nonlinear equation is replaced by many linear equations.)

Gauss-Newton method:

Choose initial vector $y^{(0)} \in \mathbb{R}^m$.

For $k = 0, 1, 2, 3, \ldots$

Compute (or approximate) $f'(y^{(k)}) \in \mathbb{R}^{L\times m}$ and $f(y^{(k)}) \in \mathbb{R}^L$

Solve the linear least-squares problem

$$\|f(y^{(k)}) + f'(y^{(k)})\Delta y^{(k)}\|_2^2 = \min_{\Delta y^{(k)}}!$$

Let $y^{(k+1)} = y^{(k)} + \Delta y^{(k)}$

In case of the problem (2.12), no explicit formula for f is available. In this case, the derivative f' has to be approximated by difference quotients.
It can be shown that the sequence $\big(y^{(k)}\big)_k$ converges linearly to a minimizer ξ if $\|f(\xi)\|_2$ is small enough. If $\|f(\xi)\|_2$ is too large, however, the method may diverge.

Solving the linear sub-problems. In each step of the method a linear least-squares problem of the type

$$\|Ax - b\|_2^2 = \min_x, \qquad A \in \mathbb{R}^{L\times m}, \quad b \in \mathbb{R}^L, \quad L > m \tag{2.13}$$

has to be solved ($A \longleftrightarrow f'(y^{(k)})$, $x \longleftrightarrow \Delta y^{(k)}$, $b \longleftrightarrow -f(y^{(k)})$). If $\operatorname{rank}(A) = m$, then the solution can, in principle, be obtained via the QR decomposition of A; see lecture "Numerische Mathematik 1+2" for details. In our situation, however, A is typically "nearly rank deficient", i.e. some of the singular values of A are close to zero. Hence, it is advisable to solve the linear least-squares problem via the Moore-Penrose pseudoinverse. Assume that $A \in \mathbb{R}^{L\times m}$ with $\operatorname{rank}(A) = p < m$. Then, the solution of (2.13) is not unique. Let $A = U\Sigma V^T$ be the singular value decomposition of $A \in \mathbb{R}^{L\times m}$ with $L > m$, i.e.

$$U \in \mathbb{R}^{L\times L} \text{ orthogonal}, \quad V \in \mathbb{R}^{m\times m} \text{ orthogonal}, \quad \Sigma = \left[\frac{\widehat{\Sigma}}{0}\right] \in \mathbb{R}^{L\times m}$$

and

$$\widehat{\Sigma} = \begin{pmatrix} \sigma_1 & & & & & \\ & \ddots & & & & \\ & & \sigma_p & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix} \in \mathbb{R}^{m \times m}$$

Then, the Moore-Penrose-Pseudoinverse A^+ is given by $A^+ = V\Sigma^+ U^T$ with

$$\Sigma^+ = \left[\widehat{\Sigma}^+ \middle| \, 0 \right] \in \mathbb{R}^{m \times L}, \qquad \widehat{\Sigma}^+ = \begin{pmatrix} \sigma_1^{-1} & & & & & \\ & \ddots & & & & \\ & & \sigma_p^{-1} & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix} \in \mathbb{R}^{m \times m},$$

and $x_\star := A^+ b$ is the solution of (2.13) which has the smallest norm. Nearly rank deficient problems can be treated in the same way, after replacing singular values below a chosen tolerance by zero.

Matlab command: `pinv`

Chapter 3

Jump-diffusion processes and integro-differential equations

Up to now, the value of the underlying has been modelled by an SDE of the form

$$dS_t = f(t, S_t)dt + g(t, S_t)dW_t \tag{3.1}$$

If $f(t, S_t) = \mu S_t$ and $g(t, S_t) = \sigma S_t$ are constant, this yields the classical geometric Brownian motion $dS_t = \mu S_t dt + \sigma S_t dW_t$ with solution

$$S_t = S_0 \exp\left(\gamma t + \sigma W_t\right), \qquad \gamma := \mu - \frac{\sigma^2}{2}. \tag{3.2}$$

Real market data show, however, that prices sometimes change very quickly. Such a behaviour should be modelled by a discontinuous jump.

Ansatz: Replace (3.2) by

$$S_t = S_0 \exp\left(\gamma t + \sigma W_t + Z_t\right)$$

with a suitable jump process Z_t. This ansatz ensures posititity of the price process.

picture

Goals of this chapter:

- Develop a more realistic model for the price dynamics:
  Combine Wiener processes with jump processes.
- Derive the associated differential equation for $V(t, S)$:
  Black-Scholes equation with additional integral term.
- Construct numerical methods for this equation.

3.1 Jump-diffusion processes

Let $\{J_t,\ t \geq 0\}$ be a stochastic process with $J_t \in \mathbb{N}_0$ and $J_0 = 0$ with probability one. Assume that for every path of J_t there are times $0 < T_1 < T_2 < \ldots$ such that

$$J_t = n \qquad \Longleftrightarrow \qquad t \in [T_n, T_{n+1}).$$

At time T_n, the value of J_t jumps from $n-1$ to n.

picture

Let $\lambda > 0$ be the **intensity** of the process, i.e. assume that

$$\begin{aligned}
&\mathbb{P}(J_{t+\delta} - J_t = 1) = \delta\lambda + \mathcal{O}\big(\delta^2\big) && \text{one jump in } (t, t+\delta] \\
&\mathbb{P}(J_{t+\delta} - J_t = 0) = 1 - \delta\lambda + \mathcal{O}\big(\delta^2\big) && \text{no jump in } (t, t+\delta] \\
&\mathbb{P}(J_{t+\delta} - J_t > 1) = \mathcal{O}\big(\delta^2\big) && \text{more than one jump in } (t, t+\delta]
\end{aligned}$$

for sufficiently small $\delta > 0$. For $t > 0, \delta = \frac{t}{N}$ and sufficiently large $N \in \mathbb{N}$ it follows that

$$\begin{aligned}
\mathbb{P}(J_t = k) &= \binom{N}{k} (\delta\lambda)^k (1-\delta\lambda)^{N-k} + \mathcal{O}(N\delta^2) \\
&= \binom{N}{k} \left(\frac{t\lambda}{N}\right)^k \left(1 - \frac{t\lambda}{N}\right)^{N-k} + \mathcal{O}(t\delta),
\end{aligned}$$

and for $N \longrightarrow \infty$, $\delta = \frac{t}{N} \longrightarrow 0$ and fixed $t > 0$ we obtain

$$\begin{aligned}
\mathbb{P}(J_t = k) &= \lim_{N\longrightarrow\infty} \binom{N}{k} \left(\frac{t\lambda}{N}\right)^k \left(1 - \frac{t\lambda}{N}\right)^{N-k} \\
&= \lim_{N\longrightarrow\infty} \frac{N!}{(N-k)!} \frac{(t\lambda)^k}{k!} \frac{1}{N^k} \left(1 - \frac{t\lambda}{N}\right)^N \left(\frac{N}{N - t\lambda}\right)^k \\
&= \frac{(t\lambda)^k}{k!} \lim_{N\longrightarrow\infty} \underbrace{\frac{N!}{(N-k)!}}_{\sim N^k} \underbrace{\frac{1}{(N-t\lambda)^k}}_{\sim N^{-k}} \underbrace{\left(1 - \frac{t\lambda}{N}\right)^N}_{\longrightarrow e^{-t\lambda}} \\
&= \frac{(t\lambda)^k}{k!} e^{-t\lambda}.
\end{aligned}$$

Definition 3.1.1 (Poisson process) *A stochastic process J_t on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a* ***Poisson process*** *if*

- $J_0 = 0$,
- $J_t - J_s \in \mathbb{N}_0$ *and* $\mathbb{P}(J_t - J_s = k) = \dfrac{\lambda^k (t-s)^k}{k!} e^{-\lambda(t-s)}$ *for* $0 \le s \le t < \infty$, *and*
- *the increments* $J_{t_2} - J_{t_1}$ *and* $J_{t_4} - J_{t_3}$ *are independent for all* $0 \le t_1 < t_2 \le t_3 < t_4$.

Lemma 3.1.2 *A Poisson process has the following properties:*

1. J_t *is nondecreasing:* $J_t \ge J_s$ *for all* $t \ge s \ge 0$.
2. $\mathbb{E}(J_t) = \lambda t, \qquad \mathbb{V}(J_t) = \sum_{k=0}^{\infty} (\lambda t - k)^2 \mathbb{P}(J_t = k) = \lambda t.$
3. J_t *is a càdlàg process (càdlàg = continue à droite, limite à gauche): It is right-continuous with left limits, i.e.*

$$\lim_{s \searrow t} J_s = J_t, \qquad \lim_{s \nearrow t} J_s \text{ exists} \qquad \text{for all } t \ge 0.$$

4. *We have*

$$\mathbb{P}(J_t = n \mid J_s = n) = e^{-\lambda(t-s)} \qquad \text{for all } t > s,\ n \in \mathbb{N}_0.$$

As a consequence, the waiting times $\tau_n := T_n - T_{n-1}$ *between successive jumps are independent and exponentially distributed with parameter* λ, *i.e. its probability density is* $\lambda e^{-\lambda x}$, *and its probability distribution function is*

$$\mathbb{P}(\tau_n \le t) = 1 - e^{-\lambda t}.$$

Notation: $\tau_n \sim \exp(\lambda)$.

Proof: Properties 1, 2 and 3 follow directly from the definition. In order to prove property 4, we calculate for $t > s$ and $m \ge n$

$$\begin{aligned}
\mathbb{P}(J_s = n, J_t = m) &= \mathbb{P}(J_s = n, J_t - J_s = m - n) \\
&= \mathbb{P}(J_s = n)\mathbb{P}(J_t - J_s = m - n) && \text{(indep. incr.)} \\
&= \frac{(\lambda s)^n}{n!} e^{-\lambda s} \frac{(\lambda(t-s))^{m-n}}{(m-n)!} e^{-\lambda(t-s)} \\
&= \lambda^m \frac{s^n}{n!} \frac{(t-s)^{m-n}}{(m-n)!} e^{-\lambda t} \\
&= \frac{(\lambda t)^m}{m!} e^{-\lambda t} \cdot \binom{m}{n} \left(\frac{s}{t}\right)^n \left(1 - \frac{s}{t}\right)^{m-n}.
\end{aligned}$$

Hence, the probability that there is no jump in the interval $(s, t]$ is

$$\begin{aligned}
\mathbb{P}(J_t = n \mid J_s = n) &= \frac{\mathbb{P}(J_s = n, J_t = n)}{\mathbb{P}(J_s = n)} \\
&= \frac{(\lambda t)^n}{n!} e^{-\lambda t} \left(\frac{s}{t}\right)^n \cdot \left(\frac{(\lambda s)^n}{n!} e^{-\lambda s}\right)^{-1} = e^{-\lambda(t-s)}.
\end{aligned}$$

The other assertion follows from

$$\mathbb{P}(\tau_n \leq t) = \mathbb{P}(J_{T_{n-1}+t} > J_{T_{n-1}}) = 1 - \mathbb{P}(J_{T_{n-1}+t} = J_{T_{n-1}}) = 1 - e^{-\lambda t}. \qquad \blacksquare$$

The jump times of the Poisson process are random, but the height of each jump is +1. Too simple. Next step: Construct a process with random height of the jumps.

Definition 3.1.3 (Compound Poisson process) *Let J_t be a Poisson process with intensity $\lambda > 0$. If $\{Y_t,\ t \geq 0\}$ are independent and identically distributed random variables with density ϕ, then*

$$X_t = \sum_{j=1}^{J_t} Y_{T_j}$$

is called a ***compound Poisson process*** *with intensity $\lambda > 0$ and jump size density ϕ.*

picture

The compound Poisson process is a pure jump process. For applications in finance, a jump diffusion process is often more suitable.

Definition 3.1.4 (Jump diffusion process) *Let $\sum_{j=1}^{J_t} Y_{T_j}$ be a compound Poisson process, and let W_t be the Wiener process. If $\gamma, \sigma \in \mathbb{R}$ with $\sigma \neq 0$, then*

$$X_t = \gamma t + \sigma W_t + \sum_{j=1}^{J_t} Y_{T_j} \tag{3.3}$$

is called a ***jump-diffusion process****.*

Idea: We can use

$$S_t = S_0 e^{X_t} = S_0 \exp\left(\gamma t + \sigma W_t + \sum_{j=1}^{J_t} Y_{T_j}\right)$$

as a new model for the price process.

3.2 Jump-diffusion models in finance

(a) Merton's model

R. Merton 1976

In this model, the jump size is assumed to be proportional to the current price, i.e.

$$S_{T_j} = q_{T_j} S_{T_j^-}, \qquad j \in \mathbb{N},$$

where $S_{T_j^-} = \lim_{t \nearrow T_j} S_t$ is the value "right before the jump". Hence, the size of the jump is $S_{T_j} - S_{T_j^-} = (q_{T_j} - 1) S_{T_j^-}$. The factors $\{q_t,\ t \geq 0\}$ are independent log-normal random variables, i.e. $q_t = \exp(Y_t)$ with normal $Y_t \sim \mathcal{N}(\alpha, \beta)$, $\mathbb{E}(Y_t) = \alpha$, $\mathbb{V}(Y_t) = \mathbb{E}((Y_t - \alpha)^2) = \beta$ and $\mathbb{E}(q_t) = e^{\alpha + \beta/2}$ for all t; cf. Definition 3.1.2 in part I.
Between two jumps, S_t is modelled by the geometric Brownian motion

$$dS_t = \mu S_t \, dt + \sigma S_t \, dW_t, \qquad t \in [T_j, T_{j+1}). \tag{3.4}$$

Combining both parts yields

$$S_t = S_0 + \int_0^t \mu S_\theta \, d\theta + \int_0^t \sigma S_\theta \, dW_\theta + \sum_{j=1}^{J_t} (q_{T_j} - 1) S_{T_j^-} \, . \tag{3.5}$$

Short-hand notation:

$$dS_t = \mu S_t \, dt + \sigma S_t \, dW_t + (q_t - 1) S_{t^-} \, dJ_t. \tag{3.6}$$

The three processes W_t, q_t and J_t are independent.

(b) Explicit solution

Before the first jump, S_t simply evolves according to the SDE (3.4). Thus, for $t \in [0, T_1)$ the solution of (3.5) is the standard geometric Brownian motion

$$S_t = S_0 \exp\left(\gamma t + \sigma W_t\right), \qquad t \in [0, T_1)$$

with $\gamma := \mu - \frac{\sigma^2}{2}$. At the jump time T_1, the price jumps from $S_{T_1^-}$ to $S_{T_1} = \exp(Y_{T_1}) S_{T_1^-}$. Then, in the interval $t \in [T_1, T_2)$, S_t evolves again according to (3.4), but now with initial value S_{T_1}. Hence,

$$S_t = S_{T_1} \exp\Big(\gamma(t - T_1) + \sigma(W_t - W_{T_1})\Big), \qquad t \in [T_1, T_2),$$

and with

$$S_{T_1} = S_{T_1^-} \exp(Y_{T_1}) = S_0 \exp\left(\gamma T_1 + \sigma W_{T_1}\right) \exp(Y_{T_1}),$$

we obtain

$$S_t = S_0 \exp\left(\gamma t + \sigma W_t\right) \exp(Y_{T_1}), \qquad t \in [T_1, T_2).$$

Repeating this procedure yields the representation

$$S_t = S_0 \exp\left(\gamma t + \sigma W_t\right) \prod_{j=1}^{J_t} \exp(Y_{T_j}), \qquad t \geq 0 \tag{3.7}$$

for the solution of (3.5). Hence, $S_t = S_0 \exp(X_t)$ is the exponential of the jump-diffusion process

$$X_t = \gamma t + \sigma W_t + \sum_{j=1}^{J_t} Y_{T_j}.$$

(c) Choosing μ under the risk-neutral measure

Under the risk-neutral measure, $e^{-rt}S_t$ must be a martingale, i.e.

$$\mathbb{E}\left(e^{-rt}S_t \mid \mathcal{F}_s\right) = e^{-rs}S_s \qquad \text{for all } t \geq s. \tag{3.8}$$

where $\{\mathcal{F}_t, t \geq 0\}$ is a filtration such that the process $e^{-rt}S_t$ is adapted to $\{\mathcal{F}_t, t \geq 0\}$. For the geometric Brownian motion $S_t = S_0 \exp\left(\gamma t + \sigma W_t\right)$ with $\gamma := \mu - \frac{\sigma^2}{2}$, this is true if $\mu = r$; see chapter 3 of part 1.

Question: What is the risk-neutral μ for the exponential jump-diffusion process (3.7)?

Reminder: If X is $\mathcal{F}_s$-measurable, then

$$\mathbb{E}(XY \mid \mathcal{F}_s) = X\,\mathbb{E}(Y \mid \mathcal{F}_s) \qquad \text{and} \qquad \mathbb{E}(X \mid \mathcal{F}_s) = X. \tag{$\star$}$$

If X is independent of $\mathcal{F}_s$, then

$$\mathbb{E}(X \mid \mathcal{F}_s) = \mathbb{E}(X). \tag{$\star\star$}$$

We decompose

$$\begin{aligned} e^{-rt}S_t &= e^{-rt} \exp\left(\left(\mu - \tfrac{\sigma^2}{2}\right) t + \sigma W_t + \sum_{j=1}^{J_t} Y_{T_j}\right) S_0 \\ &= \underbrace{e^{-r(t-s)} \exp\left(\left(\mu - \tfrac{\sigma^2}{2}\right)(t-s) + \sigma(W_t - W_s) + \sum_{j=J_s+1}^{J_t} Y_{T_j}\right)}_{=:Z(t,s)} e^{-rs}S_s. \end{aligned}$$

We know that $e^{-rs}S_s$ is $\mathcal{F}_s$-measurable because $e^{-rt}S_t$ is adapted to $\{\mathcal{F}_t, t \geq 0\}$. Applying $(\star)$ gives

$$\mathbb{E}\left(e^{-rt}S_t \mid \mathcal{F}_s\right) = \mathbb{E}\left(Z(t,s) \mid \mathcal{F}_s\right) e^{-rs}S_s.$$

This means that $e^{-rt}S_t$ can only be a martingale if $\mathbb{E}\left(Z(t,s) \mid \mathcal{F}_s\right) = 1$ for all $t \geq 0$, i.e. if

$$\begin{aligned} 1 &= \mathbb{E}\left(\exp\left(\left(-r+\mu-\tfrac{\sigma^2}{2}\right)(t-s) + \sigma(W_t - W_s) + \sum_{j=J_s+1}^{J_t} Y_{T_j}\right) \mid \mathcal{F}_s\right) \\ &= \exp\Big(\Big(-r+\mu-\tfrac{\sigma^2}{2}\Big)(t-s)\Big)\,\mathbb{E}\left(\exp\left(\sigma(W_t-W_s) + \sum_{j=J_s+1}^{J_t} Y_{T_j}\right)\right). \qquad (3.9) \\ &= \exp\Big(\Big(-r+\mu-\tfrac{\sigma^2}{2}\Big)(t-s)\Big)\,\mathbb{E}\Big(\exp\big(\sigma(W_t-W_s)\big)\Big)\mathbb{E}\left(\prod_{j=J_s+1}^{J_t} q_{T_j}\right) \qquad (3.10) \end{aligned}$$

In (3.9), we have used that W_t and J_t have independent increments such that $(\star\star)$ applies. Equality (3.10) follows from the fact that the three processes W_t, J_t, q_t are independent, and that $q_{T_j} = \exp(Y_{T_j})$. From Lemma 3.1.1 in part I of the lecture we know that

$$\mathbb{E}\Big(\exp\big(\sigma(W_t-W_s)\big)\Big) = \exp\Big(\tfrac{\sigma^2}{2}(t-s)\Big).$$

For the last factor in (3.10) we use that $J_t - J_s = J_{t-s}$ and that $\{q_t,\ t \geq 0\}$ are independent and identically distributed with constant expectation $\mathbb{E}(q_t) = e^{\alpha+\beta/2}$:

$$\begin{aligned} \mathbb{E}\left(\prod_{j=J_s+1}^{J_t} q_{T_j}\right) &= \mathbb{E}\left(\prod_{j=1}^{J_{t-s}} q_{T_j}\right) \\ &= \sum_{k=0}^{\infty} \mathbb{P}(J_{t-s}=k)\,\mathbb{E}\left(\prod_{j=1}^{k} q_{T_j}\right) \\ &= \sum_{k=0}^{\infty} e^{-\lambda(t-s)}\frac{\lambda^k(t-s)^k}{k!}\mathbb{E}(q_t)^k \\ &= \exp\Big(\lambda(t-s)\big(\mathbb{E}(q_t)-1\big)\Big). \end{aligned}$$

Substituting this into (3.10) taking the logarithm and dividing by $(t-s)$ yields the condition

$$-r+\mu+\lambda(\mathbb{E}(q_t)-1) = 0.$$

Hence, the risk-neutral μ is

$$\mu = r - \lambda\big(\mathbb{E}(q_t)-1\big).$$

(d) Numerical simulation of a single path of Merton's model

Input: parameters μ, σ, nonnegative parameters $\alpha, \beta, \lambda \geq 0$, initial value $S_0 > 0$, maturity $T > 0$

1. Let $\mu = r - \lambda(\mathbb{E}(q_t) - 1) = r - \lambda\left(e^{\alpha+\beta/2} - 1\right)$ and $\gamma = \mu - \frac{\sigma^2}{2}$.
   Generate a path of $\tilde{X}_t = \gamma t + \sigma W_t$ on the interval $[0, T]$.
2. Simulate a path of the Poisson process J_t with intensity λ on the interval $[0, T]$:
   (a) Set $T_0 = 0$, $j = 0$.
   (b) Generate a random number $\tau_j \sim \exp(\lambda)$, i.e. exponentially distributed with parameter λ (cf. 4. in Lemma 3.1.2).
   (c) Set $T_{j+1} = \min\{T_j + \tau_j, T\}$.
   (d) Set $J_t = j$ for all $t \in [T_j, T_{j+1})$.
   (e) If $T_{j+1} < T$: Set $j = j + 1$ and go to step (b).
   Else: Set $J_T = j$ and go to step 3.
3. For $j = 1, \ldots, J_T$
   (a) Simulate independent random variables $Z_j \sim \mathcal{N}(0, 1)$ with standard normal distribution ($j = 1, \ldots, J_T$).
   (b) Set $q_{T_j} = \exp(\alpha + \sqrt{\beta} Z_j)$ $\quad (\Longrightarrow \ln(q_{T_j}) \sim \mathcal{N}(\alpha, \beta)$, i.e. q_{T_j} is log-normal).
4. Set

$$S_t = S_0 \exp(\tilde{X}_t) \prod_{j=1}^{J_t} q_{T_j}.$$

Remarks. (1) Simulating waiting times $\tau_j \sim \exp(\lambda)$ is easy: If $U \sim \mathcal{U}(0, 1)$ is uniformly distributed, then $\tau_j := -\ln(1 - U)/\lambda \sim \exp(\lambda)$; cf. page 33 in [KKK10].
(2) Of course, the paths of W_t, $\tilde{X}_t$, J_t can only be computed in finitely many points $0 < t_1 < t_2 < \ldots < t_N = T$. The discretization points t_n must not be confused with the jump times T_j. There are two possibilities to choose the discretization:

1. Choose $N \in \mathbb{N}$, set $\tau = T/N$ and $t_n = n\tau$ as usual. In this case, the jump times usually fall between the discretization points.
2. If it is important to choose the discretization in such a way that the jump times T_j belong to the set of discretization points, then one can first determine T_1 and discretize $[0, T_1]$ with $N_1 \in \mathbb{N}$ steps and step-size $\tau_1 = T_1/N_1$. Then, T_2 is determined, and $[T_1, T_2]$ is discretized with $N_2 \in \mathbb{N}$ steps and step-size $\tau_2 = (T_2 - T_1)/N_2$, and so on.

(e) More general jump-diffusion models

Instead of constant drift and diffusion, one may assume that $\mu = \mu(t)$ and $\sigma = \sigma(t, S)$. In this case, the geometric Brownian motion (3.4) is replaced by the SDE

$$dS_t = \mu(t) S_t \, dt + \sigma(t, S_t) S_t \, dW_t, \qquad t \in [T_j, T_{j+1}).$$

Moreover, a different distribution can be used for the q_{T_j}, but q_{T_j} must be nonnegative, because otherwise the price of the underlying can have negative values.

Combining jumps and SDE yields the jump-diffusion process

$$S_t = S_0 + \int_0^t \mu(\theta) S_\theta \, d\theta + \int_0^t \sigma(\theta, S_\theta) S_\theta \, dW_\theta + \sum_{j=1}^{J_t} (q_{T_j} - 1) S_{T_j^-}. \tag{JDP}$$

Short-hand notation:

$$dS_t = \mu(t) S_t \, dt + \sigma(t, S_t) S_t \, dW_t + (q_t - 1) S_{t^-} \, dJ_t. \tag{JDP'}$$

As before, the three processes W_t, q_t and J_t are independent.

The model can be simulated in a similar way as Merton's model. Since no explicit formula for the SDE part is available now, a numerical method such as Euler-Maruyama has to be used for the SDE part.

3.3 From jump-diffusion processes to integro-differential equations

Question: How can we compute the value of an option if the price of the underlying is modelled by (JDP')?

Goal: Derive a counterpart of the Black-Scholes equation for the jump-diffusion model.

Reminder: Derivation of the classical Black-Scholes equation in part I, section 3.2 (SDE instead of JDP, i.e $\lambda = 0$, $J_t \equiv 0$)

- Replication strategy: Consider a portfolio containing $a_t \in \mathbb{R}$ underlyings and $b_t \in \mathbb{R}$ bonds such that

  $$V(t, S_t) = a_t S_t + b_t B_t$$

  where S_t and B_t are the values of the underlying and the bond, respectively.
- Assumption: Portfolio is self-financing.
- Apply the Itô formula, replace dS_t by SDE, equate terms with dW_t and dt, respectively.
  $\Longrightarrow$ Black-Scholes equation

If the underlying is modelled by (JDP'), however, then the market is not complete, i.e. not every option can be replicated.[1] Therefore, the above strategy has to be modified.

Consider a portfolio $P_t = a_t S_t + b_t B_t$ where S_t evolves according to (JDP') and where $dB_t = r(t) B_t \, dt$.

[1] The assumption that the market is not complete is more realistic, because options are redundant in a complete market. However, pricing is more difficult in incomplete markets, because if an equivalent martingale measure exists, then it is in general not unique.

Assume that the portfolio is self-financing: no cash inflow or outflow, i.e. buying an item must be financed by selling another one. Consequence:

$$dP_t = a_t\, dS_t + b_t\, dB_t.$$

Substituting (JDP') and $dB_t = r(t)B_t dt$ yields

$$dP_t = a_t\Big(\mu(t)S_t\, dt + \sigma(t,S_t)S_t\, dW_t + (q_t - 1)S_{t^-}\, dJ_t\Big) + b_t r(t)B_t\, dt. \tag{3.11}$$

On the other hand, we can apply the Itô formula to $V(t,S_t)$ on each interval $[T_n, T_{n+1})$ between two jumps and afterwards add the jumps of the option price:

$$\begin{aligned} dV(t,S_t) = &\left(\partial_t V(t,S_t) + \mu(t)S_t\partial_S V(t,S_t) + \frac{\sigma^2(t,S_t)}{2}S_t^2\partial_S^2 V(t,S_t)\right) dt \\ &+ \sigma(t,S_t)S_t\partial_S V(t,S_t)\, dW_t + \Delta V\, dJ_t \end{aligned} \tag{3.12}$$

with $\Delta V = V(t,S_t) - V(t,S_{t^-}) = V(t,q_tS_{t^-}) - V(t,S_{t^-})$. In the absence of jumps, P_t is supposed to replicate $V(t,S_t)$. Therefore, the dW_t terms in (3.11) and (3.12) must be the same:

$$a_t\sigma(t,S_t)S_t \stackrel{!}{=} \sigma(t,S_t)S_t\partial_S V(t,S_t) \qquad \Longrightarrow \qquad a_t = \partial_S V(t,S_t).$$

Because of the jump process, we cannot determine b_t in such a way that P_t equals $V(t,S)$. Therefore, we impose the weaker condition

$$\mathbb{E}\big(V(t,S_t)\big) \stackrel{!}{=} \mathbb{E}(P_t), \qquad \text{and hence} \qquad \mathbb{E}\big(dV(t,S_t)\big) \stackrel{!}{=} \mathbb{E}(dP_t)$$

Substituting (3.11) and (3.12) yields

$$\begin{aligned} &\mathbb{E}\Big(a_t(q_t-1)S_{t^-}\, dJ_t\Big) + r(t)\mathbb{E}(b_t)B_t\, dt \\ &\stackrel{!}{=} \mathbb{E}\left(\partial_t V(t,S_t) + \frac{\sigma^2(t,S_t)}{2}S_t^2\partial_S^2 V(t,S_t)\right) dt + \mathbb{E}(\Delta V\, dJ_t) \end{aligned}$$

because the dW_t terms and the terms $a_t\mu(t)S_t = \mu(t)S_t\partial_S V(t,S_t)$ cancel. With

- $c := \mathbb{E}(q_t - 1)$
- $\mathbb{E}\left(a_t(q_t-1)S_{t^-}dJ_t\right) = \mathbb{E}\big(S_{t^-}\partial_S V(t,S_t)\big)\underbrace{\mathbb{E}(q_t-1)}_{=:c}\cdot\underbrace{\mathbb{E}(dJ_t)}_{\lambda dt} = c\lambda\mathbb{E}\big(S_{t^-}\partial_S V(t,S_t)\big)\, dt$
- $\mathbb{E}\left(\Delta V\, dJ_t\right) = \mathbb{E}(\Delta V)\cdot\underbrace{\mathbb{E}(dJ_t)}_{\lambda dt} = \Big[\mathbb{E}\big(V(t,q_tS_{t^-})\big) - \mathbb{E}\big(V(t,S_{t^-})\big)\Big]\,\lambda dt$
- $\mathbb{E}(b_t)B_t = \mathbb{E}(b_tB_t) = \mathbb{E}(P_t) - \mathbb{E}(a_tS_t) = \mathbb{E}\big(V(t,S_t)\big) - \mathbb{E}\big(S_t\partial_S V(t,S_t)\big)$

it follows that

$$\begin{aligned} &c\lambda\mathbb{E}\big(S_{t^-}\partial_S V(t,S_t)\big)\, dt + r(t)\Big(\underbrace{\mathbb{E}\big(V(t,S_t)\big) - \mathbb{E}\big(S_t\partial_S V(t,S_t)\big)}_{=\mathbb{E}(b_t)B_t}\Big)\, dt \\ &\stackrel{!}{=} \mathbb{E}\left(\partial_t V(t,S_t) + \frac{\sigma^2(t,S_t)}{2}S_t^2\partial_S^2 V(t,S_t)\right) dt + \Big[\mathbb{E}\big(V(t,q_tS_{t^-})\big) - \mathbb{E}\big(V(t,S_{t^-})\big)\Big]\,\lambda dt. \end{aligned}$$

Now we assume that $V = V(t, S)$ is a deterministic function and consider $S = S_t$ as a *variable*, i.e. $\mathbb{E}(S_t V(t, S_t)) = SV(t, S)$, etc. We thus obtain

$$cS\lambda\partial_S V + r(t)(V - S\partial_S V) = \partial_t V + \frac{\sigma^2(t,S)}{2} S^2 \partial_S^2 V + \lambda \mathbb{E}\left(V(t, q_t S)\right) - \lambda V$$

If ϕ is the density of q_t, then

$$\mathbb{E}\left(V(t, q_t S)\right) = \int_0^\infty V(t, zS)\phi(z)\, dz,$$

and we obtain the **partial integro-differential equation (PIDE)**

$$\begin{aligned}\partial_t V(t,S) + \frac{\sigma^2(t,S)}{2} S^2 \partial_S^2 V(t,S) + \big(r(t) - c\lambda\big) S\partial_S V(t,S) \\ -\big(r(t) + \lambda\big) V(t,S) + \lambda \int_0^\infty V(t, zS)\phi(z)\, dz = 0\end{aligned}$$

for $t \in [0, T]$, $S > 0$ and with $c = \mathbb{E}(q_t - 1) = \int_0^\infty z\phi(z)dz - 1$.

Remarks.

1. This is a PDE with an additional integral term.
2. For $\lambda = 0$ we obtain the Black-Scholes equation (with time-dependent interest rate and local volatility).
3. Our derivation is, of course, very sketchy. A rigorous proof can be found in [CT04, chapter 12].

Example: European call in the Merton model

In the Merton model $r(t) = r > 0$ and $\sigma(t, S_t) = \sigma > 0$ are constant, and ϕ is the density of the log-normal distribution:

$$\phi(z) = \phi(z, \alpha, \beta) = \begin{cases} \frac{1}{\sqrt{2\pi\beta} z} \exp\left(-\frac{(\ln(z)-\alpha)^2}{2\beta}\right) & \text{for } z > 0 \\ 0 & \text{else} \end{cases}$$

(cf. part I). This means that jumps in the log-price $\ln(S_t)$ have a normal distribution. For this model, we have

$$c = \int_0^\infty z\phi(z)dz - 1 = \exp\left(\alpha + \frac{\beta}{2}\right) - 1. \tag{3.13}$$

3.4 Numerical approximation

(a) Change of variables and localization

Let $V(t,S)$ be the value of a European option with an underlying modelled by the jump-diffusion process (JDP'). For simplicity, we assume that σ and r are constant. As before, J_t is a Poisson process with intensity λ and ϕ is the density of q_t, but we do neither assume nor exclude that ϕ is log-normal. Hence, $V(t,S)$ solves the PIDE

$$\begin{aligned}\partial_t V(t,S) + \frac{\sigma^2}{2}S^2\partial_S^2 V(t,S) + \big(r - c\lambda\big)S\partial_S V(t,S) \\ -\big(r+\lambda\big)V(t,S) + \lambda\int_0^\infty V(t,zS)\phi(z)\,dz = 0\end{aligned} \tag{3.14}$$

with $c = \mathbb{E}(q_t - 1) = \int_0^\infty z\phi(z)dz - 1$. We change variables as follows:

$$\begin{aligned} x &= \ln(S/S_0) = \ln(S) - \ln(S_0), & S &= e^x S_0, \\ \theta &= T - t, & u(\theta,x) &= e^{r(T-t)}V(t,S). \end{aligned}$$

With similar computations as in Chapter 3 of part 1 we obtain:

- $\partial_\theta u(\theta,x) = -e^{r(T-t)}\partial_t V(t,S) + ru(\theta,x)$
- $\partial_x u(\theta,x) = Se^{r(T-t)}\partial_S V(t,S)$
- $\partial_x^2 u(\theta,x) = \left(S\partial_S V(t,S) + S^2\partial_S^2 V(t,S)\right)e^{r(T-t)}$.

In order to transform the integral part, we let

$$\widehat{\phi}(y) = e^y\phi(e^y).$$

Substituting $z = e^y$ and $dz = e^y dy$ yields

$$\begin{aligned}\int_0^\infty V(t,zS)\phi(z)\,dz &= \int_{-\infty}^\infty V\left(t, e^{x+y}S_0\right)\phi(e^y)e^y dy \\ &= e^{-r(T-t)}\int_{-\infty}^\infty u(\theta,x+y)\widehat{\phi}(y)\,dy.\end{aligned}$$

By replacing these formulas in (3.14), we obtain

$$\begin{aligned}\partial_\theta u(\theta,x) = \mathcal{L}u(\theta,x) := \frac{\sigma^2}{2}\left(\partial_x^2 u(\theta,x) - \partial_x u(\theta,x)\right) + (r - c\lambda)\partial_x u(\theta,x) - \lambda u(\theta,x) \\ + \lambda\int_{-\infty}^\infty u(\theta,x+y)\widehat{\phi}(y)\,dy\end{aligned} \tag{3.15}$$

for $x \in (-\infty,\infty)$, $\theta \in (0,T]$ and initial condition

$$u(0,x) = u_0(x) = V(T,e^x S_0)$$

Our goal is to approximate the solution $u(\theta, x)$ numerically. Of course, such an approximation can only be computed on a bounded (but sufficiently large) interval $[x_{\min}, x_{\max}]$. As usual, boundary conditions have to be imposed at the artificial boundaries. In order to evaluate the integral term in (3.15), however, we also have to impose the value of the solution *beyond* the boundaries, i.e. we need a condition of the type

$$u(\theta, x) = g(\theta, x) \qquad \text{for } x \notin (x_{\min}, x_{\max}) \text{ and all } \theta \in [0, T]$$

with some suitably chosen function $g(\theta, x)$. For European calls/puts, a simple choice is the payoff function $g(\theta, x) = u_0(x)$.

Hence, we will consider the problem

$$\partial_\theta u(\theta, x) = \mathcal{L}u(\theta, x) \qquad \theta \in (0, T], \quad x \in (x_{\min}, x_{\max}) \tag{3.16a}$$

$$u(0, x) = u_0(x) \qquad x \in \mathbb{R} \tag{3.16b}$$

$$u(\theta, x) = g(\theta, x) \qquad \theta \in [0, T], \quad x \notin (x_{\min}, x_{\max}) \tag{3.16c}$$

Proposition 4.1 in [CV05] states an error bound for the truncation error under the condition that $u_0(x)$ is bounded. This applies to a put, but not to a call. For calls, the put-call parity can be used.

(b) Space and time discretization

1. Split the operator $\mathcal{L}$ defined in (3.15) into the differential part and the integral part: $\mathcal{L} = \mathcal{A} + \mathcal{B}$ with

$$\begin{aligned} \mathcal{A}u(\theta, x) &= \frac{\sigma^2}{2}\partial_x^2 u(\theta, x) + \Big(r - c\lambda - \frac{\sigma^2}{2}\Big)\partial_x u(\theta, x) - \lambda u(\theta, x) \\ \mathcal{B}u(\theta, x) &= \lambda \int_{-\infty}^{\infty} u(\theta, x + y)\widehat{\phi}(y)\, dy \end{aligned} \tag{3.17}$$

2. Choose $1 < m \in \mathbb{N}$, define mesh-size $h = (x_{\max} - x_{\min})/m$ and mesh points $x_k = x_{\min} + kh$.
   Our goal is to compute approximations

$$v_k(\theta) \approx u(\theta, x_k), \quad k \in \{1, \dots, m-1\}, \qquad v(\theta) = \big(v_1(\theta), \dots, v_{m-1}(\theta)\big)^T.$$

   For $k \notin \{1, \dots, m-1\}$ we set $u(\theta, x_k) = g(\theta, x_k)$ according to (3.16c). Of course, we can only consider finitely many points. Hence, we choose $\widehat{m} \in \mathbb{N}$, let $y_{\min} = x_{-\widehat{m}}$, $y_{\max} = x_{m+\widehat{m}}$ and let

$$u(\theta, x_k) = g(\theta, x_k) \qquad \text{for } k = -\widehat{m}, \dots, 0 \text{ and } k = m, \dots, m + \widehat{m}.$$

   These values are compiled in the vectors

$$g^L(\theta) = \begin{pmatrix} g(\theta, x_{-\widehat{m}}) \\ \vdots \\ g(\theta, x_0) \end{pmatrix} \in \mathbb{R}^{\widehat{m}+1}, \qquad g^R(\theta) = \begin{pmatrix} g(\theta, x_m) \\ \vdots \\ g(\theta, x_{m+\widehat{m}}) \end{pmatrix} \in \mathbb{R}^{\widehat{m}+1}$$

picture

3. Approximate the integral $\mathcal{B}u(\theta, x_j)$:

$$\begin{aligned}
\mathcal{B}u(\theta, x_j) &= \lambda \int_{-\infty}^{\infty} u(\theta, x_j + y)\widehat{\phi}(y)\, dy \qquad (\xi = x_j + y) \\
&= \lambda \int_{-\infty}^{\infty} u(\theta, \xi)\widehat{\phi}(\xi - x_j)\, d\xi \\
&\approx \lambda \int_{y_{\min}-h/2}^{y_{\max}+h/2} u(\theta, \xi)\widehat{\phi}(\xi - x_j)\, d\xi \\
&= \lambda \sum_{k=-\widehat{m}}^{m+\widehat{m}} \int_{x_k-h/2}^{x_k+h/2} u(\theta, \xi)\widehat{\phi}(\xi - x_j)\, d\xi \\
&\approx \sum_{k=-\widehat{m}}^{m+\widehat{m}} u(\theta, x_k) \underbrace{\lambda h \widehat{\phi}(x_k - x_j)}_{=:\nu_{jk}}.
\end{aligned}$$

Warning: If $x_{\min} \notin h\mathbb{Z}$, then the sum or difference of two grid points is *not* a grid point:

$$x_k - x_j = x_{\min} + kh - (x_{\min} + jh) = (k - j)h \notin x_{\min} + h\mathbb{Z}.$$

Hence, the density $\widehat{\phi}(\cdot)$ has (possibly) to be evaluated on a *different* grid than the function $u(\theta, \cdot)$.

Matrix-vector notation: Define

$$B := \begin{pmatrix} \nu_{1,-\widehat{m}} & \nu_{1,-\widehat{m}+1} & \cdots & \cdots & \nu_{1,m+\widehat{m}} \\ \nu_{2,-\widehat{m}} & \nu_{2,-\widehat{m}+1} & \cdots & \cdots & \nu_{2,m+\widehat{m}} \\ \vdots & & \ddots & & \vdots \\ \nu_{m-1,-\widehat{m}} & \cdots & \cdots & \cdots & \nu_{m-1,m+\widehat{m}} \end{pmatrix} \in \mathbb{R}^{(m-1)\times(m+1+2\widehat{m})}$$

and decompose this matrix into blocks:

$$B = \left[\begin{array}{c|c|c} B_{-1} & B_0 & B_1 \\ (m-1) \times (\widehat{m}+1) & (m-1) \times (m-1) & (m-1) \times (\widehat{m}+1) \end{array}\right]$$

Then, we have

$$\left[\sum_{k=-\widehat{m}}^{m+\widehat{m}} u(\theta, x_k)\nu_{jk}\right]_{j=1,\dots,m-1} \approx B_{-1}g^L(\theta) + B_0 v(\theta) + B_1 g^R(\theta)$$

4. Approximate $\partial_x u$ and $\partial_x^2 u$ by finite differences:

$$\partial_x^2 u(\theta, x_k) \approx \frac{u(\theta, x_{k+1}) - 2u(\theta, x_k) + u(\theta, x_{k-1})}{h^2}$$
$$\partial_x u(\theta, x_k) \approx \frac{u(\theta, x_{k+1}) - u(\theta, x_{k-1})}{2h}$$

It can be shown that the upwind discretization

$$\partial_x u(\theta, x_k) \approx \begin{cases} \dfrac{u(\theta, x_{k+1}) - u(\theta, x_k)}{h} & \text{if } r - c\lambda - \dfrac{\sigma^2}{2} > 0 \\ \dfrac{u(\theta, x_k) - u(\theta, x_{k-1})}{h} & \text{if } r - c\lambda - \dfrac{\sigma^2}{2} \leq 0 \end{cases}$$

is more stable if σ is so small that the dominating term in (3.15) is

$$\Big(r - c\lambda - \frac{\sigma^2}{2}\Big)\partial_x u(\theta, x).$$

This approximation, however, has only order one. Without loss of generality, we assume henceforth that $r - c\lambda - \frac{\sigma^2}{2} > 0$ and use the corresponding upwind discretization.
Matrix-vector notation: Define the matrices

$$A_1 := \frac{1}{h}\begin{pmatrix} -1 & 1 & & \\ 0 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & 0 & -1 \end{pmatrix} \in \mathbb{R}^{(m-1)\times(m-1)}$$
$$A_2 := \frac{1}{h^2}\begin{pmatrix} -2 & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & -2 \end{pmatrix} \in \mathbb{R}^{(m-1)\times(m-1)}.$$

5. After truncation and semi-discretization in space, we obtain the ordinary differential equation

$$\begin{aligned} \dot{v}(\theta) = \frac{\sigma^2}{2}A_2 v(\theta) + \left(r - c\lambda - \frac{\sigma^2}{2}\right) A_1 v(\theta) - \lambda v(\theta) \\ + B_{-1}g^L(\theta) + B_0 v(\theta) + B_1 g^R(\theta) + a(\theta) \end{aligned} \tag{3.18}$$

as an approximation for (3.15). The vector $a(\theta) \in \mathbb{R}^{m-1}$ corrects the boundary data missing in the finite differences:

$$a_k(\theta) = \begin{cases} \dfrac{\sigma^2}{2h^2} g(\theta, x_0) & \text{if } k = 1 \\ \dfrac{\sigma^2}{2h^2} g(\theta, x_m) + \left(r - c\lambda - \dfrac{\sigma^2}{2} \right) \dfrac{g(\theta, x_m)}{h} & \text{if } k = m-1 \\ 0 & \text{else.} \end{cases}$$

After defining

$$A := \frac{\sigma^2}{2} A_2 - \left(\frac{\sigma^2}{2} - r + c\lambda \right) A_1 - \lambda I$$
$$b(\theta) := B_{-1} g^L(\theta) + B_1 g^R(\theta)$$

(3.18) simplifies to

$$\dot{v}(\theta) = Av(\theta) + B_0 v(\theta) + a(\theta) + b(\theta) =: F(\theta, v) \tag{3.19}$$

6. Time-discretization: Choose $N \in \mathbb{N}$, let $\tau = T/N$ and $\theta_n = n\tau$.

   Goal: Approximate $v(\theta_n) \approx w_n \in \mathbb{R}^{m-1}$

   As we have seen in part I (section 7.3), explicit methods are not even suitable for the standard Black-Scholes equation ($\lambda = 0$). We consider A-stable implicit methods, namely the implicit Euler method

$$\begin{aligned} & & w_{n+1} &= w_n + \tau F(\theta_{n+1}, w_{n+1}) \\ \iff & & \big(I - \tau(A + B_0)\big) w_{n+1} &= w_n + \tau a(\theta_{n+1}) + \tau b(\theta_{n+1}) \end{aligned}$$

   and the trapezoidal rule

$$\begin{aligned} w_{n+1} &= w_n + \frac{\tau}{2} F(\theta_n, w_n) + \frac{\tau}{2} F(\theta_{n+1}, w_{n+1}) \\ \left(I - \frac{\tau}{2}(A + B_0) \right) w_{n+1} &= \left(I + \frac{\tau}{2}(A + B_0) \right) w_n \\ &\quad + \frac{\tau}{2} \big(a(\theta_n) + a(\theta_{n+1}) \big) + \frac{\tau}{2} \big(b(\theta_n) + b(\theta_{n+1}) \big). \end{aligned}$$

   As usual, a linear system has to be solved in each time-step of these methods. The matrix A is tridiagonal (cf. part I), but the matrix B_0 is **not sparse**, i.e. all entries of B_0 are in general nonzero! This is due to the fact that the integral (3.17) is non-local. Solving a linear system with such a matrix is much more expensive than in the tridiagonal case.

Roughly speaking, the stiffness originates from the terms $1/h$ and $1/h^2$ in the difference quotients. This observation suggests to treat only the differential part implicitly, whereas the integral part remains explicit. This yields the first-order **IMEX (implicit-explicit) method**

$$(I - \tau A)w_{n+1} = w_n + \tau B_0 w_n + \tau a(\theta_{n+1}) + \tau b(\theta_n). \tag{IMEX1}$$

This idea can be generalized to construct a IMEX method with classical order 2:

$$\begin{aligned} \Big(I - \frac{\tau}{2}A\Big)\, w^\star_{n+1} &= w_n + \frac{\tau}{2}Aw_n + \tau B_0 w_n + \frac{\tau}{2}\big(a(\theta_n) + a(\theta_{n+1})\big) + \tau b(\theta_n) \\ w_{n+1} &= w_n + \frac{\tau}{2}F(\theta_n, w_n) + \frac{\tau}{2}F(\theta_n, w^\star_{n+1}); \end{aligned} \tag{IMEX2}$$

cf. IV.4.3, Eq. (4.12) in [HV03].

Convergence. If the solution is sufficiently regular, then the convergence of these methods can be analyzed in a similar way as in part I, chapter 7. Some jump processes considered in mathematical finance, however, are pure jump models with diffusion coefficient $\sigma = 0$. In this case, the term $\partial_x^2 u(\theta, x)$ in (3.16a) vanishes, such that the PDE part is hyperbolic instead of parabolic, and classical (smooth) solutions do typically not exist because there is no diffusion part which improves the regularity of the solution as time evolves (cf. 7.6 (a) in part I). Hence, the word "solution" has then to be interpreted in a weaker sense of "viscosity solution". The classical order of the methods can, of course, not be expected for such a low regularity, but convergence (without specification of order) of the method (IMEX1) to a viscosity solution of (3.16) can still be shown; cf. Proposition 12.13, Section 12.4.4 in [CT04] and 6.3 in [CV05].

3.5 More general Lévy processes

The jump-diffusion models considered in this chapter belong to a much larger class of processes:

Definition 3.5.1 *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space with a filtration $\{\mathcal{F}_t : t \geq 0\}$. An adapted càdlàg process $X = \{X_t : t \geq 0\}$ with $X_t \in \mathbb{R}$ is called a* ***Lévy process*** *if it has the following properties:*

(i) $X_0 = 0$

(ii) Independent increments: $X_t - X_s$ is independent of $\mathcal{F}_s$ for all $0 \leq s < t < \infty$.

(iii) Stationary increments: $X_t - X_s$ has the same distribution as X_{t-s} for all $0 \leq s < t < \infty$.

(iv) Stochastically continuous: $\lim_{t \to s} \mathbb{P}(|X_t - X_s| > \varepsilon) = 0$ for all $\varepsilon > 0$ (limit in probability).

Examples: The Poisson process, the compound Poisson process, the Wiener process, and the jump-diffusion process (3.3) are Lévy processes.

The class of Lévy processes contains processes which can have infinitely many jumps in every interval; an example is the variance gamma process. Most of these jumps are small in the sense that for any given $C > 0$, the number of jumps with absolute size larger than C is finite. Such processes are sometimes used to model the underlying. The option price can again be determined by solving a PIDE numerically. Details: [CT04, CV05].

Chapter 4

The Finite Element Method for elliptic PDEs

4.1 Motivation

Up to now: Options with a single underlying, $V = V(t, S)$
Now: Options with $d > 1$ underlyings, $V = V(t, S_1, \ldots, S_d)$

Suppose that the prices of the underlyings are modelled by a system of SDEs

$$dS_i(t) = \mu_i S_i(t)dt + \sigma_i S_i(t)dW_i(t) \qquad (i = 1, \ldots, d)$$

with $S_i(t) \geq 0$ and $\mu_i, \sigma_i \in \mathbb{R}$. $W_1, \ldots, W_d$ are correlated
Wiener processes, i.e. $W(t) \sim \mathcal{N}(0, t\rho)$ with a symmetric, positive definite correlation matrix

$$\rho \in \mathbb{R}^{d\times d}, \qquad \rho_{ii} = 1, \qquad \rho_{ij} = \rho_{ji} \in [-1, 1]$$

(cf. 6.1.3 in part I). With similar arguments as in the one-dimensional case one can derive the ***d*-dimensional Black-Scholes equation**

$$\partial_t V + \frac{1}{2}\sum_{i,j=1}^{d} \rho_{ij}\sigma_i\sigma_j S_i S_j \partial_{S_i}\partial_{S_j} V + r\sum_{i=1}^{d} S_i \partial_{S_i} V - rV = 0 \qquad \text{(BSE)}$$

with value $V = V(t, S_1, \ldots, S_d)$, interest rate $r > 0$, σ_i volatilities, ρ_{ij} correlation coefficients.
Generalization: $r = r(t)$, $\sigma_i = \sigma_i(t, S_1, \ldots, S_d)$
Terminal condition of a European basket option:

$$V(T, S_1, \ldots, S_d) = \begin{cases} (K - \sum_{i=1}^{d} \alpha_i S_i)^+ & \text{Put} \\ (\sum_{i=1}^{d} \alpha_i S_i - K)^+ & \text{Call} \end{cases}$$

with weights $\alpha_1, \ldots, \alpha_d > 0$.
Now consider a European basket double barrier call option with $d = 2$ underlyings and payoff $V(T, S_1, S_2) = (S_1 + S_2 - K)^+$ (i.e. $\alpha_1 = \alpha_2 = 1$ for simplicity).
"Double barrier" means: The option becomes worthless if

$$\begin{aligned} & a_1 S_1 + a_2 S_2 \leq a_0 && \text{(down-and-out barrier)} \\ \text{or} \quad & b_1 S_1 + b_2 S_2 \geq b_0 && \text{(up-and-out barrier)} \end{aligned}$$

for given parameters $a_i, b_i > 0$. As a consequence, the Black-Scholes equation (BSE) has to be solved on the domain

$$\Omega = \{(S_1, S_2) \in \mathbb{R}^2 : a_1 S_1 + a_2 S_2 > a_0, \quad b_1 S_1 + b_2 S_2 < b_0\}$$

with boundary conditions

$$V(t, S_1, S_2) = 0 \qquad \text{if} \quad a_1 S_1 + a_2 S_2 = a_0 \quad \text{or} \quad b_1 S_1 + b_2 S_2 = b_0$$

and natural boundary conditions for $S_1 = 0$ and $S_2 = 0$, respectively.

picture

Due to the shape of the domain, a space discretization with finite differences is not suitable for such a problem.
The main advantages of the Finite Element Method (FEM) over finite-difference methods are:

- Discretization of domains with complex shape is possible.
- Error bounds can be shown under much lower regularity assumptions.

The FEM is mathematically more difficult and relies on sophisticated tools from analysis. Therefore, the FEM will be introduced for **elliptic** PDEs in this chapter and then extended to **parabolic** PDEs such as (BSE) in the next chapter.
An important motivation to analyze elliptic problems is the following: If we formally apply the implicit Euler method or the trapezoidal rule to (BSE), then we have to solve an elliptic boundary value problem in each time step.

4.2 Variational formulation of elliptic boundary value problems

As a model problem, we consider the **Poisson equation**

$$\left\{ \begin{aligned} -\Delta u &= f && \text{in } \Omega \\ u &= 0 && \text{on } \Gamma \end{aligned} \right. \tag{P}$$

$\Omega \subset \mathbb{R}^d$ is a bounded domain (connected, open, non-empty) with piecewise smooth boundary Γ. The (given) function $f : \Omega \longrightarrow \mathbb{R}$ is supposed to be continuous. The solution u is a scalar-valued function on $\bar{\Omega} = \Omega \cup \Gamma$. Δ denotes the **Laplace operator**

$$\Delta u = \partial^2_{x_1} u + \cdots + \partial^2_{x_d} u$$

Consider the space

$$V := \Big\{ v : \bar{\Omega} \longrightarrow \mathbb{R} \in C(\bar{\Omega}) \text{ and piecewise } C^1 \text{ with } v(x) = 0 \text{ for all } x \in \Gamma \Big\}.$$

Define

$$\begin{aligned} \ell : V \longrightarrow \mathbb{R}, \qquad & \ell(v) = \int_\Omega f(x) v(x) \; dx \\ a : V \times V \longrightarrow \mathbb{R}, \qquad & a(u,v) = \sum_{i=1}^{d} \int_\Omega \partial_{x_i} u(x) \cdot \partial_{x_i} v(x) \; dx. \end{aligned}$$

Notation: $\partial_{x_i} u(x) \cdot \partial_{x_i} v(x) = \Big(\partial_{x_i} u(x)\Big) \partial_{x_i} v(x) \neq \partial_{x_i} \Big(u(x) \partial_{x_i} v(x) \Big)$

It is easy to check that $\ell : V \longrightarrow \mathbb{R}$ is a linear form, and that $a : V \times V \longrightarrow \mathbb{R}$ is a bilinear form, i.e.

$$\begin{aligned} \ell(c_1 v_1 + c_2 v_2) &= c_1 \ell(v_1) + c_2 \ell(v_2) \\ a(c_1 v_1 + c_2 v_2, w) &= c_1 a(v_1, w) + c_2 a(v_2, w) \\ a(v, c_1 w_1 + c_2 w_2) &= c_1 a(v, w_1) + c_2 a(v, w_2) \end{aligned}$$

for all $c_1, c_2 \in \mathbb{R}$ and $v, w, v_i, w_i \in V$.
The bilinear form $a : V \times V \longrightarrow \mathbb{R}$ is even a scalar product on V, because it is

- symmetric, i.e.

$$a(w,v) = a(v,w) \qquad \text{for all } v, w \in V$$

- positive definite, i.e.

$$a(v,v) > 0 \qquad \text{for all } 0 \neq v \in V$$

Proof:

$$a(v,v) = 0 \Longleftrightarrow \int_\Omega \sum_{i=1}^{d} (\underbrace{\partial_{x_i} v}_{\geq 0})^2 \; dx = 0 \Longleftrightarrow v(x) \equiv const = 0,$$

because $v(x) = 0$ for all $x \in \Gamma$ and $v \in V$.
The induced norm $\|v\|_a = \sqrt{a(v,v)}$ is called the **energy norm**.

Proposition 4.2.1 (Variational formulation) *Define the functional*

$$J(v) := \frac{1}{2}a(v,v) - \ell(v)$$

and consider the following problems:

(i) Find a (classical) solution $u \in C^2(\Omega) \cap C(\bar{\Omega})$ of (P).

(ii) Find $u \in V$ such that $a(u,v) = \ell(v)$ for all $v \in V$.

(iii) Find $u \in V$ such that u is the solution of the minimization problem $J(v) = \min_{v \in V}!$, i.e. $J(u) \leq J(v)$ for all $v \in V$.

Then $(i) \Longrightarrow (ii) \Longleftrightarrow (iii)$. If $u \in C^2(\Omega) \cap C(\bar{\Omega})$, then $(ii) \Longrightarrow (i)$

For the proof we need the following tool:

Lemma 4.2.2 (Green's identity) *If $v, w \in C^1(\Omega) \cap C(\bar{\Omega})$, then*

$$\int_\Omega \partial_{x_i} w(x) \cdot v(x)\, dx = \int_\Gamma w(x)v(x)\eta_i(x)d\sigma(x) - \int_\Omega w(x) \cdot \partial_{x_i} v(x)\, dx$$

where $\eta(x) = (\eta_1(x), \ldots, \eta_d(x))^T$ denotes the outer unit normal vector in $x \in \Gamma$.

Remark: This is a generalization of integration by parts:

$$\int_a^b v'(x) \cdot w(x)\, dx = [v(x)w(x)]_{x=a}^b - \int_a^b v(x) \cdot w'(x)\, dx$$

Proof of Proposition 4.2.1:

(i) $\Longrightarrow$ (ii): For all $v \in V$ we have

$$\ell(v) = \int_\Omega f(x)v(x)\, dx \overset{(P)}{=} -\int_\Omega \Delta u(x)v(x)\, dx = -\sum_{i=1}^d \int_\Omega \partial_{x_i}^2 u(x) \cdot v(x)\, dx$$

Apply Green's identity:

$$-\int_\Omega \partial_{x_i}^2 u(x) \cdot v(x)\, dx = -\int_\Gamma \underbrace{\partial_{x_i} u(x) \cdot v(x)\eta_i(x)d\sigma(x)}_{=0 \text{ since } v(x)=0 \text{ for all } x \in \Gamma} + \int_\Omega \partial_{x_i} u(x) \cdot \partial_{x_i} v(x)\, dx$$

$$\Longrightarrow \qquad \ell(v) = a(u,v) \qquad \text{for all } v \in V.$$

(ii) $\Longleftrightarrow$ (iii): Pick an arbitrary $0 \neq v \in V$ and define $F_v : \mathbb{R} \longrightarrow \mathbb{R}$, $F_v(s) = J(u+sv)$.

$$\begin{aligned} F_v'(s) &= \frac{1}{2}\frac{d}{ds} \underbrace{a(u+sv, u+sv)}_{=a(u,u)+2sa(u,v)+s^2a(v,v)} - \frac{d}{ds}\underbrace{\ell(u+sv)}_{=\ell(u)+s\ell(v)} \\ &= a(u,v) + sa(v,v) - \ell(v) \\ F_v''(s) &= a(v,v) > 0 \qquad \text{for all } 0 \neq v \in V. \end{aligned}$$

u solves (iii) $\Longleftrightarrow$ $F_v(s)$ has a minimum in $s = 0$ for every $0 \neq v \in V$, and this is equivalent to

$$0 = F_v'(0) = a(u,v) - \ell(v) \qquad \text{for all } 0 \neq v \in V.$$

(ii) $\Longrightarrow$ (i) for $u \in C^2(\Omega) \cap C(\bar{\Omega})$: Since $a(u,v) = \ell(v)$ for all $v \in V$, it follows via Green's identity that

$$\int_\Omega (f(x) + \Delta u(x))v(x)\, dx = 0, \text{ for all } v \in V \supseteq C_c^\infty(\Omega)$$

Applying the fundamental lemma of calculus of variations (Lemma 2.3.2) yields $f(x) + \Delta u(x) = 0$ for all $x \in \Omega \Longrightarrow (i)$. ∎

4.3 Concept of the Finite Element Method

Instead of the original problem (P), we solve the variational formulation:

$$\text{Find } u \in V \text{ such that } a(u,v) = \ell(v) \text{ for all } v \in V.$$

Remark: The space V will later be replaced by more suitable space.

Choose a N-dimensional subspace $V_N \subset V$ and approximate u by the solution of the following finite-dimensional problem:

$$\text{Find } u_N \in V_N \text{ such that } a(u_N, v_N) = \ell(v_N) \text{ for all } v_N \in V_N. \tag{4.1}$$

Equivalent:

$$\text{Find } u_N \in V_N \text{ such that } J(u_N) \leq J(v_N) \text{ for all } v_N \in V_N. \tag{4.2}$$

Remarks: 1. Equivalence of (4.1) and (4.2) can be shown by using V_N instead of V in the part "(ii) $\Longleftrightarrow$ (iii)" in the proof of Proposition 4.2.1.
2. It follows from (4.1) that

$$a(u_N - u, v_N) = a(u_N, v_N) - a(u, v_N) = \ell(v_N) - \ell(v_N) = 0$$

because $v_N \in V_N \subset V$. Hence, the approximation u_N has the property that the error $u_N - u$ is orthogonal with respect to $a(\cdot,\cdot)$ to all elements of the space V_N. This is called the **Galerkin condition**.

Let $\{\varphi_1, ..., \varphi_n\}$ be a basis of V_N and assume that $u_N = \sum_{i=1}^N \widehat{u}_i \varphi_i$ with coefficients $\widehat{u}_i \in \mathbb{R}$ and $v_N = \sum_{j=1}^N \widehat{v}_j \varphi_j$ with coefficients $\widehat{v}_i \in \mathbb{R}$. Substitute this into (4.1) and use (bi-)linearity:

$$\sum_{i=1}^N \sum_{j=1}^N \widehat{u}_i \widehat{v}_j a(\varphi_i, \varphi_j) = \sum_{j=1}^N \widehat{v}_j \ell(\varphi_j) \qquad \text{for all } \widehat{v}_i \in \mathbb{R}.$$

In matrix-vector-notation (use that $A^T = A$):

$$\widehat{v}^T A \widehat{u} = \widehat{v}^T b \qquad \text{for all } \widehat{v} \in \mathbb{R}^N$$

with

$$\widehat{v} = (\widehat{v}_1, ..., \widehat{v}_N)^T, \qquad A = \Big(a(\varphi_i, \varphi_j)\Big)_{i,j} \in \mathbb{R}^{N \times N}$$
$$\widehat{u} = (\widehat{u}_1, ..., \widehat{u}_N)^T, \qquad b = \Big(\ell(\varphi_1), ..., \ell(\varphi_N)\Big)^T \in \mathbb{R}^N.$$

In particular, this must be true for the canonical basis vectors $\widehat{v} = e_k = (0, ..., 0, 1, 0, ..., 0)^T$ (the k-th entry is 1). Hence, we seek the solution $\widehat{u} \in \mathbb{R}^N$ of the linear equation

$$A\widehat{u} = b.$$

$a : V \times V \to \mathbb{R}$ is symmetric $\Longrightarrow A \in \mathbb{R}^{N \times N}$ is symmetric.
$a : V \times V \to \mathbb{R}$ is positive definite $\Longrightarrow A \in \mathbb{R}^{N \times N}$ is positive definite.

Miniproof:

$$\widehat{v}^T A \widehat{v} = \sum_{i=1}^N \sum_{j=1}^N \widehat{v}_i \widehat{v}_j a(\varphi_i, \varphi_j) = a\left(\sum_{i=1}^N \widehat{v}_i \varphi_i, \sum_{j=1}^N \widehat{v}_j \varphi_j\right) > 0 \qquad \text{for all } 0 \neq \widehat{v} \in \mathbb{R}^N.$$

Hence, the matrix A is invertible, and we have shown:

Proposition 4.3.1 (Existence and uniqueness of the numerical solution) *The finite-dimensional problem* (4.1) *has a unique solution* $u_N \in V_N$.

Open question: How to choose V_N? Will be answered later.

4.4 The Lax-Milgram lemma

Question: Does the variational formulation of the problem have a unique solution? Is there a unique $u \in V$ such that

$$a(u,v) = \ell(v) \quad \forall v \in V \qquad \text{or equivalently} \qquad J(u) \le J(v) \quad \forall v \in V \quad ?$$

Reminder: A vector space V with inner product $\langle \cdot, \cdot \rangle$ is a Hilbert space if V is complete with respect to the induced norm $\| \cdot \| = \sqrt{\langle \cdot, \cdot \rangle}$, i.e. if every Cauchy sequence converges.

Definition 4.4.1 (V-elliptic bilinear form) *A symmetric bilinear form $a : V \times V \to \mathbb{R}$ on a Hilbert space V with inner product $\langle \cdot, \cdot \rangle$ is called **V-elliptic** if the following holds:*

- *a is bounded: There is a constant $M < \infty$ such that $|a(u,v)| \le M \|u\| \cdot \|v\|$ for all $u, v \in V$.*
- *a is coercive: There is a constant $\alpha > 0$ such that $a(v,v) \ge \alpha \|v\|^2$ for all $v \in V$.*

Remark: If $a(\cdot, \cdot)$ is V-elliptic, then

$$M\|v\|^2 \ge \underbrace{a(v,v)}_{\|v\|_a^2} \ge \alpha \|v\|^2.$$

Hence, the norms $\| \cdot \|$ and $\| \cdot \|_a$ are equivalent, $\| \cdot \| \sim \| \cdot \|_a$.
Consequence: V is also a Hilbert space with respect to the inner product $a(\cdot, \cdot)$.

Theorem 4.4.2 (Lax-Milgram) *Let V be a Hilbert space with inner product $\langle \cdot, \cdot \rangle$, let $a : V \times V \to \mathbb{R}$ be V-elliptic, and let $\ell : V \to \mathbb{R}$ be a continuous linear form (i.e. there is a constant $c > 0$ such that $|\ell(v)| \le c\|v\|$ for all $v \in V$). Then, the minimization problem*

$$\textit{Find } u \in V \textit{ such that } J(u) \le J(v) \textit{ for all } v \in V$$

and the equivalent problem

$$\textit{Find } u \in V \textit{ such that } a(u,v) = \ell(v) \textit{ for all } v \in V$$

have a unique solution $u \in V$.

Proof.

- $J : V \to \mathbb{R}$ is bounded from below, because

$$J(v) = \frac{1}{2}a(v,v) - \ell(v) \ge \frac{1}{2}\alpha\|v\|^2 - c\|v\| = \frac{1}{2\alpha}(\alpha\|v\| - c)^2 - \frac{c^2}{2\alpha} \ge -\frac{c^2}{2\alpha}.$$

- Let $(v_n)_n$ be a minimal sequence, i.e. $v_n \in V$ and

$$\lim_{n\to\infty} J(v_n) = \inf_{v\in V} J(v) =: J_0.$$

Then, it follows that

$$\begin{aligned} 0 \le \alpha\|v_n - v_m\|^2 &\le a(v_n - v_m, v_n - v_m) \\ &= 4\underbrace{J(v_n)}_{\to J_0} + 4\underbrace{J(v_m)}_{\to J_0} - 8\underbrace{J\left(\frac{v_n + v_m}{2}\right)}_{\ge J_0} \longrightarrow 0. \end{aligned}$$

Hence, $(v_n)_n$ is a Cauchy sequence in V, and since V is a Hilbert space, the limit $\lim_{n\to\infty} v_n =: u \in V$ exists.

- Show that u is a solution of the minimization problem:

$$\begin{aligned} |J(u) - J_0| &= \lim_{n\to\infty} |J(u) - J(v_n)| \\ &\le \frac{1}{2} \lim_{n\to\infty} |\underbrace{a(u,u) - a(v_n, v_n)}_{=a(u-v_n, u+v_n)}| + \lim_{n\to\infty} |\ell(u - v_n)| \\ &\le \frac{1}{2} \lim_{n\to\infty} M \underbrace{\|u - v_n\|}_{\downarrow 0} \underbrace{\|u + v_n\|}_{\downarrow 2\|u\|} + c\underbrace{\|u - v_n\|}_{\downarrow 0} \to 0 \end{aligned}$$

Hence, $J(u) = J_0$ and thus $J(u) \le J(v)$ for all $v \in V$.

- Show uniqueness: Suppose that $w \in V$ is a solution, too:

$$\begin{aligned} & a(u,v) = \ell(v) = a(w,v) && \text{for all } v \in V \\ \Longrightarrow \quad & a(u - w, v) = 0 && \text{for all } v \in V. \end{aligned}$$

Choosing $v := u - w$ yields

$$0 = a(u - w, u - w) \ge \alpha\|u - w\|^2$$

which implies $u = w$. ∎

Back to the Poisson equation (P) and its variational formulation in Proposition 4.2.1:

$$V := \left\{ v \in \bar{\Omega} \longrightarrow \mathbb{R} \in C(\bar{\Omega}) \text{ and piecewise } C^1 \text{ with } v(x) = 0 \text{ for all } x \in \Gamma \right\}$$

$$a(u,v) = \sum_{i=1}^{d} \int_\Omega \partial_{x_i} u(x) \partial_{x_i} v(x)\, dx.$$

$$\ell(v) = \int_\Omega f(x) v(x)\, dx$$

Let $\langle \cdot, \cdot \rangle$ be an inner product on V with induced norm $\|\cdot\|$. Assume that $a(\cdot, \cdot)$ is V-elliptic when V is equipped with $\|\cdot\|$. By definition, this means that $\|\cdot\| \sim \|\cdot\|_a$. Hence, $(V, \langle \cdot, \cdot \rangle)$ is a Hilbert space if and only if $(V, a(\cdot, \cdot))$ is a Hilbert space. But here comes the bad news:

$$\textbf{\textit{(V, a(\cdot, \cdot))} is not a Hilbert space.}$$

Example: Consider the function $g(x) = (x^2 - 2)^2$. Then, the minimization problem

$$\text{Find } x_0 \text{ such that } g(x_0) \leq g(x) \text{ for all } x \in V$$

does **not** have a solution if $V = \mathbb{Q}$, whereas for $V = \mathbb{R}$ two solutions exists ($x_0 = \pm\sqrt{2}$).

4.5 Sobolev spaces

In this section, we compile a number of important definitions and results from analysis without proofs.

Goal: Construct completions of "classical" function spaces (such as $C^1(\Omega) \cap C(\overline{\Omega})$).

Proposition 4.5.1 (Completions) *If $V = (V, \langle \cdot, \cdot \rangle)$ is a vector space with inner product $\langle \cdot, \cdot \rangle$, then there is a unique (up to isometric isomorphisms) Hilbert space $(\overline{V}, \langle\langle \cdot, \cdot \rangle\rangle)$ such that V is dense in $\overline{V}$ and*

$$\langle\langle u, v \rangle\rangle = \langle u, v \rangle \qquad \text{for all } u, v \in V \subseteq \overline{V}.$$

The space $(\overline{V}, \langle\langle \cdot, \cdot \rangle\rangle)$ is called the ***completion*** *of $(V, \langle \cdot, \cdot \rangle)$. The completion of V is the smallest complete space which contains V.*

Idea of the proof: Consider equivalence classes of Cauchy sequences in V.

$$(v_n)_n \sim (u_n)_n \iff \lim_{n \to \infty} \|v_n - u_n\| = 0$$

Similar to the completion $\mathbb{Q} \longrightarrow \mathbb{R}$. ∎

Assumption for the rest of this chapter: $\Omega \subset \mathbb{R}^d$ is a bounded Lipschitz domain with boundary Γ. This means:

- Ω is open, bounded, connected, and non-empty.
- For every boundary point $x \in \Gamma$, there is a neighbourhood U such that, after an affine change of coordinates (translation and/or rotation), $\Gamma \cap U$ is described by the equation $x_d = \chi(x_1, \ldots, x_{d-1})$ with a uniformly Lipschitz continuous function χ. Moreover, $\Omega \cap U$ is on one side of $\Gamma \cap U$.

Example: Circles and rectangles are Lipschitz domains, but a circle with a cut is not.

Definition 4.5.2 ($L_p(\Omega)$ spaces)
For $p \in \mathbb{N}$:

$$L_p(\Omega) := \left\{ v : \Omega \to \mathbb{R} \text{ measurable and } \int_\Omega |v(x)|^p \, dx < \infty \right\}$$

$$\|v\|_{L_p(\Omega)} := \left(\int_\Omega |v(x)|^p \, dx \right)^{\frac{1}{p}}$$

For $p = \infty$:

$$L_\infty(\Omega) := \{ v : \Omega \to \mathbb{R} \text{ measurable and } |v(x)| < \infty \text{ a.e.} \}$$

$$\|v\|_{L_\infty(\Omega)} := \inf\{ c > 0 : \ |v(x)| \leq c \text{ a.e.} \}$$

The integral is the Lebesgue integral, and the abbreviation "a.e." means "almost everywhere", i.e. for all $x \in \Omega \setminus N$ for null sets N.
For convenience, we write $\|v\|_{L_p}$ instead of $\|v\|_{L_p(\Omega)}$ if it is clear which domain is meant.

The elements of $L_p(\Omega)$ are not functions but **equivalence classes** of functions: $u, v \in L_p(\Omega)$ are equivalent if $u = v$ a.e.. It is common usage to speak of "L_p functions", but some care is required: It does not make sense to speak about "the value of a L_p function in a point x" because a single point is a null set.

Important properties:

- $\|\cdot\|_{L_p}$ is a norm on $L_p(\Omega)$ for every $p \in \{1, 2, ..., \infty\}$, and $(L_p(\Omega), \|\cdot\|_{L_p})$ is a Banach space (complete).
- Special case $p = 2$: The space $L_2(\Omega)$ is a Hilbert space with inner product

$$\langle u, v \rangle_{L_2(\Omega)} = \int_\Omega u(x) v(x) \, dx.$$

Definition 4.5.3

$$L_1^{loc}(\Omega) := \left\{ u : \Omega \longrightarrow \mathbb{R} \text{ is measurable and locally integrable} \right\}$$

"Locally integrable" means that $u \in L^1(S)$ for all compact subsets $S \subset \Omega$.

Definition 4.5.4 (Weak derivative) *A function $u \in L_1^{loc}(\Omega)$ is **weakly differentiable** with respect to x_i if there is a $v \in L_1^{loc}(\Omega)$ such that*

$$\int_\Omega v(x)\phi(x) \, dx = -\int_\Omega u(x) \partial_{x_i}\phi(x) \, dx \qquad \text{for all } \phi \in C_c^\infty(\Omega)$$

or equivalently

$$\langle v, \phi \rangle_{L_2(\Omega)} = -\langle u, \partial_{x_i}\phi \rangle_{L_2(\Omega)} \qquad \text{for all } \phi \in C_c^\infty(\Omega).$$

*Then, v is called the **weak (partial) derivative** and is denoted by $v = \partial_{x_i} u$.*

Uniqueness: If $v = \partial_{x_i} u \in L_1^{loc}(\Omega)$ and $\widetilde{v} = \partial_{x_i} u \in L_1^{loc}(\Omega)$ are both weak partial derivatives of $u \in L_1^{loc}(\Omega)$, then $v = \widetilde{v}$ a.e.

Examples.

1. If $u \in C^1(\Omega) \cap C(\overline{\Omega})$, then Green's formula yields for all $\phi \in C_c^\infty(\Omega)$

$$\int_\Omega \partial_{x_i} u(x)\phi(x)\, dx \overset{\text{Green}}{=} \underbrace{\int_\Gamma u(x)\phi(x)\eta_i(x)d\sigma(x)}_{=0,\text{ since } \phi \in C_c^\infty(\Omega)} - \int_\Omega u(x)\partial_{x_i}\phi(x)\, dx.$$

   Hence, the weak derivative coincides with the classical derivative.

2. The function $u(x) = |x|$ is not differentiable on $\Omega := (-1, 1)$ in the classical sense. However, $u(x)$ is weakly differentiable with weak derivative

$$v(x) = \begin{cases} -1 & \text{for } x \le 0 \\ 1 & \text{for } x > 0 \end{cases}$$

   because for every $\phi \in C_c^\infty(\Omega)$ integration by parts yields

$$\begin{aligned} \int_{-1}^1 u(x)\phi'(x)\, dx &= \int_{-1}^0 u(x)\phi'(x)\, dx + \int_0^1 u(x)\phi'(x)\, dx \\ &= [u(x)\phi(x)]_{-1}^0 - \int_{-1}^0 (-1)\cdot\phi(x)\, dx + [u(x)\phi(x)]_0^1 - \int_0^1 1\cdot\phi(x)\, dx \\ &= -u(-1)\underbrace{\phi(-1)}_{=0} + u(1)\underbrace{\phi(1)}_{=0} - \int_{-1}^1 v(x)\phi(x)\, dx. \end{aligned}$$

   The function

$$\widetilde{v}(x) = \begin{cases} -1 & \text{for } x < 0 \\ 1 & \text{for } x \ge 0 \end{cases}$$

   is a weak derivative of $u(x)$, too. Both functions are identical as elements of $L_1^{loc}(\Omega)$.

3. Let $\Omega = (0, 2)$ and

$$u(x) = \begin{cases} x & \text{for } x \in (0, 1], \\ 2 & \text{for } x \in (1, 2). \end{cases}$$

   This function is not weakly differentiable.

Notation: For a multi-index $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ we set

$$\partial^\alpha u := \partial_{x_1}^{\alpha_1} \ldots \partial_{x_d}^{\alpha_d} u, \qquad |\alpha|_1 = \alpha_1 + \ldots + \alpha_d.$$

The above definition can now be extended to higher-order weak derivatives: $v(x) = \partial^\alpha u(x)$ is the weak partial derivative of $u(x)$ if

$$\int_\Omega v(x)\phi(x)\, dx = (-1)^{|\alpha|_1} \int_\Omega u(x)\partial^\alpha \phi(x)\, dx \qquad \text{for all } \phi \in C_c^\infty(\Omega).$$

Definition 4.5.5 (Sobolev space $H^m(\Omega)$) *For $m \in \mathbb{N}$ the Sobolev space $H^m(\Omega)$ is the set*

$$H^m(\Omega) := \Big\{ v \in L_2(\Omega) : \partial^\alpha v \in L_2(\Omega) \text{ exists for all } \alpha \in \mathbb{N}_0^d \text{ with } |\alpha|_1 \le m \Big\}$$

with inner product

$$\langle u, v \rangle_{H^m(\Omega)} := \sum_{|\alpha|_1 \le m} \langle \partial^\alpha u, \partial^\alpha v \rangle_{L_2(\Omega)}.$$

and norm $\|v\|_{H^m(\Omega)} = \sqrt{\langle v, v \rangle_{H^m(\Omega)}}$.

We write $\langle u, v \rangle_{H^m} = \langle u, v \rangle_{H^m(\Omega)}$ and $\|\cdot\|_{H^m} = \|\cdot\|_{H^m(\Omega)}$ if it is clear which Ω is meant.

Example ($m = 1$):

$$H^1(\Omega) := \Big\{ v \in L_2(\Omega) : \partial_{x_i} v \in L_2(\Omega) \text{ exists for all } i = 1, ..., d \Big\}$$

$$\langle u, v \rangle_{H^1(\Omega)} := \langle u, v \rangle_{L_2(\Omega)} + \sum_{i=1}^d \langle \partial_{x_i} u, \partial_{x_i} v \rangle_{L_2(\Omega)}$$

$$\|v\|_{H^1(\Omega)} = \left(\|v\|_{L_2(\Omega)}^2 + \sum_{i=1}^d \|\partial_{x_i} v\|_{L_2(\Omega)}^2 \right)^{1/2}.$$

Important properties:
$\big(H^m(\Omega), \langle \cdot, \cdot \rangle_{H^m}\big)$ is a Hilbert space for every $m \in \mathbb{N}_0$, and in particular $H^0(\Omega) = L_2(\Omega)$.
The space

$$C^{\infty,m}(\Omega) := \Big\{ v \in C^\infty(\Omega) : \int_\Omega |\partial^\alpha v(x)|^2\, dx < \infty \text{ for all } \alpha \in \mathbb{N}_0^d \text{ with } |\alpha|_1 \le m \Big\}.$$

is dense in $H^m(\Omega)$ with respect to $\|\cdot\|_{H^m}$: For every $u \in H^m(\Omega)$ and every $\varepsilon > 0$ there is a $v_\varepsilon \in C^{\infty,m}(\Omega)$ such that $\|v_\varepsilon - u\|_{H^m} < \varepsilon$.
It can be shown that $H^m(\Omega)$ is the completion of $C^{\infty,m}(\Omega)$ with respect to $\|\cdot\|_{H^m}$, i.e.

$$u \in H^m(\Omega) \Longleftrightarrow \text{There are } v_n \in C^{\infty,m}(\Omega) \text{ such that } \lim_{n\to\infty} \|u - v_n\|_{H^m(\Omega)} = 0.$$

Example: Let $\Omega = (-1, 1)$ and $u(x) = |x|$. Then $u \in H^1(\Omega)$, but $u \notin C^{\infty,1}(\Omega)$. However, we can define $v_n := \sqrt{x^2 + \frac{1}{n^2}} \in C^{\infty,1}(\Omega)$ and show that $\|v_n - u\|_{H^1(\Omega)} \longrightarrow 0$ for $n \longrightarrow \infty$ (exercise).

Definition 4.5.6 *The* ***Sobolev space*** $\boldsymbol{H_0^m(\Omega)}$ *is the completion of* $C_c^\infty(\Omega)$ *with respect to* $\|\cdot\|_{H^m(\Omega)}$*, i.e.*

$$u \in H_0^m(\Omega) \iff \text{There are } v_n \in C_c^\infty(\Omega) \text{ such that } \lim_{n\to\infty} \|u - v_n\|_{H^m(\Omega)} = 0.$$

$H_0^m(\Omega)$ is a closed subspace of $H^m(\Omega)$. If the boundary Γ is C^1, then $v \in C(\overline{\Omega}) \cap H_0^m(\Omega)$ implies that $v(x) = 0$ for all $x \in \Gamma$.

Proposition 4.5.7 (Poincaré-Friedrichs inequality) *Let*

$$|v|_{H^1(\Omega)} := \left(\sum_{i=1}^{d} \|\partial_{x_i} v\|_{L_2(\Omega)}^2 \right)^{\frac{1}{2}}$$

denote the Sobolev seminorm, i.e. $\|v\|_{H^1(\Omega)}^2 = \|v\|_{L_2(\Omega)}^2 + |v|_{H^1(\Omega)}^2$. *There is a constant* $c_\Omega > 0$ *such that*

$$\|v\|_{H^1(\Omega)} \le c_\Omega |v|_{H^1(\Omega)} \qquad \text{for all } v \in H_0^1(\Omega).$$

Proof. See 1.5, page 29 in [Bra07].

Remark: On $H_0^1(\Omega)$, the seminorm $|\cdot|_{H^1(\Omega)}$ is even a norm and $|\cdot|_{H^1(\Omega)} \sim \|\cdot\|_{H^1(\Omega)}$ (exercise).

If we want to solve a PDE with inhomogeneous boundary conditions, we need a condition such as, e.g., "$u(x) = g(x)$ for all $x \in \Gamma$". Since Γ is a null set, however, we have to specify how such a condition has to be understood if $u \in L_2(\Omega)$.

Theorem 4.5.8 (Trace theorem) *There is a bounded linear map*

$$\gamma : H^1(\Omega) \longrightarrow L_2(\Gamma), \qquad \|\gamma(v)\|_{L_2(\Gamma)} \le c\|v\|_{H^1(\Omega)} \qquad (c > 0)$$

such that $\gamma(v) = v|_\Gamma$ *for all* $v \in C^1(\bar{\Omega})$

Proof: page 45-47 in [Bra07]

Remark: It can be shown that

$$H_0^1(\Omega) := \big\{ v \in H^1(\Omega) : \gamma(v) = 0 \big\}$$

Theorem 4.5.9 (Sobolev's embedding theorem) *Let $\Omega \subset \mathbb{R}^d$ and let $r, s \in \mathbb{N}_0$ with $r > s + d/2$. Then*

$$H^r(\Omega) \subset C^s(\overline{\Omega})$$

and there is a constant $c > 0$ such that

$$\|u\|_{C^s(\overline{\Omega})} \leq c\|u\|_{H^r(\Omega)}, \qquad \text{where} \qquad \|u\|_{C^s(\overline{\Omega})} := \sum_{|\alpha|_1 \leq s} \sup_{x \in \overline{\Omega}} |\partial^\alpha u(x)|.$$

Examples:

$$\begin{aligned} d = 1 \quad &\Longrightarrow \quad H^1(\Omega) \subset C(\overline{\Omega}), \qquad H^2(\Omega) \subset C^1(\overline{\Omega}) \\ d \in \{2, 3\} \quad &\Longrightarrow \quad H^1(\Omega) \not\subset C(\overline{\Omega}), \qquad H^2(\Omega) \subset C(\overline{\Omega}) \end{aligned}$$

4.6 Variational formulation of more general elliptic boundary value problems

Let $\Omega \subseteq \mathbb{R}^d$ be a bounded Lipschitz domain with $d \in \{2, 3\}$ and assume that Γ is piecewise C^1. Consider the elliptic PDE

$$\mathcal{A}u = f$$

for $f \in L_2(\Omega)$ and with the second-order differential operator

$$\begin{aligned} \mathcal{A}u(x) &= -\mathrm{div}\big(\kappa(x)\nabla u(x)\big) + \kappa_0(x)u(x) \\ &= -\sum_{i,j=1}^{d} \partial_{x_i}\Big(\kappa_{ij}(x)\partial_{x_j}u(x)\Big) + \kappa_0(x)u(x) \end{aligned}$$

with $\kappa(x) = \big(\kappa_{ij}(x)\big)_{i,j} \in \mathbb{R}^{d\times d}$ for all $x \in \Omega$. The coefficient functions $\kappa_{ij} : \Omega \longrightarrow \mathbb{R}$ and $\kappa_0 : \Omega \longrightarrow \mathbb{R}$ are supposed to have the following properties:

1. $|\kappa_0(x)|, \quad |\kappa_{ij}(x)| \leq M \qquad$ for all $x \in \Omega$
2. $\kappa_{ij} = \kappa_{ji} \qquad$ for all $i, j = 1, ..., d$
3. There are constants $c_0 \geq 0$ and $c_1 > 0$ such that $\kappa_0(x) \geq c_0$ and

$$\sum_{i=1}^{d}\sum_{j=1}^{d} \kappa_{ij}(x)\xi_i\xi_j \geq c_1 \sum_{i=1}^{d} \xi_i^2 \qquad \text{for all } x \in \Omega \text{ and } \xi = (\xi_1, \ldots, \xi_d)^T \in \mathbb{R}^d. \quad (4.3)$$

The second assertion is equivalent to

$$\xi^T\kappa\xi \geq c_1|\xi|_2^2, \qquad \xi = (\xi_1, ..., \xi_d)^T, \qquad |\xi|_2 = \left(\sum_{i=1}^{d} \xi_i^2\right)^{1/2}.$$

A differential operator with the properties 2 and 3 is called **elliptic**.

Example: For $\kappa_0(x) = 0$, $\kappa_{ii} = 1$ and $\kappa_{ij} = 0$ for $i \neq j$, we obtain $\mathcal{A} = -\Delta$.

1. Homogeneous Dirichlet boundary conditions. Consider the boundary value problem

$$\begin{aligned} \mathcal{A}u &= f \qquad \text{in } \Omega \\ u &= 0 \qquad \text{on } \Gamma. \end{aligned}$$

Assume that $u \in C^2(\Omega) \cap C(\overline{\Omega})$ is a classical solution. Multiply both sides of the first line with a test function $v \in C_c^\infty(\Omega)$ and integrate over Ω:

$$-\sum_{i,j=1}^{d} \int_\Omega \partial_{x_i}\Big(\kappa_{ij}(x)\partial_{x_j}u(x)\Big)v(x)\,dx + \int_\Omega \kappa_0(x)u(x)v(x)\,dx = \int_\Omega f(x)v(x)\,dx$$

Apply Green's formula:

$$\underbrace{\sum_{i,j=1}^{d} \int_\Omega \kappa_{ij}(x)\partial_{x_j}u(x)\cdot\partial_{x_i}v(x)\,dx + \int_\Omega \kappa_0(x)u(x)v(x)\,dx}_{=:\,a(u,v)} = \underbrace{\int_\Omega f(x)v(x)\,dx}_{=:\,\ell(v)}$$

Equivalent notation:

$$a(u,v) = \int_\Omega (\nabla v(x))^T \kappa(x) \nabla u(x)\,dx + \int_\Omega \kappa_0(x)u(x)v(x)\,dx$$

Variational (weak) formulation: Find $u \in H_0^1(\Omega)$ such that

$$a(u,v) = \ell(v) \qquad \text{for all } v \in H_0^1(\Omega). \tag{4.4}$$

Such an u is called a **weak solution** of the boundary value problem.
Since we seek a solution in $H_0^1(\Omega)$, the boundary condition is fulfilled (in the sense of the trace theorem).

Show that the assumptions of the Lax-Milgram theorem (Theorem 4.4.2) are true:

- Show that $\ell : H_0^1(\Omega) \to \mathbb{R}$ is a continuous linear form:

$$|\ell(v)| = |\langle f, v\rangle_{L_2}| \overset{\text{Cauchy-Schwarz}}{\leq} \|f\|_{L_2}\|v\|_{L_2} \leq \underbrace{\|f\|_{L_2}}_{=:c}\|v\|_{H^1}.$$

- Show that $a(\cdot,\cdot)$ is a $H_0^1(\Omega)$-elliptic bilinear form. Boundedness:

$$\begin{aligned}
|a(u,v)| &\leq \sum_{i,j=1}^{d} \int_\Omega \underbrace{|\kappa_{ij}(x)|}_{\leq M} \cdot |\partial_{x_j} u(x)| \cdot |\partial_{x_i} v(x)| + \underbrace{|\kappa_0(x)|}_{\leq M} \cdot |u(x)| \cdot |v(x)| \, dx \\
&\leq M \sum_{i,j=1}^{d} \left\langle |\partial_{x_j} u|, |\partial_{x_i} v| \right\rangle_{L_2} + M \left\langle |u|, |v| \right\rangle_{L_2} \\
&\overset{\text{Cauchy-Schwarz}}{\leq} M \sum_{i,j=1}^{d} \|\partial_{x_j} u\|_{L_2} \cdot \|\partial_{x_i} v\|_{L_2} + M \|u\|_{L_2} \cdot \|v\|_{L_2} \\
&\leq M(d^2+1) \cdot \|u\|_{H^1} \|v\|_{H^1}
\end{aligned}$$

  Coercivity:

$$\begin{aligned}
a(v,v) &= \int_\Omega \sum_{i,j=1}^{d} \kappa_{ij}(x) \partial_{x_i} v(x) \cdot \partial_{x_j} v(x) + \kappa_0(x) v^2(x) \, dx \\
&\geq \int_\Omega c_1 \sum_{i=1}^{d} (\partial_{x_i} v(x))^2 + c_0 v^2(x) \, dx && (\mathcal{A} \text{ is elliptic}) \\
&= c_1 |v|_{H^1}^2 + c_0 \|v\|_{L_2}^2 \geq c_1 |v|_{H^1}^2 \\
&\geq \frac{c_1}{c_\Omega^2} \|v\|_{H^1}^2 && (\text{Poincaré-Friedrichs})
\end{aligned}$$

Applying Lax-Milgram (Theorem 4.4.2) with $V = H_0^1(\Omega)$ yields that the problem (4.4) has a unique (weak) solution $u \in H_0^1(\Omega)$.

Remark: The definition of the differential operator $\mathcal{A}$ only makes sense if $\kappa_{ij} \in C^1(\Omega)$, whereas only $\kappa_{ij} \in L_\infty(\Omega)$ is required in the weak formulation.

2. Inhomogeneous Dirichlet boundary conditions. For a given function $g : \Gamma \to \mathbb{R}$ we consider the problem

$$\begin{aligned}
\mathcal{A}u &= f \qquad \text{in } \Omega \\
u &= g \qquad \text{on } \Gamma
\end{aligned}$$

Assumption: There is a $u_* \in H^1(\Omega)$ such that $g = u_*\big|_\Gamma = \gamma(u_*)$ in the sense of the trace theorem 4.5.8.

Variational (weak) formulation: Find $u \in H^1(\Omega)$ (instead of $H_0^1(\Omega)$) such that $w := u - u_* \in H_0^1(\Omega)$ and

$$a(u,v) = \ell(v) \qquad \text{for all } v \in H_0^1(\Omega) \tag{4.5}$$

(same derivation as before). Equivalent: Find $w \in H_0^1(\Omega)$ such that

$$a(w,v) = \ell(v) - a(u_*, v) =: \widetilde{\ell}(v) \qquad \text{for all } v \in H_0^1(\Omega). \tag{4.6}$$

Verify the conditions of the Lax-Milgram theorem: $\widetilde{\ell} : H_0^1(\Omega) \to \mathbb{R}$ is continuous, because

$$\begin{aligned} |\widetilde{\ell}(v)| &\le \underbrace{|\ell(v)|}_{=|\langle f,v\rangle_{L_2}|} + |a(u_*,v)| \\ &\le \|f\|_{L_2}\|v\|_{H^1} + M(d^2+1)\|u_*\|_{H^1}\|v\|_{H^1} \\ &= \left(\|f\|_{L_2} + M(d^2+1)\|u_*\|_{H^1}\right)\|v\|_{H^1} \end{aligned}$$

The Lax-Milgram theorem yields that (4.6) has a unique solution $w \in H_0^1(\Omega)$. Hence, (4.5) has a unique solution $u = u_* + w \in H^1(\Omega)$.

3. Neumann boundary conditions. For every $x \in \Gamma$, we define the **conormal derivative**

$$\frac{\partial u}{\partial \eta_\kappa}(x) := \sum_{i,j=1}^d \kappa_{ij}(x)\partial_{x_j} u(x) \cdot \eta_i(x) = \eta(x)^T \kappa(x) \nabla u(x),$$

where $\eta : \Gamma \to\in \mathbb{R}^d$ is again the outer unit normal. Let $f \in L_2(\Omega)$ and $g \in L_2(\Gamma)$ and consider the boundary value problem

$$\mathcal{A}u = f \qquad \text{in } \Omega \tag{4.7a}$$

$$\frac{\partial u}{\partial \eta_\kappa} = g \qquad \text{on } \Gamma. \tag{4.7b}$$

Variational (weak) formulation? Let $u \in C^2(\Omega) \cap C(\overline{\Omega})$ is a classical solution and proceed as in 1.: Multiply both sides of the first line with a test function $v \in C^\infty(\Omega)$, integrate over Ω and apply Green's formula. This yields

$$\begin{aligned} \ell(v) &= \int_\Omega f(x)v(x)\,dx = \int_\Omega \mathcal{A}u(x)v(x)\,dx \\ &= -\sum_{i,j=1}^d \int_\Omega \partial_{x_i}\Big(\kappa_{ij}(x)\partial_{x_j}u(x)\Big)\cdot v(x)\,dx + \int_\Omega \kappa_0(x)u(x)v(x)\,dx \\ &= -\int_\Gamma \underbrace{\left(\sum_{i,j=1}^d \kappa_{ij}(x)\partial_{x_j}u(x)\cdot\eta_i(x)\right)}_{=\frac{\partial u}{\partial\eta_\kappa}(x)=g(x)} v(x)d\sigma(x) \\ &\quad + \int_\Omega \sum_{i,j=1}^d \kappa_{ij}(x)\partial_{x_j}u(x)\cdot\partial_{x_i}v(x) + \int_\Omega \kappa_0(x)u(x)v(x)\,dx \\ &= -\int_\Gamma g(x)v(x)d\sigma(x) + a(u,v) \end{aligned}$$

Weak formulation: Find $u \in H^1(\Omega)$ (instead of $H^1_0(\Omega)$) such that

$$a(u,v) = \ell(v) + \int_\Gamma g(x)v(x)\,d\sigma(x) =: \widetilde{\ell}(v) \qquad \text{for all } v \in H^1(\Omega) \tag{4.8}$$

with $a(\cdot,\cdot)$ as before. Lax-Milgram conditions:

- Show that $\ell : H^1(\Omega) \to \mathbb{R}$ is continuous: The trace theorem (Theorem 4.5.8) yields

$$\begin{aligned} |\widetilde{\ell}(v)| &\leq \|f\|_{L_2(\Omega)} \cdot \|v\|_{H^1(\Omega)} + \|g\|_{L_2(\Gamma)} \cdot \underbrace{\|\gamma(v)\|_{L_2(\Gamma)}}_{\leq c\|v\|_{H^1(\Omega)}} \\ &\leq C\|v\|_{H^1(\Omega)}. \end{aligned}$$

- Show that $a(\cdot,\cdot)$ is $H^1(\Omega)$-elliptic. Boundedness can be shown as before. To prove coercivity, it can be shown as in 1. that

$$a(v,v) \geq c_1|v|^2_{H^1} + c_0\|v\|^2_{L_2}.$$

  This time, the Poincaré-Friedrichs inequality cannot be applied. In order to obtain coercivity, we have to assume that $c_0 > 0$ (instead of $c_0 \geq 0$) and use that

$$c_1|v|^2_{H^1} + c_0\|v\|^2_{L_2} \geq \min\{c_1, c_0\} \cdot \|v\|^2_{H^1}.$$

Lax-Milgram $\longrightarrow$ unique solution.

Remark. Let $\kappa_0(x) = 0$ for all $x \in \Omega$, i.e. $c_0 = 0$. In this case, the solution of the boundary value problem (4.7) is not unique (exercise). This indicates that the case $c_0 = 0$ may also cause problems in the original formulation of the problem.

4.7 Linear finite elements

Consider the boundary value problem

$$\begin{aligned} \mathcal{A}u &= f \qquad \text{in } \Omega \\ u &= 0 \qquad \text{on } \Gamma \end{aligned}$$

from the previous section in two dimensions ($\Omega \subset \mathbb{R}^2$). The variational formulation is: Find $u \in H^1_0(\Omega)$ such that

$$a(u,v) = \ell(v) \qquad \text{for all } v \in H^1_0(\Omega).$$

where

$$\begin{aligned} a(u,v) &= \int_\Omega (\nabla v(x))^T \kappa(x) \nabla u(x)\,dx + \int_\Omega \kappa_0(x)u(x)v(x)\,dx \\ \ell(v) &= \int_\Omega f(x)v(x)\,dx. \end{aligned}$$

Numerical approximation: Choose an N-dimensional subspace $V_N \subset H_0^1(\Omega)$ and find $u_N \in V_N$ such that

$$a(u_N, v_N) = \ell(v_N) \qquad \text{for all } v_N \in V_N$$

(cf. section 4.3). Question: How to choose V_N?
In this lecture, we will focus on *linear* elements.

- **Triangulation.** Approximate the boundary Γ by a polygon $\widetilde{\Gamma}$ with interior $\widetilde{\Omega}$. Subdivide $\widetilde{\Omega}$ into $m \in \mathbb{N}$ triangles $T_k \subset \mathbb{R}^2$ with vertices P_k^1, P_k^2, $P_k^3 \in \mathbb{R}^2$. Conditions:
  - If $P_k^i \in T_j$, then $P_k^i \in \{P_j^1, P_j^2, P_j^3\}$. This means in particular that two triangles can have common points or common edges, but their interiors must not intersect.
  - If $x \in \widetilde{\Gamma}$, then x must lie on an edge of a triangle.

  Let $P^1, P^2, \ldots, P^N$ be the set of all inner points.

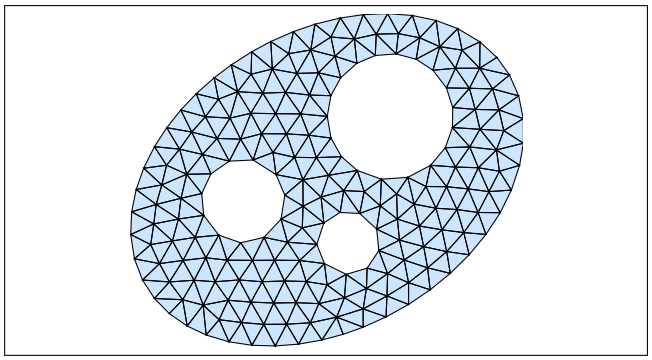


Figure 4.1: Example of a triangulation.

- **Basis.** Let φ_i be a **hat function**, i.e.

$$\varphi_i\big|_{T_k} \text{ is linear for all } k, \text{ and } \varphi_i(P_j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{else.} \end{cases}$$

Picture

  Choose $V_N = \operatorname{span}\{\varphi_1, \ldots, \varphi_N\}$. It can be shown that $V_N \subset H_0^1(\widetilde{\Omega})$. Seek approximation

$$u_N(x) = \sum_{i=1}^{N} \widehat{u}_i \varphi_i(x).$$

  The coefficients $\widehat{u}_i$ are simply the values of u_N in the vertices, because by definition $u_N(P_j) = \sum_{i=1}^{N} \widehat{u}_i \varphi_i(P_j) = \widehat{u}_j$.

- **Compiling the linear system.** If we let

$$A = \Big(a(\varphi_i, \varphi_j)\Big)_{i,j} \in \mathbb{R}^{N\times N} \qquad b = \Big(\ell(\varphi_1), ..., \ell(\varphi_N)\Big)^T \in \mathbb{R}^N,$$

then $\widehat{u} = (\widehat{u}_1, ..., \widehat{u}_N)^T$ is the solution of the linear problem $A\widehat{u} = b$ (cf. section 4.3). The **stiffness matrix** A is symmetric, positive definite and sparse, because $T_k \cap T_j = \emptyset$ for many pairs $j, k \in \{1, \ldots, N\}$; cf. Figure 4.2.

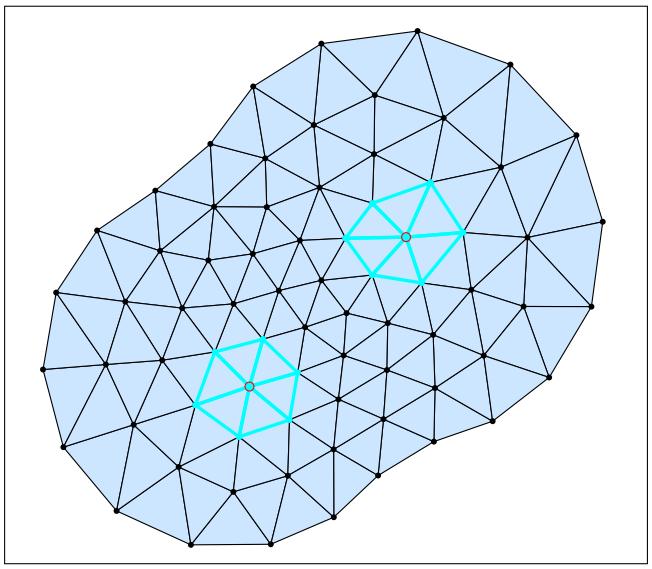


Figure 4.2: The supports of the two basis elements corresponding to the two red dots do not intersect. Hence, the corresponding entry in the stiffness matrix is zero.

Element-wise computation:

$$\begin{aligned} a(\varphi_i, \varphi_j) &\approx \sum_{k=1}^{m} a_k(\varphi_i, \varphi_j) \\ a_k(\varphi_i, \varphi_j) &= \int_{T_k} (\nabla\varphi_j(x))^T \kappa(x) \nabla\varphi_i(x)\, dx + \int_{T_k} \kappa_0(x)\varphi_i(x)\varphi_j(x)\, dx \\ \ell(\varphi_j) &\approx \sum_{k=1}^{m} \int_{T_k} f(x)\varphi_j(x)\, dx. \end{aligned}$$

Consider a fixed triangle T_k for some $k \in \{1, \ldots, m\}$ and call the coordinates (x, y) instead of $x = (x_1, x_2)$.
Change from global to local indexing: Let $P_i^k = (x_i^k, y_i^k)$ be the three corners of the triangle, and let

$$\varphi_i^k(x, y) = \alpha_i + \beta_i x + \gamma_i y$$

be the associated basis functions ($i = 1, 2, 3$). The coefficients $\alpha_i, \beta_i, \gamma_i$ depend on k. In order to determine the coefficients $\alpha_i, \beta_i, \gamma_i$, we use that the interpolation property

$$\alpha_i + \beta_i x_j^k + \gamma_i y_j^k = \varphi_i^k(x_j^k, y_j^k) = \varphi_i^k(P_j^k) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

for $i, j = 1, 2, 3$ is equivalent to

$$\begin{pmatrix} 1 & x_1^k & y_1^k \\ 1 & x_2^k & y_2^k \\ 1 & x_3^k & y_3^k \end{pmatrix} \begin{pmatrix} \alpha_1 & \alpha_2 & \alpha_3 \\ \beta_1 & \beta_2 & \beta_3 \\ \gamma_1 & \gamma_2 & \gamma_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Hence, it follows that

$$\begin{pmatrix} \alpha_1 & \alpha_2 & \alpha_3 \\ \beta_1 & \beta_2 & \beta_3 \\ \gamma_1 & \gamma_2 & \gamma_3 \end{pmatrix} = \begin{pmatrix} 1 & x_1^k & y_1^k \\ 1 & x_2^k & y_2^k \\ 1 & x_3^k & y_3^k \end{pmatrix}^{-1}.$$

Now the gradient

$$\nabla \varphi_i(x, y) = \begin{pmatrix} \beta_i \\ \gamma_i \end{pmatrix}$$

can easily be obtained. Note that $\nabla \varphi_i(x)$ is constant on each T_k. The integrals in $a_k(\varphi_i, \varphi_j)$ can be approximated by quadrature or computed exactly if the functions $\kappa_{ij}(x)$ and $\kappa_0(x)$ are simple enough.

- **Solving the linear system:** Cholesky decomposition if N is small enough, multigrid methods or Krylov methods if N is large.

Extensions: Higher-order finite elements, problems in higher dimension, other boundary conditions, ... see [Bra07, BS08, LT09]
Details about the implementation: [Goc06]

4.8 Accuracy

Notation: Let $|\cdot|_2$ be the Euclidean vector norm of a vector $v \in \mathbb{R}^N$, i.e. $|v|_2 = \sqrt{v^T v}$ (not to be confused with the Sobolev seminorm $|\cdot|_{H^1}$).

Let $u \in V$ be the solution of

$$a(u, v) = \ell(v) \qquad \text{for all } v \in V.$$

Let $u_h \in V_h$ be the solution of

$$a(u_h, v_h) = \ell(v_h) \qquad \text{for all } v_h \in V_h,$$

where V_h is the space of piecewise linear finite elements with maximal edge of length

$$h := \max_{k=1,\dots,m} \max \Big\{ |P_k^1 - P_k^2|_2, |P_k^1 - P_k^3|_2, |P_k^2 - P_k^3|_2 \Big\}.$$

How good is the approximation $u_h \approx u$?

Remark. The error $u - u_h$ is the Galerkin error. In applications, additional errors can be caused by

- numerical approximation of integrals in $a(\cdot,\cdot)$ and $\ell(\cdot)$ by quadrature,
- approximation $\widetilde{\Omega} \approx \Omega$
- solving $A\widehat{u} = b$by iterative methods (Krylov, multigrid), and
- roundoff errors.

We will only analyze the Galerkin error.

(a) Céa's Lemma

Lemma 4.8.1 (Céa's Lemma) *Among all $v_h \in V_h$ the Galerkin approximation u_h yields the best approximation with respect to the energy norm $\|w\|_a = \sqrt{a(w,w)}$, i.e.*

$$\|u_h - u\|_a \leq \|v_h - u\|_a \qquad \textit{for all } v_h \in V_h.$$

Proof. We have seen in Section 4.3 that

$$a(u - u_h, v_h) = 0 \qquad \text{for all } v_h \in V_h.$$

Hence, it follows that

$$\|u - u_h\|_a^2 = a(u - u_h, u - u_h) = a(u - u_h, u) - \underbrace{a(u - u_h, u_h)}_{=0}$$

$$= a(u - u_h, u - v_h) \overset{\text{C.S.}}{\leq} \|u - u_h\|_a \|u - v_h\|_a.$$

Dividing by $\|u - u_h\|_a$ yields the assertion. ∎

Next question: How accurately can u be approximated in V_h?

(b) Bounds for the interpolation error

Let $\Omega \subseteq \mathbb{R}^2$ be a bounded Lipschitz domain, and assume that Γ is a polygon. Assume that $u \in H^2(\Omega)$. Since $d = 2$, it follows by Sobolev embedding (Theorem 4.5.9), that $u \in C^0(\bar{\Omega})$. Hence, we can define the interpolant

$$\psi(x) = \sum_{j=1}^{N} u(P_j)\varphi_j(x), \qquad \psi(P_j) = u(P_j) \qquad \text{for all } j = 1, ..., N \tag{4.9}$$

Strategy: Choose $v_h = \psi$ in Céas Lemma and derive a suitable upper bound for the difference $u - \psi$.

Theorem 4.8.2 (Error bound for the interpolation error) *Under these assumptions, the interpolation error can be bounded by*

$$\|u - \psi\|_{L_2} \leq \sqrt{\frac{3}{8}} h^2 \|u\|_{H^2} \tag{4.10}$$

$$|u - \psi|_{H^1} \leq \frac{3}{\sqrt{8}\sin^2(\beta)} h \|u\|_{H^2} \tag{4.11}$$

where β is the smallest interior angle of all triangles.

Sketch of the proof (see XVI, Satz 91.6 in [HB09] for the full proof). Show results only for $u \in C^2(\overline{\Omega})$ and use that $C^2(\overline{\Omega})$ is dense in $H^2(\Omega)$.

(a) Let T be a fixed triangle with vertices P_1, P_2, P_3. (Choose suitable enumeration of the points if necessary.) Let $x = (x_1, x_2) \in T$ and $d_k := P_k - x$ for $k \in \{1, 2, 3\}$. Taylor expansion:

$$u(P_k) = u(x) + (\nabla u(x))^T d_k + R_k(x) \tag{4.12}$$

$$R_k(x) := \int_0^1 d_k^T \left[u''(x + s d_k)\right] d_k (1 - s) ds, \qquad (u'' \text{ Hessian})$$

Substitute into (4.9) and use that $\varphi_k(x) = 0$ for all $k > 3$ (because x is in the triangle with vertices P_1, P_2, P_3):

$$\begin{aligned} \psi(x) &= \sum_{j=1}^{N} u(P_j)\varphi_j(x) \\ &= u(x) \sum_{k=1}^{3} \varphi_k(x) + (\nabla u(x))^T \sum_{k=1}^{3} d_k \varphi_k(x) + \sum_{k=1}^{3} R_k(x)\varphi_k(x) \end{aligned} \tag{4.13}$$

Some terms can be simplified:

$$\sum_{k=1}^{3} \varphi_k(x) = 1 \quad \text{for all } x \in T \quad \text{(piecewise linear interpolant of } g(x) \equiv 1) \tag{4.14}$$

$$\sum_{k=1}^{3} P_k \varphi_k(x) = x \quad \text{for all } x \in T \quad (g \text{ piecewise linear and } g(P_k) = P_k \implies g(x) = x) \tag{4.15}$$

Hence:

$$\sum_{k=1}^{3} d_k \varphi_k(x) \overset{\text{Def.}}{=} \underbrace{\sum_{k=1}^{3} P_k \varphi_k(x)}_{=x} - \sum_{k=1}^{3} x \varphi_k(x) = x - x = 0 \qquad \text{for all } x \in T$$

Together with (4.13), this yields

$$\psi(x) - u(x) = \sum_{k=1}^{3} R_k(x) \varphi_k(x).$$

Cauchy-Schwarz in $\mathbb{R}^3$ yields

$$|\psi(x) - u(x)|^2 \leq \sum_{k=1}^{3} R_k^2(x) \cdot \sum_{k=1}^{3} \underbrace{\varphi_k^2(x)}_{\leq \varphi_k(x)} \overset{(4.14)}{\leq} \sum_{k=1}^{3} R_k^2(x)$$

and hence

$$\|\psi - u\|_{L_2(T)}^2 \leq \sum_{k=1}^{3} \int_T R_k^2(x)\, dx = I_1 + I_2 + I_3 \qquad \text{with } I_k = \int_T R_k^2(x)\, dx \tag{4.16}$$

(b) Choose $k \in \{1, 2, 3\}$. Extend the triangle T to a sector S with radius $h_T := \max_{i,j \in \{1,2,3\}} |P_i - P_j|_2$ as in Figure 4.3.

Extend u'' and hence the integrand $R_k^2(x)$ to the entire sector S by defining

$$u''(x) = 0 \qquad \text{for all } x \in S \setminus T.$$

Polar coordinates:

$$x = P_k + r x_\theta, \qquad x_\theta = \begin{pmatrix} \cos(\theta) \\ \sin(\theta) \end{pmatrix}, \qquad \theta \in [\theta_0, \theta_0 + \gamma_k], \qquad r \in [0, h_T].$$

With $d_k = P_k - x = -r x_\theta$ this yields

$$\begin{aligned} I_k &= \int_S R_k^2(x)\, dx \\ &= \int_{\theta_0}^{\theta_0 + \gamma_k} \int_0^{h_T} \underbrace{\left(\int_0^1 r x_\theta^T \Big[u''(P_k + (1-s) r x_\theta) \Big] r x_\theta (1-s)\, ds \right)^2}_{=(\star)} r\, dr d\theta \end{aligned}$$

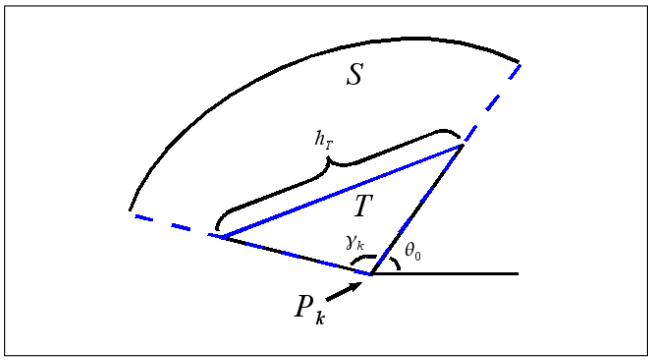


Figure 4.3: Triangle T and sector S.

For the inner term $(\star)$, we substitute $q = (1-s)r$ and use Cauchy-Schwarz in $L_2(T)$:

$$\begin{aligned}(\star) &= \left(-\int_0^r x_\theta^T \Big[u''(P_k + qx_\theta)\Big] x_\theta \sqrt{q} \cdot \sqrt{q}\, dq\right)^2 \\ &\le \int_0^r \left(x_\theta^T \Big[u''(P_k + qx_\theta)\Big] x_\theta\right)^2 q\, dq \cdot \underbrace{\int_0^r q\, dq}_{=\frac{1}{2}r^2}\end{aligned}$$

Let $|M|_F = \sqrt{\sum_{i,j} M_{ij}^2}$ be the Frobenius norm of a matrix M with entries M_{ij}. Then $|Mv|_2^2 \le |M|_F^2 \cdot |v|_2^2$ (exercise) and hence

$$\left(x_\theta^T \Big[u''(P_k + qx_\theta)\Big] x_\theta\right)^2 \le \underbrace{|x_\theta|_2^2}_{=1} \cdot \Big|\Big[u''(P_k + qx_\theta)\Big] x_\theta\Big|_2^2 \le |u''(P_k + qx_\theta)|_F^2.$$

Altogether, this yields

$$\begin{aligned}I_k &\le \frac{1}{2}\int_{\theta_0}^{\theta_0+\gamma_k} \int_0^{h_T} \Big(\underbrace{\int_0^r |u''(P_k + qx_\theta)|_F^2 q\, dq}_{\le \int_0^{h_T} \ldots dq}\Big) r^3\, dr d\theta \\ &\le \frac{1}{2}\int_{\theta_0}^{\theta_0+\gamma_k} \int_0^{h_T} |u''(P_k + qx_\theta)|_F^2 q\, dq \cdot \underbrace{\int_0^{h_T} r^3 dr}_{=\frac{h_T^4}{4}}\, d\theta \\ &\le \frac{h_T^4}{8}\int_T |u''(x)|_F^2\, dx \\ &\le \frac{1}{8} h_T^4 \|u\|_{H^2(T)}^2 \qquad (4.17)\end{aligned}$$

This yields the first bound (4.10), because with $h_T \le h$ we obtain

$$\|\psi - u\|_{L_2}^2 = \sum_T \|\psi - u\|_{L_2(T)}^2 \le \frac{3}{8} h^4 \|u\|_{H^2}^2$$

(c) Prove a bound for the gradient. Law of sines

$$\frac{|P_2 - P_3|_2}{\sin(\alpha_1)} = \frac{|P_1 - P_3|_2}{\sin(\alpha_2)} = \frac{|P_1 - P_2|_2}{\sin(\alpha_3)} \geq h_T = \max_{i,j \in \{1,2,3\}} |P_i - P_j|_2$$

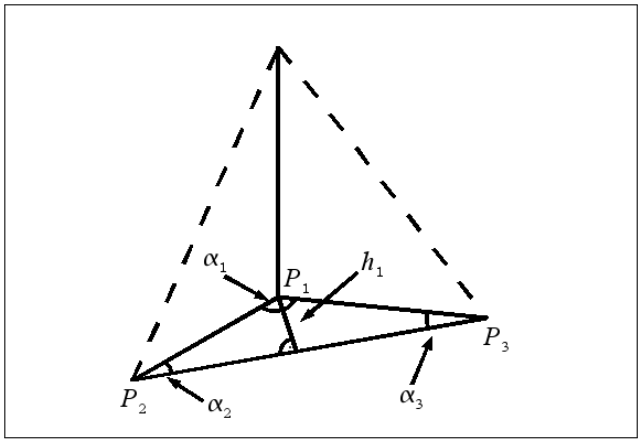


Figure 4.4: Hat function φ_1

Hence, it follows that

$$|\nabla \varphi_1|_2 \leq \frac{1}{h_1} = \frac{1}{\sin(\alpha_2) \cdot |P_1 - P_2|_2} \leq \frac{1}{\sin(\alpha_2) \cdot \sin(\alpha_3) \cdot h_T} \leq \frac{1}{\sin^2(\beta) \cdot h_T} \tag{4.18}$$

Same estimate for φ_2 and φ_3.

Since x lies inside the triangle with vertices P_1, P_2, P_3, it follows that

$$\begin{aligned}
\nabla \psi(x) &\overset{(4.9)}{=} \sum_{k=1}^{3} u(P_k) \nabla \varphi_k(x) \\
&\overset{(4.12)}{=} \sum_{k=1}^{3} \Big(u(x) + (\nabla u(x))^T d_k + R_k(x) \Big) \nabla \varphi_k(x) \\
&= u(x) \sum_{k=1}^{3} \nabla \varphi_k(x) + \underbrace{\left[\sum_{k=1}^{3} \nabla \varphi_k(x) d_k^T \right]}_{2 \times 2} \nabla u(x) + \sum_{k=1}^{3} R_k(x) \nabla \varphi_k(x)
\end{aligned} \tag{4.19}$$

With $\sum_{k=1}^{3} \varphi_k(x) \overset{(4.14)}{=} 1$ and $\sum_{k=1}^{3} P_k \varphi_k(x) \overset{(4.15)}{=} x$ we obtain that

$$\sum_{k=1}^{3} \nabla \varphi_k(x) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad \sum_{k=1}^{3} \nabla \varphi_k(x) P_k^T = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

and hence

$$\sum_{k=1}^{3} \nabla\varphi_k(x) d_k^T = \sum_{k=1}^{3} \nabla\varphi_k(x) P_k^T - \left(\sum_{k=1}^{3} \nabla\varphi_k(x) \right) x^T = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

This simplifies (4.19) to

$$\nabla\psi(x) = \nabla u(x) + \sum_{k=1}^{3} R_k(x) \nabla\varphi_k(x).$$

The error can now be bounded as follows:

$$\begin{aligned} |\nabla\psi(x) - \nabla u(x)|_2^2 &\le \Big| \sum_{k=1}^{3} R_k(x) \nabla\varphi_k(x) \Big|_2^2 \\ &\le \sum_{k=1}^{3} R_k^2(x) \cdot \sum_{k=1}^{3} |\nabla\varphi_k(x)|_2^2 \qquad \text{Cauchy-Schwarz in } \mathbb{R}^3 \\ &\overset{(4.18)}{\le} \sum_{k=1}^{3} R_k^2(x) \cdot \frac{3}{\sin^4(\beta) \cdot h_T^2} \end{aligned}$$

Integrate:

$$|\psi - u|_{H^1(T)}^2 \le \frac{3}{\sin^4(\beta)} \cdot \frac{1}{h_T^2} \cdot \sum_{k=1}^{3} I_k \overset{(4.17)}{\le} \frac{9}{8\sin^4(\beta)} h_T^2 \|u\|_{H^2(T)}^2$$

Summing up over all triangles yields

$$|\psi - u|_{H^1}^2 = \sum_T |\psi - u|_{H^1(T)}^2 \le \frac{9}{8\sin^4(\beta)}\ h^2\ \|u\|_{H^2}^2$$

Taking the square root yields the assertion (4.11). ∎

(c) Error bounds for a model problem

Consider the boundary value problem

$$\begin{aligned} -\partial_{x_1}(\kappa\partial_{x_1} u) - \partial_{x_2}(\kappa\partial_{x_2} u) + \kappa_0 u &= f \qquad \text{in } \Omega \\ u &= 0 \qquad \text{on } \Gamma \end{aligned}$$

with $\kappa, \kappa_0 : \Omega \longrightarrow \mathbb{R}$. This is a special case of the problem considered in Section 4.6.

Variational formulation: Find $u \in H_0^1(\Omega)$ such that

$$a(u, v) = \ell(v) \qquad \text{for all } v \in H_0^1(\Omega). \tag{4.20}$$

where

$$a(u,v) = \int_\Omega \kappa\big(\partial_{x_1} u \cdot \partial_{x_1} v + \partial_{x_2} u \cdot \partial_{x_2} v\big)\, dx + \int_\Omega \kappa_0 uv\, dx$$
$$\ell(v) = \int_\Omega fv\, dx$$

Assumptions:

- $\Omega \subset \mathbb{R}^2$ bounded, convex, Γ polygon
- $\kappa, \kappa_0 \in L_\infty(\Omega)$ and there are constants $c_{\min}, c_{\max} < \infty$ such that
  - $\kappa, \kappa_0 \leq c_{\max}$ a.e.,
  - $0 \leq \kappa_0(x)$ a.e. and $0 < c_{\min} \leq \kappa(x)$ a.e..
- $f \in L_2(\Omega)$

We know from Section 4.6 that $a(\cdot,\cdot)$ is H_0^1-elliptic, and that Lax-Milgram can be applied $\Longrightarrow$ unique solution. For the error analysis, however, we need more regularity:

Lemma 4.8.3 *In addition to the previous assumptions, we assume that $\kappa \in C^1(\overline{\Omega})$. Then, the weak formulation (4.20) has a unique solution $u \in H^2(\Omega) \cap H_0^1(\Omega)$, and there is a constant $c = c(\Omega) > 0$ such that*

$$\|u\|_{H^2} \leq c \cdot \|f\|_{L_2}$$

(without proof)

Theorem 4.8.4 (H^1 error of the FEM approximation)
Let $\kappa \in C^1(\overline{\Omega})$ and let $\Omega, f, \kappa_0, a, \ell$ be as before. Let V_h be the finite element space of piecewise linear functions on a triangulation, where $h > 0$ is the length of the longest edge. Let $u_h \in V_h$ be the finite element approximation of u, i.e. $u_h \in V_h$ is the solution of

$$a(u_h, v_h) = \ell(v_h) \qquad \text{for all } v_h \in V_h.$$

Then, there is a constant $c > 0$ such that

$$\|u - u_h\|_{H^1} \leq ch\|u\|_{H^2}.$$

Proof. Let $\psi(x) = \sum_{k=1}^N u(P_k)\varphi_k(x) \in V_h$ be the interpolant from Theorem 4.8.2.

$$\begin{aligned}
\|u - u_h\|_{H^1} &\leq C\|u - u_h\|_a && (H_0^1\text{-ellipticity}) \\
&= C \min_{v_h \in V_h} \|u - v_h\|_a && (\text{Céa's Lemma, Lemma 4.8.1}) \\
&\leq C\|u - \psi\|_a && (\text{choose } v_h = \psi \in V_h) \\
&\leq \tilde{C}\|u - \psi\|_{H^1} && (H_0^1\text{-ellipticity}) \\
&\leq \widehat{C}h\|u\|_{H^2} && (\text{Theorem 4.8.2})
\end{aligned}$$

■

In this result the error is measured in the norm $\|\cdot\|_{H^1}$. In applications, however, often the error in $\|\cdot\|_{L_2}$ is more important.

Theorem 4.8.5 (L_2 error of the FEM approximation)
Under the assumptions of Theorem 4.8.2, there is constant $C > 0$ such that

$$\|u - u_h\|_{L_2} \le Ch^2 \|u\|_{H^2}.$$

Remark: Note that now a different order of convergence is obtained:

Error in $\|\cdot\|_{H^1}$	$\longleftrightarrow$	order 1
Error in $\|\cdot\|_{L_2}$	$\longleftrightarrow$	order 2

Proof ("Nitsche's trick"). Let $w \in H_0^1(\Omega)$ and $w_h \in V_h$ be the solutions of

$$\begin{aligned} a(w, v) &= \widetilde{\ell}(v) \qquad \text{for all } v \in H_0^1(\Omega) \\ a(w_h, v_h) &= \widetilde{\ell}(v_h) \qquad \text{for all } v_h \in V_h \end{aligned}$$

with

$$\widetilde{\ell}(v) = \int_\Omega \big(u(x) - u_h(x)\big) v(x)\, dx.$$

Since $a(u - u_h, v_h) = 0$ for all $v_h \in V_h$ we have in particular

$$a(u - u_h, w_h) = 0 \tag{4.21}$$

and hence

$$\begin{aligned} \|u - u_h\|_{L_2}^2 &= \int_\Omega (u(x) - u_h(x))^2\, dx \\ &= \underbrace{\int_\Omega (u(x) - u_h(x)) u(x)\, dx}_{=\widetilde{\ell}(u)} - \underbrace{\int_\Omega (u(x) - u_h(x)) u_h(x)\, dx}_{=\widetilde{\ell}(u_h)} \\ &= a(w, u) - a(w, u_h) \\ &= a(w, u - u_h) - \underbrace{a(w_h, u - u_h)}_{=0, \quad \text{see (4.21)}} \\ &= a(w - w_h, u - u_h) \\ &\le C \cdot \|w - w_h\|_{H^1} \cdot \|u - u_h\|_{H^1} \end{aligned}$$

because $a(\cdot, \cdot)$ is H_0^1-elliptic. Theorem 4.8.4 yields

$$\|w - w_h\|_{H^1} \le C \cdot h \cdot \|w\|_{H^2}, \qquad \|u - u_h\|_{H^1} \le C \cdot h \cdot \|u\|_{H^2},$$

and by Lemma 4.8.3

$$\|w\|_{H^2} \leq C \cdot \|u - u_h\|_{L_2}.$$

Hence, in total we have

$$\|u - u_h\|_{L_2}^2 \leq Ch^2 \|w\|_{H^2} \cdot \|u\|_{H^2} \leq Ch^2 \|u - u_h\|_{L_2} \cdot \|u\|_{H^2},$$

and dividing by $\|u - u_h\|_{L_2}$ yields the assertion. ■

Remark: If Ω is a rectangle, then the boundary value problem could also be solved with finite differences. The convergence order would be the same for classical second-order finite differences with mesh width $h > 0$. However, convergence proofs for finite differences require $u \in C^4(\overline{\Omega})$, which is a considerably stronger regularity assumption.

Chapter 5

The Finite Element Method for parabolic PDEs

5.1 Model problem and weak formulation

Consider the parabolic problem

$$\partial_t u = \partial_{x_1}\big(\kappa\partial_{x_1}u\big) + \partial_{x_2}\big(\kappa\partial_{x_2}u\big) - \kappa_0 u + f \qquad \forall x \in \Omega,\ t \in (0,T] \tag{5.1a}$$

$$u(t,x) = 0 \qquad \forall x \in \Gamma,\ t \in (0,T] \tag{5.1b}$$

$$u(0,x) = u_0(x) \qquad \forall x \in \Omega \tag{5.1c}$$

with sufficiently smooth functions $\kappa(x)$, $\kappa_0(x)$, and $f(t,x)$. For $\kappa(x) \equiv 1$, $\kappa_0(x) \equiv f(t,x) \equiv 0$ we obtain the heat equation. For $f(t,x) = f(x)$ the stationary solution ($\partial_t u = 0$) is the solution of the elliptic boundary value problem considered in the previous chapter.
A function $u = u(t,x)$ is called a **classical solution** of (5.1) if $\partial_t u,\ \partial_{x_i} u,\ \partial^2_{x_i} u,\ i \in \{1,2\}$ exist in the classical sense and (5.1a)-(5.1c) are true.

x Multiply both sides of (5.1a) with a test function $v \in C_c^\infty(\Omega)$ and apply Green's formula:

$$\int_\Omega \partial_t u \cdot v\ dx = -\int_\Omega \kappa(\partial_{x_1}u \cdot \partial_{x_1}v + \partial_{x_2}u \cdot \partial_{x_2}v)\ dx - \int_\Omega \kappa_0 uv\ dx + \int_\Omega fv\ dx. \tag{5.2a}$$

For fixed t, the function $x \mapsto u(t,x)$ is denoted by $u(t,\cdot)$ or simply $u(t)$. This function is considered as a parameter-dependent element in a suitable function space such as, e.g., $H^1(\Omega)$.

Definition 5.1.1 (weak solution) *Let $a(\cdot,\cdot)$ be the bilinear form*

$$a(w,v) = \int_\Omega \kappa\big(\partial_{x_1}w \cdot \partial_{x_1}v + \partial_{x_2}w \cdot \partial_{x_2}v\big)\ dx + \int_\Omega \kappa_0 wv\ dx.$$

A function $u : [0,T] \times \Omega \longrightarrow \mathbb{R}$ is called weak solution of (5.1), if

- $u(t) \in L_2(\Omega)$ *for all* $t \in [0,T]$
- $u(t) \in H_0^1(\Omega)$ *for almost all* $t \in (0,T]$*, and for those* t

$$\langle \partial_t u(t), v\rangle_{L_2} = -a\big(u(t), v\big) + \langle f(t), v\rangle_{L_2} \qquad \text{for all } v \in H_0^1(\Omega),$$

- $\lim_{t \searrow 0} \|u(t) - u_0\|_{L_2} = 0.$

Assumptions:
(A1) $\Omega \subset \mathbb{R}^2$ bounded, convex, Γ polygon
(A2) $\kappa \in C^1(\bar{\Omega})$ with $0 < c_{\min} \leq \kappa(x) \leq c_{\max} < \infty$ a.e..
(A3) $\kappa_0 \in L_\infty(\Omega)$ and $0 \leq \kappa_0(x) \leq c_{\max}$.
(A4) $f \in L_2\big((0,T)\times\Omega\big)$
Under these conditions it can be shown that for every $f \in L_2\big((0,T)\times\Omega\big)$ and $u_0 \in H_0^1(\Omega)$, there is a weak solution of (5.1) such that

- $u(t) \in H_0^1(\Omega)$ for all $t \in [0,T]$
- $u(t) \in H^2(\Omega)$ for almost all $t \in [0,T]$
- $\partial_t u(t) \in L_2(\Omega)$ for almost all $t \in [0,T]$

(see Satz 97.2 in [HB09]).

Remark. The existence of a **classical** solution requires stronger assumptions.

5.2 Approximation with finite elements

Method of lines (cf. part I of the lecture):

$$\textbf{PDE} \longrightarrow \textbf{weak formulation} \xrightarrow[\text{discretization}]{\text{space}} \textbf{ODE} \xrightarrow[\text{discretizaion}]{\text{time}} \textbf{numerical approximation}$$

Consider a triangulation with triangles $T_1, \dots, T_m$, interior points $P^1, \dots, P^N$ and maximal edge of length $h > 0$. Let φ_j be the usual hat functions

$$\varphi_i\big|_{T_k} \text{ is linear for all } k, \text{ and } \varphi_i(P_j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{else,} \end{cases}$$

and let $V_h = \operatorname{span}\{\varphi_1, \dots, \varphi_N\}$ be the finite element space of piecewise linear functions.

Galerkin ansatz: Find $u_h(t,x) = \sum_{j=1}^N \widehat{u}_j(t)\varphi_j(x) \in V_h$ such that

$$\langle \partial_t u_h(t), v_h\rangle_{L_2} = -a\big(u_h(t), v_h\big) + \langle f(t), v_h\rangle_{L_2} \qquad \forall v_h \in V_h,\ t \in [0,T] \tag{5.3a}$$

$$u_h(0) = \widetilde{u}_0 \tag{5.3b}$$

where $\widetilde{u}_0 \in V_h$ is the interpolation of $u_0 \notin V_h$:

$$\widetilde{u}_0(x) = \sum_{j=1}^{N} u_0(P_j)\varphi_j(x).$$

With the representation $v_h = \sum_{j=1}^{N} \hat{v}_j \varphi_j$, (5.3) is equivalent to the ODE

$$M\widehat{u}'(t) = -L\widehat{u}(t) + b(t) \tag{5.4a}$$

$$\widehat{u}(0) = \widehat{u}^{(0)} \tag{5.4b}$$

with

$$\begin{aligned}
\widehat{u}(t) &= (\widehat{u}_1(t), \dots, \widehat{u}_N(t))^T \\
\widehat{u}^{(0)} &= \big(u_0(P_1), \dots, u_0(P_N)\big)^T \\
M &= (M_{ij})_{i,j=1}^N, & M_{ij} &= \langle \varphi_i, \varphi_j \rangle_{L_2} & &\text{mass matrix, s.p.d} \\
L &= (L_{ij})_{i,j=1}^N, & L_{ij} &= a(\varphi_i, \varphi_j) & &\text{stiffness matrix, s.p.d.} \\
b(t) &= \big(b_1(t), \dots, b_N(t)\big)^T, & b_i(t) &= \langle f(t), \varphi_i \rangle_{L_2}
\end{aligned}$$

5.3 Accuracy

(a) Space discretization

Theorem 5.3.1 (Error of the space discretization with FEM)
Assume (A1)-(A4) and, in addition, that $u_0 \in H_0^1(\Omega) \cap H^2(\Omega)$ and $\partial_t u \in H^2(\Omega)$ for almost all $t \in [0,T]$. Let u be the weak solution, i.e.

$$\langle \partial_t u(t), v \rangle_{L_2} = -a\big(u(t), v\big) + \langle f(t), v \rangle_{L_2} \quad \text{for all } v \in H_0^1(\Omega) \text{ and a.a.t} \in [0,T],$$
$$\lim_{t \searrow 0} \|u(t) - u_0\|_{L_2} = 0.$$

Let u_h be the solution of (5.3a)-(5.3b). Then, there is a constant $C > 0$ such that the error of the space discretization is bounded by

$$\|u_h(t) - u(t)\|_{L_2} \le Ch^2 \left(\|u_0\|_{H^2} + \int_0^t \|\partial_t u(s)\|_{H^2}\, ds \right).$$

Proof.

1. For given $w \in H_0^1(\Omega)$ let $w_h \in V_h$ be the unique solution of

   $$a(w_h, v_h) = a(w, v_h) \qquad \forall v_h \in V_h.$$

   The existence and uniqueness follows from Lax-Milgram, because the linear form $v_h \mapsto a(w, v_h)$ is continuous. The mapping

   $$R_h : H_0^1(\Omega) \to V_h, \qquad R_h : w \mapsto w_h$$

   is called the **Ritz projection**.

Picture

If in addition $w \in H^2(\Omega)$, then Theorem 4.8.5 implies

$$\|w - R_h w\|_{L_2} \leq Ch^2 \|w\|_{H^2}$$

(choose $\ell(v) = a(w, v)$ for given w).

2. Let $R_h u(t)$ be the Ritz projection of $u(t)$, i.e.

$$a(R_h u(t), v_h) = a(u(t), v_h) = \langle f(t) - \partial_t u(t), v_h \rangle_{L_2} \qquad \forall v_h \in V_h. \tag{5.5}$$

By assumption the time derivative $\partial_t u$ exists in $H^2(\Omega)$ for almost all t, and

$$a(\partial_t R_h u(t), v_h) = a(\partial_t u(t), v_h) \qquad \forall v_h \in V_h.$$

Hence, $R_h \partial_t u(t) = \partial_t R_h u(t)$ is the Ritz projection of $\partial_t u$.

3. Define $d_h(t) = R_h u(t) - u_h(t)$. Then for all $v_h \in V_h$ we have

$$\begin{aligned} &\langle \partial_t d_h(t), v_h \rangle_{L_2} + a(d_h(t), v_h) \\ &= \langle \partial_t R_h u(t) - \partial_t u_h(t), v_h \rangle_{L_2} + a(R_h u(t), v_h) - a(u_h(t), v_h). \end{aligned}$$

Substituting

$$\begin{aligned} a(R_h u(t), v_h) &\overset{(5.5)}{=} \langle f(t) - \partial_t u(t), v_h \rangle_{L_2} \\ a(u_h(t), v_h) &\overset{(5.3a)}{=} \langle f(t) - \partial_t u_h(t), v_h \rangle_{L_2} \end{aligned}$$

and cancelling terms yields

$$\langle \partial_t d_h(t), v_h \rangle_{L_2} + a(d_h(t), v_h) = \langle \partial_t R_h u(t) - \partial_t u(t), v_h \rangle_{L_2}.$$

4. If we choose $v_h := d_h(t)$, then

$$\langle \partial_t d_h(t), d_h(t) \rangle_{L_2} + \underbrace{a(d_h(t), d_h(t))}_{\geq 0} = \langle \partial_t R_h u(t) - \partial_t u(t), d_h(t) \rangle_{L_2}.$$

It follows that for almost all $t \in [0, T]$

$$\begin{aligned} \|d_h(t)\|_{L_2} \cdot \frac{d}{dt} \|d_h(t)\|_{L_2} &= \frac{1}{2} \frac{d}{dt} \|d_h(t)\|_{L_2}^2 \\ &= \langle \partial_t d_h(t), d_h(t) \rangle_{L_2} \\ &\leq \langle \partial_t R_h u(t) - \partial_t u(t), d_h(t) \rangle_{L_2} \\ &\leq \|\partial_t R_h u(t) - \partial_t u(t)\|_{L_2(\Omega)} \cdot \|d_h(t)\|_{L_2} \end{aligned}$$

Without loss of generality, we assume that $\|d_h(t)\|_{L_2} > 0$ for all $t \in (0, T]$ and divide by $\|d_h(t)\|_{L_2}$. This yields

$$\frac{d}{dt}\|d_h(t)\|_{L_2} \leq \|\partial_t R_h u(t) - \partial_t u(t)\|_{L_2},$$

and after integrating from 0 to t, we obtain

$$\|d_h(t)\|_{L_2} \leq \|d_h(0)\|_{L_2} + \int_0^t \|\partial_t R_h u(s) - \partial_t u(s)\|_{L_2}\, ds$$

5. According to 1. the inequalites

$$\begin{aligned} \|u(t) - R_h u(t)\|_{L_2} &\leq Ch^2 \|u(t)\|_{H^2} \\ \|\partial_t u(t) - \partial_t R_h u(t)\|_{L_2} &\leq Ch^2 \|\partial_t u\|_{H^2} \end{aligned}$$

hold for almost all $t \in [0, T]$. By definition of $d_h(t) = R_h u(t) - u_h(t)$ and by 4., we obtain

$$\begin{aligned} \|u(t) - u_h(t)\|_{L_2} &\leq \|u(t) - R_h u(t)\|_{L_2} + \|d_h(t)\|_{L_2} \\ &\leq Ch^2 \|u(t)\|_{H^2} + \|d_h(0)\|_{L_2} + Ch^2 \int_0^t \|\partial_t u(s)\|_{H^2}\, ds \end{aligned}$$

Since

$$\begin{aligned} \|d_h(0)\|_{L_2} &= \big\| \big(R_h u(0) - u_0\big) + \big(u_0 - u_h(0)\big) \big\|_{L_2} \\ &\leq \underbrace{\|R_h u_0 - u_0\|_{L_2}}_{\leq Ch^2 \|u_0\|_{H^2}} + \underbrace{\|u_0 - \widetilde{u}_0\|_{L_2}}_{\leq Ch^2 \|u_0\|_{H^2} \text{ (Th.4.8.2)}} \leq Ch^2 \|u_0\|_{H^2} \end{aligned}$$

and

$$u(t) = u_0 + \int_0^t \partial_t u(s)\, ds$$

we finally obtain

$$\|u(t) - u_h(t)\|_{L_2} \leq Ch^2 \left(\|u_0\|_{H^2} + \int_0^t \|\partial_t u(s)\|_{H^2}\, ds \right). \qquad \blacksquare$$

(b) Time discretization

The space discretization turns the parabolic PDE into the ODE (5.4):

$$\begin{aligned} M\widehat{u}'(t) &= -L\widehat{u}(t) + b(t) \qquad\qquad t \in (0, T) \\ \widehat{u}(0) &= \widehat{u}^0 \end{aligned} \tag{5.4}$$

$M, L \in \mathbb{R}^{N\times N}$ are symmetric positive definite matrices and $b(t) \in \mathbb{R}^N$. Apply the implicit Euler method

$$(M + \tau L)\widehat{u}^{n+1} = M\widehat{u}^n + \tau b(t_{n+1}) \qquad n = 0, \ldots, n_{\max} - 1 \tag{5.6}$$

with step-size $\tau := T/n_{\max} > 0$ to obtain approximations $\widehat{u}^n \approx \widehat{u}(t_n)$ at $t_n = n\tau$. The linear problems have a unique solution because $M + cL$ is symmetric and positive definite for all $c \geq 0$.

Accuracy? In 7.3 (c) of part I of the lecture, we have shown the following:

Theorem 5.3.2 (Error bound for the implicit Euler method) *Let $(\cdot \mid \cdot)$ be an arbitrary scalar product on $\mathbb{R}^d$ with induced norm $|||z||| = \sqrt{(z \mid z)}$. Let $F : [0,T]\times\mathbb{R}^d \longrightarrow \mathbb{R}^d$ be C^1 and assume that F satisfies the one-sided Lipschitz condition*

$$\Big(F(t,z) - F(t,\tilde{z}) \mid z - \tilde{z}\Big) \leq \lambda \,|||\, z - \tilde{z} \,|||^2 \qquad \forall t \in [0,T] \quad \forall z, \tilde{z} \in \mathbb{R}^d \tag{5.7}$$

with a constant $\lambda < 0$. Let $y(t)$ be the exact solution of the initial-value problem

$$\dot{y} = F(t,y), \quad t \in [0,T], \qquad y(0) = y_0,$$

and let $y^n \approx y(t_n)$ be the approximations computed with the implicit Euler method with step-size $\tau = T/n_{max}$. Then, the error is bounded by

$$\max_{n=0,\ldots,n_{max}} |||y^n - y(t_n)||| \leq \frac{C}{2}\tau \cdot \max_{t\in[0,T]} |||y''(t)|||$$

with $C := \min\{T, 1/|\lambda|\}$.

Remarks: In part I of the lecture, we have also considered the cases $\lambda = 0$ and $\lambda > 0$. The proof was given for the Euclidean scalar product, but the arguments remain true for arbitrary inner products.

Apply Theorem 5.3.2 to (5.4). Show that

$$F(t,\widehat{v}) = M^{-1}\big(b(t) - L\widehat{v}\big)$$

fulfills the one-sided Lipschitz condition if we choose

$$(z \mid \tilde{z}) := z^T M\tilde{z}, \qquad\qquad |||z||| = \sqrt{z^T Mz}.$$

For every $\widehat{v}, \widehat{w} \in \mathbb{R}^N$ it follows from $M^T = M$ and $L^T = L = \big(a(\varphi_i, \varphi_j)\big)_{ij}$ that

$$\begin{aligned}\big(F(t,\widehat{v}) - F(t,\widehat{w}) \mid \widehat{v} - \widehat{w}\big) &= -\big(M^{-1}L(\widehat{v} - \widehat{w}) \mid \widehat{v} - \widehat{w}\big) \\ &= -(\widehat{v} - \widehat{w})^T L^T M^{-T} M(\widehat{v} - \widehat{w}) \\ &= -(\widehat{v} - \widehat{w})^T L(\widehat{v} - \widehat{w}) \\ &= -a(v - w, v - w)\end{aligned}$$

with

$$v := \sum_{j=1}^{N} \widehat{v}_j \varphi_j \qquad \text{and} \qquad w := \sum_{j=1}^{N} \widehat{w}_j \varphi_j.$$

Since $a(\cdot,\cdot)$ is $H_0^1(\Omega)$-elliptic, we know that there is an $\alpha > 0$ such that

$$\begin{aligned} a(v-w, v-w) &\geq \alpha \|v-w\|_{H^1}^2 \geq \alpha \|v-w\|_{L_2}^2 \\ &= \alpha \sum_{i=1}^{N} \sum_{j=1}^{N} (\widehat{v}_i - \widehat{w}_i)(\widehat{v}_j - \widehat{w}_j) \int_\Omega \varphi_i(x)\varphi_j(x)\, dx \\ &= \alpha (\widehat{v} - \widehat{w})^T M (\widehat{v} - \widehat{w}) \\ &= \alpha \left|\!\left|\!\left| \widehat{v} - \widehat{w} \right|\!\right|\!\right|^2 . \end{aligned}$$

This yields the one-sided Lipschitz bound

$$\big(F(t,\widehat{v}) - F(t,\widehat{w}), \widehat{v} - \widehat{w}\big) \leq -\alpha \left|\!\left|\!\left| \widehat{u} - \widehat{w} \right|\!\right|\!\right|^2 .$$

If $b(t)$ is sufficiently smooth, then Theorem 5.3.2 can be applied and yields

$$\max_{n=0,\ldots,n_{\max}} \left|\!\left|\!\left| \widehat{u}^n - \widehat{u}(t_n) \right|\!\right|\!\right| \leq c\tau \cdot \max_{t\in[0,T]} \left|\!\left|\!\left| \widehat{u}''(t) \right|\!\right|\!\right|$$

with a constant $c > 0$ which does not depend on the space discretization. Note that the norm $|||\cdot|||$ depends on the space discretization, but the constant α does *not*.

Now let $u_h(t,x) = \sum_{j=1}^{N} \widehat{u}_j(t)\varphi_j(x) \in V_h$ be the solution of (5.3), i.e.

$$\begin{aligned} \langle \partial_t u_h(t), v_h \rangle_{L_2} &= -a\big(u_h(t), v_h\big) + \langle f(t), v_h \rangle_{L_2} \qquad \forall v_h \in V_h,\ t \in [0,T] \\ u_h(0) &= \widetilde{u}_0, \end{aligned}$$

and let

$$u_h^n(x) = \sum_{j=1}^{N} \widehat{u}_j^n \varphi_j(x) \approx u_h(t_n, x)$$

be the corresponding approximation where $\widehat{u}_j^n \approx \widehat{u}_j(t_n)$ are approximated with the implicit Euler method (5.6).

Corollary 5.3.3 (Error of the time discr. with the implicit Euler method) *In addition to (A1)-(A4), assume that $t \mapsto f(t,\cdot)$ is in $C^1([0,T], L_2(\Omega))$, and that $t \mapsto \partial_t^2 u_h(t,\cdot)$ is a continuous mapping from $[0,T]$ to $L_2(\Omega)$. Then there is a constant c which does not depend on h such that*

$$\|u_h^n(\cdot) - u_h(t_n,\cdot)\|_{L_2} \leq c\tau \cdot \max_{t\in[0,T]} \|\partial_t^2 u_h(t,\cdot)\|_{L_2}.$$

Proof. This follows directly from Theorem 5.3.2 because by definition

$$\|u_h^n(\cdot) - u_h(t_n,\cdot)\|_{L_2} = \left|\!\left|\!\left| \widehat{u}^n - \widehat{u}(t_n) \right|\!\right|\!\right|, \qquad \|\partial_t^2 u_h(t,\cdot)\|_{L_2} = \left|\!\left|\!\left| \widehat{u}''(t) \right|\!\right|\!\right| . \qquad \blacksquare$$

Summary: Combining Theorem 5.3.1 and Corollary 5.3.3 yields (under suitable assumptions) the error bound

$$\|u_h^n(\cdot) - u(t_n, \cdot)\|_{L_2} \leq C(\tau + h^2)$$

for the total error in space and time.

Higher-order methods for the time-discretization. Of course, the implicit Euler method could be replaced by higher-order Runge-Kutta or multistep methods in order to improve the accuracy. However, some care is required. Classical theory states that if such a method is applied to the ODE $\dot{y} = F(t, y)$ and F is sufficiently smooth, then

$$|y(t_n) - y^n| \leq C\tau^p \qquad \text{for all } n = 0, \ldots, n_{\max} \tag{5.8}$$

where p is the order of the method. Unfortunately, such a result **cannot be applied** if the ODE originates from the space discretization of a PDE, because then F depends on the variable h of the space discretization. If $h \to 0$ (i.e. if the spatial approximation is improved), the constant $C = C_h$ in (5.8) will typically tend to infinity, such that the error bound becomes worthless. A reasonable error analysis has to account for the fact that F is related to an (unbounded) differential operator. In this situation, order reduction can be observed for certain methods, i.e. the observed order of convergence is lower than the order predicted by ODE theory (cf. XVII, 101, pages 743-754 in [HB09]).

5.4 Application to a double barrier basket call

Let $V(t, S_1, \ldots, S_d)$ be the value of a European double barrier basket call with d underlyings; cf. Section 4.1. Hence, V is the solution of the d-dimensional Black-Scholes equation

$$\partial_t V + \frac{1}{2}\sum_{i,j=1}^{d} \beta_{ij} S_i S_j \partial_{S_i}\partial_{S_j} V + r\sum_{j=1}^{d} S_j \partial_{S_j} V - rV = 0, \qquad \begin{array}{l} S = (S_1, \ldots, S_d) \in \Omega \\ t \in [0, T] \end{array}$$

with interest rate $r \geq 0$, volatilities $\sigma_i \geq 0$, correlation coefficients ρ_{ij} and $\beta_{ij} := \rho_{ij}\sigma_i\sigma_j$. The PDE has to be solved on the domain

$$\Omega = \left\{ S \in \mathbb{R}_+^d : \sum_{j=1}^{d} a_j S_j > a_0 \text{ and } \sum_{j=1}^{d} b_j S_j < b_0 \right\}$$

with boundary

$$\Gamma = \bigcup_{i=0}^{d} \Gamma_i, \qquad \Gamma_0 := \left\{ S \in \bar{\Omega} : \sum_{j=1}^{d} a_j S_j = a_0 \text{ or } \sum_{j=1}^{d} b_j S_j = b_0 \right\},$$
$$\Gamma_i := \left\{ S \in \bar{\Omega} : S_i = 0 \right\}.$$

We impose the boundary condition

$$V(t,S) = 0 \qquad \text{for } S \in \Gamma_0.$$

On Γ_i with $i \in \{1,\dots,d\}$ we do not need to impose any boundary conditions, because all terms involving $\partial_{S_i} V$ or $\partial^2_{S_i} V$ vanish, and the PDE reduces to a Black-Scholes equation in $d-1$ spatial dimensions.
Since we consider a European basket call, the terminal condition is

$$V(T,S_1,\dots,S_d) = \Big(\sum_{i=1}^d \alpha_i S_i - K\Big)^+ \qquad \text{for } S \in \Omega$$

with strike $K > 0$ and weights $\alpha_1,\dots,\alpha_d > 0$. Note that the terminal condition does *not* agree with the boundary conditions.
We change to "time to maturity", i.e. we consider the function $u(t,S) = V(T-t,S)$. After defining

$$\begin{aligned} \kappa_{ij}(S) &= \tfrac{1}{2}\beta_{ij} S_i S_j, & \kappa &= (\kappa_{ij})_{i,j} \in \mathbb{R}^{d\times d}, \\ \mu_j(S) &= \Big(r - \beta_{jj} - \frac{1}{2}\sum_{i\neq j}\beta_{ij}\Big) S_j, & \mu &= (\mu_1,\dots,\mu_d)^T \in \mathbb{R}^d, \end{aligned}$$

u is the solution of

$$\begin{aligned} \partial_t u &= \operatorname{div}(\kappa\nabla u) + \mu^T \nabla u - ru, && S \in \Omega, t \in (0,T], \\ u(t,S) &= 0, && S \in \Gamma_0, t \in [0,T], \\ u(0,S) &= \Big(\sum_{i=1}^d \alpha_i S_i - K\Big)^+ =: u_0(S), && S \in \Omega. \end{aligned}$$

(The reader should check that this is true.) The differential operators div and ∇ act only on the spatial variables S_i, not on t.

Variational formulation. Multiply both sides of the PDE with a test function $w \in C^\infty(\bar\Omega)$ with $w(S) = 0$ for $S \in \Gamma_0$. Integrating and applying Green's identity (Lemma 4.2.2) yields

$$\begin{aligned} \int_\Omega \partial_t u \cdot w \, dS &= \int_\Gamma \eta^T(\kappa\nabla u) w \, d\sigma(S) - \int_\Omega (\nabla w)^T \kappa \nabla u \, dS + \int_\Omega \mu^T \nabla u \cdot w \, dS - r\int_\Omega uw \, dS \\ &= -a(u,w) \end{aligned}$$

where $\eta(x) = (\eta_1(x),\dots,\eta_d(x))^T$ denotes again the outer unit normal vector in $x \in \Gamma$. The boundary integral $\int_\Gamma \dots d\sigma$ vanishes on Γ_0 (because $w(S) = 0$ for $S \in \Gamma_0$), but not on Γ_i with $i \geq 1$. Note that the i-th row and columns of κ are zero on Γ_i with $i \geq 1$, but not the entire matrix.

Now we let

$$H^1_\star(\Omega) = \big\{ v \in H^1(\Omega) : v|_{\Gamma_0} = 0 \quad \text{(trace operator)} \big\}$$

and seek a function $u = u(t,x)$ such that $u(t) \in L_2(\Omega)$ for all $t \in [0,T]$, $u(t) \in H^1_\star(\Omega)$ for almost all $t \in [0,T]$, and for those t

$$\langle \partial_t u(t), w \rangle_{L_2} = -a\big(u(t), w\big) \qquad \text{for all } w \in H^1_\star(\Omega)$$

with initial data $u(0) = u_0$. After space discretization with piecewise linear elements we obtain again an ODE of the form (5.4), but with $b(t) = 0$ for all t.

Chapter 6

A short introduction to Sparse Grids

Let $V(t,S) = V(t,S_1,\dots,S_d)$ be the value of a European basket option with $d \gg 1$ (e.g. $d = 10$ or $d = 20$).
Approximating $V(t,S)$ means solving a PDE on $\Omega \subseteq \mathbb{R}^d$ or computing a d-dimensional integral.
Assume that $\Omega = [a,b] \times ... \times [a,b]$ and consider a equidistant grid: Choose $m \in \mathbb{N}$, let $h = \frac{b-a}{m}$, approximate V in the grid points $(i_1 h, ..., i_d h) \in \mathbb{R}^d$ with $i_1, ..., i_d \in \{0, ..., m\}$.

Problem: This grid contains $(m+1)^d$ points and $(m-1)^d$ inner points. Regardless of the method used to approximate $V(t,S)$, the numerical costs and the memory requirements grow at least exponentially in d! For $d \gg 1$, classical methods are too expensive or fail completely. This problem is known as the **curse of dimensionality**.

Sparse grids are special space discretizations which are suitable for high-dimensional problems. The goal of this short introduction is to illustrate this concept. We will only consider the approximation of **given** functions on sparse grids – not how to solve PDEs on sparse grids. The main reference of this chapter is [BG04].

6.1 Notation

(a) Multi-indices

Let $\alpha = (\alpha_1, ..., \alpha_d) \in \mathbb{N}_0^d$ and $\beta = (\beta_1, ..., \beta_d) \in \mathbb{N}_0^d$ be multi-indices. Define

$$|\alpha|_1 := \sum_{i=1}^{d} \alpha_i \qquad |\alpha|_\infty := \max_{i=1,...,d} \alpha_i \qquad \mathbb{1} := (1, ..., 1) \in \mathbb{N}^d$$

and the following operations:

$$\begin{aligned} \alpha \cdot \beta &:= (\alpha_1\beta_1, ..., \alpha_d\beta_d), \\ c\alpha &:= (c\alpha_1, ..., c\alpha_d) \qquad \text{for any } c \in \mathbb{R} \\ 2^\alpha &:= (2^{\alpha_1}, ..., 2^{\alpha_d}). \end{aligned}$$

Moreover, we define

$$\alpha \leq \beta :\Longleftrightarrow \alpha_j \leq \beta_j \qquad \forall j = 1, ..., d$$
$$\alpha < \beta :\Longleftrightarrow \alpha \leq \beta \text{ and } \alpha \neq \beta$$

(b) Derivatives, norms, spaces

For $d \in \mathbb{N}$ let $\Omega = (0,1)^d$ with $\bar{\Omega} = [0,1]^d$ and boundary $\Gamma = \bar{\Omega} \setminus \Omega$. Let $u : \bar{\Omega} \longrightarrow \mathbb{R}$ be a smooth function with (weak) partial derivatives

$$\partial^\alpha u := \partial_{x_1}^{\alpha_1} \dots \partial_{x_d}^{\alpha_d} u;$$

cf. Section 4.5. Moreover, we define the semi-norm

$$|u|_{\alpha,\infty} := \|\partial^\alpha u\|_{L_\infty(\Omega)}, \tag{6.1}$$

and the energy norm

$$\|u\|_E = \Big(\sum_{|\alpha|_1 = 1} \|\partial^\alpha u\|_{L_2(\Omega)}^2 \Big)^{\frac{1}{2}}.$$

Define spaces of functions with bounded mixed derivatives: For $r \in \mathbb{N}_0$ let

$$X^r(\bar{\Omega}) := \Big\{ u : \bar{\Omega} \longrightarrow \mathbb{R} : \partial^\alpha u \in L_\infty(\Omega) \text{ for all } \alpha \in \mathbb{N}_0^d \text{ with } |\alpha|_\infty \leq r \Big\}$$
$$X_0^r(\bar{\Omega}) := \big\{ u \in X^r(\bar{\Omega}) : u \mid_\Gamma = 0 \big\}$$

Warning: $X^r(\bar{\Omega})$ should not be confused with the Sobolev space

$$W^{r,\infty}(\bar{\Omega}) := \Big\{ u : \bar{\Omega} \longrightarrow \mathbb{R} : \partial^\alpha u \in L_\infty(\Omega) \text{ for all } \alpha \in \mathbb{N}_0^d \text{ with } |\alpha|_1 \leq r \Big\} \supset X^r(\bar{\Omega}).$$

(c) Grids

For $\ell \in \mathbb{N}^d$ we define the mesh width

$$h_\ell := 2^{-\ell}, \qquad h_\ell = (h_{\ell_1}, ..., h_{\ell_d}), \qquad h_{\ell_j} = 2^{-\ell_j}. \tag{6.2}$$

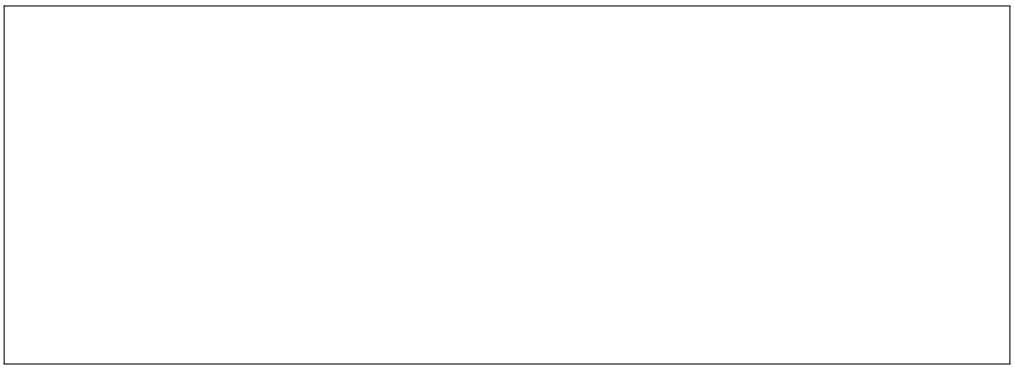
and for $i \in \mathbb{N}_0^d$ the grid points

$$x_{\ell,i} := i \cdot h_\ell \in \mathbb{R}^d, \qquad x_{\ell,i} = (x_{\ell_1,i_1}, ..., x_{\ell_d,i_d}), \qquad x_{\ell_j,i_j} = i_j \cdot h_{\ell_j}.$$

The corresponding grid on $\bar{\Omega}$ is denoted by

$$\Omega_\ell = \{ x_{\ell,i} : 0 \leq i \leq 2^\ell \}$$

Example: $d = 2$, $\ell = (2,3)$, $h_\ell = \big(\frac{1}{4}, \frac{1}{8}\big)$, $i = (3,5)$:



Our goal is to approximate given functions $u \in X_0^r(\bar{\Omega})$. Since $u\,|_\Gamma = 0$, only inner mesh points have to be considered.

(d) Piecewise linear basis functions (hat functions)

Hat functions in one dimension:

$$\varphi(\xi) := \begin{cases} 1 - |\xi| & \text{for } \xi \in [-1, 1] \\ 0 & \text{else.} \end{cases} \tag{6.3}$$

Consider dilated and shifted hat functions

$$\varphi_{\ell_j, i_j}(x_j) := \varphi\left(\frac{x_j - i_j h_{\ell_j}}{h_{\ell_j}}\right) \tag{6.4}$$

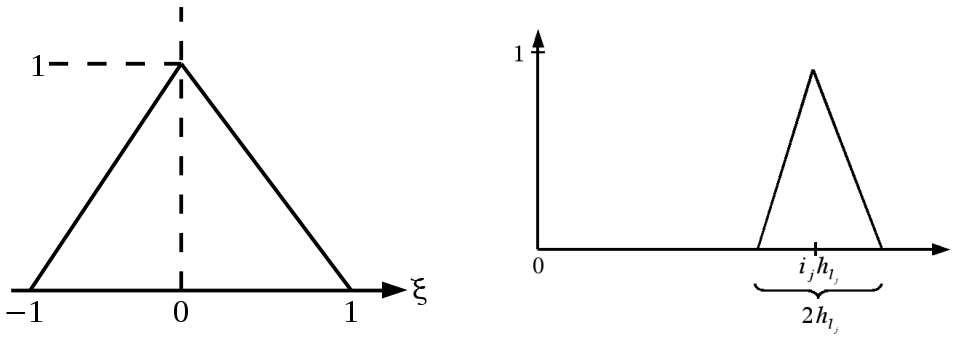


Figure 6.1: Left: Simple hat function. Right: Shifted and dilated hat function.

Multivariate basis: tensor products

$$\varphi_{\ell,i}(x) = \prod_{j=1}^{d} \varphi_{\ell_j, i_j}(x_j), \qquad \varphi_{\ell,i}(x_{\ell,j}) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{else} \end{cases}$$

(e) Nodal and hierarchical basis

Let V_ℓ be the space of piecewise d-linear functions which vanish on the boundary:

$$V_\ell = \operatorname{span}\Big\{\varphi_{\ell,i} : \mathbb{1} \leq i \leq 2^\ell - \mathbb{1}\Big\} \tag{6.5}$$

The set $\{\varphi_{\ell,i} : \mathbb{1} \leq i \leq 2^\ell - \mathbb{1}\}$ is called the **nodal basis** of V_ℓ. This is the "standard"

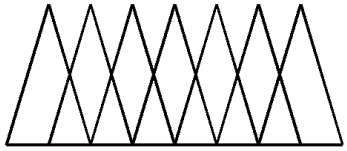


Figure 6.2: Nodal basis for $d = 1$, $l = 3$, $h = 2^{-\ell} = \frac{1}{8}$.

basis. Sparse grids, however, are based on the **hierarchical basis**. First, we define the **hierarchical increments**

$$W_\ell := \operatorname{span}\{\varphi_{\ell,i} : i \in I_\ell\}$$
$$\text{where } I_\ell := \big\{\mathbb{1} \leq i \leq 2^\ell - \mathbb{1},\ \ i_j \text{ odd},\ \ j = 1, ..., d\big\}$$

This definition implies that

$$V_\ell = \bigoplus_{k \leq \ell} W_k$$

The **hierarchical basis** of V_ℓ is the set $\{\varphi_{k,i} : i \in I_k, k \leq l\}$.

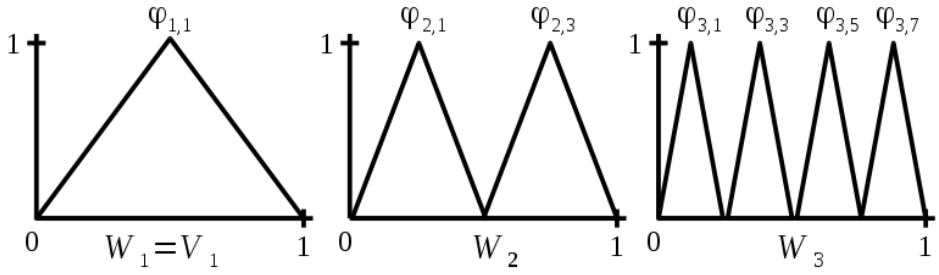


Figure 6.3: Hierarchical basis for $d = 1$, $\ell = 3$, $h = 2^{-\ell} = \frac{1}{8}$.

Next, we define the space

$$V := \sum_{\ell_1=1}^{\infty} \cdots \sum_{\ell_d=1}^{\infty} W_{(\ell_1,...,\ell_d)} = \bigoplus_{\ell \in \mathbb{N}^d} W_\ell$$

The closure of V with respect to $\|\cdot\|_{H^1(\Omega)}$ is the Sobolev space $H_0^1(\Omega)$. Every $u \in X_0^2(\bar{\Omega})$ has a unique representation

$$u(x) = \sum_{\ell \in \mathbb{N}^d} u_\ell(x) \qquad \text{with} \qquad u_\ell(x) = \sum_{i \in I_\ell} v_{\ell,i} \varphi_{\ell,i}(x) \in W_\ell. \tag{6.6}$$

Example: Approximation of continuous functions by interpolation. Let $d = 1$ and consider the function $u(x) = x(1-x)$.

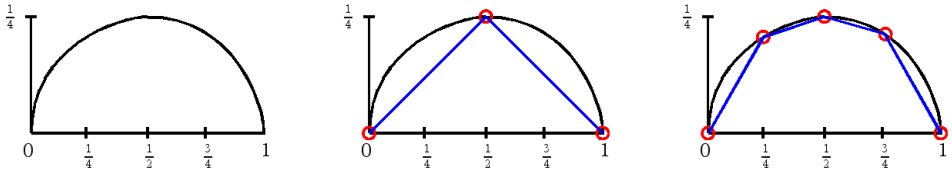


Figure 6.4: Left: The function $u(x) = x(1-x)$. Middle: Approximation in $V_1 = W_1$. Right Approximation in $V_2 = W_1 \oplus W_2$.

- Approximation in $V_1 = W_1 = \operatorname{span}\{\varphi_{1,1}\}$
  Choose $a \in \mathbb{R}$ such that

$$u(1/2) = a\varphi_{1,1}(1/2) \Longrightarrow a = 1/4 \quad \text{(piecewise linear interpolation)} \tag{6.7}$$

- Approximation in $V_2 = W_1 \oplus W_2 = \operatorname{span}\{\varphi_{1,1}, \varphi_{2,1}, \varphi_{2,3}\}$
  Choose a, b, c such that

$$u(x) = a\varphi_{1,1}(x) + b\varphi_{2,1}(x) + c\varphi_{2,3}(x) \quad \text{for } x \in \{1/4, 1/2, 3/4\}$$

$$\begin{aligned} u(1/2) &= a \cdot 1 + b \cdot 0 + c \cdot 0 && \Longrightarrow a = 1/4 \text{ (as before)} \\ \underbrace{u(1/4)}_{=3/16} &= \underbrace{a}_{=1/4} \cdot 1/2 + b \cdot 1 + c \cdot 0 && \Longrightarrow b = 1/16 \\ u(3/4) &= a \cdot (1/2) + b \cdot 0 + c \cdot 1 && \Longrightarrow c = 1/16 \end{aligned}$$

If we "add" new increment spaces W_ℓ, then the coefficients corresponding to basis elements in W_k with $k < l$ do not change. Including more basis elements means that more "details" of the function can be resolved. If the target function is smooth, the "small details" are insignificant. Hence, the corresponding coefficients have small values.

6.2 Properties of the subspaces W_ℓ

Henceforth, we consider a (given) function

$$u \in X_0^2(\bar{\Omega}) = \big\{u : \bar{\Omega} \longrightarrow \mathbb{R} : u|_\Gamma = 0 \text{ and } \partial^\alpha u \in L_\infty(\Omega) \text{ for all } |\alpha|_\infty \le 2\big\}.$$

Since $X_0^2(\bar{\Omega}) \subset V := \bigoplus_{l \in \mathbb{N}^d} W_\ell$, the function has a **hierarchical multi-level representation**

$$\boxed{u = \sum_{\ell \in \mathbb{N}^d} u_\ell \qquad \text{with} \qquad u_\ell(x) = \sum_{i \in I_\ell} v_{\ell,i} \cdot \varphi_{\ell,i}(x) \in W_\ell.} \tag{6.8}$$

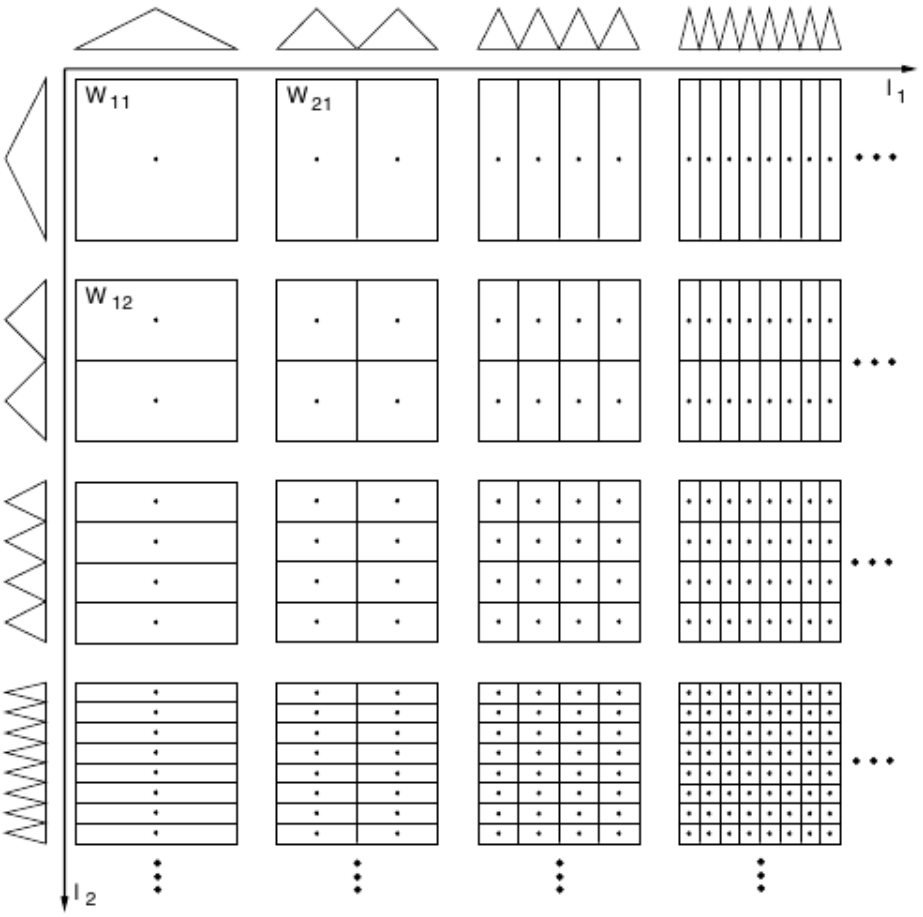


Figure 6.5: Illustration of the subspaces W_ℓ for $d = 2$. Figure stolen from [BG04].

What is the "price" of $u_\ell \in W_\ell$? What is the "value" of $u_\ell \in W_\ell$?

"Price": The dimension (= number of basis elements) of the space W_ℓ is

$$|W_\ell| = |I_\ell| = 2^{|\ell - \mathbb{1}|_1} = 2^{|\ell|_1 - d} = \prod_{i=1}^{d} 2^{(\ell_i - 1)} \tag{6.9}$$

"Value": How important is the contribution of $u_\ell \in W_\ell$ in the representation

$$u(x) = \sum_{\ell \in \mathbb{N}^d} u_\ell(x) \quad ?$$

Lemma 6.2.1 (Norm of the basis elements)

1. $\|\varphi_{\ell,i}\|_{L_\infty(\Omega)} = 1$

2. $\|\varphi_{\ell,i}\|_{L_1(\Omega)} = 2^{-|\ell|_1}$

3. $\|\varphi_{\ell,i}\|_E = \sqrt{2}\left(\frac{2}{3}\right)^{\frac{d-1}{2}} 2^{-\frac{|\ell|_1}{2}} \left(\sum_{j=1}^d 2^{2\ell_j}\right)^{\frac{1}{2}}$

Proof: Assertion 1. is clear, assertion 2. is straightforward, assertion 3. requires a long calculation; cf. [BG04].

Lemma 6.2.2 *Let*

$$\psi_{\ell_j,i_j}(x_j) := -2^{-(\ell_j+1)}\varphi_{\ell_j,i_j}(x_j) \qquad \text{and} \qquad \psi_{\ell,i}(x) := \prod_{j=1} \psi_{\ell_j,i_j}(x_j). \tag{6.10}$$

If $u \in X_0^2(\bar{\Omega})$, then the coefficients $v_{\ell,i}$ of the hierarchical multi-level representation (6.8) *are given by*

$$v_{\ell,i} = \int_\Omega \psi_{\ell,i}(x)\, \partial^{(2,\ldots,2)} u(x)\, dx.$$

Proof: Consider first the case $d = 1$, i.e. $\Omega = (0,1)$. Let $\ell \in \mathbb{N}$ and $i \in I_\ell$, and let $g(x) : [x_{\ell,i} - h_\ell, x_{\ell,i} + h_\ell] \longrightarrow \mathbb{R}$ be the linear interpolation between

$$\big(x_{\ell,i} - h_\ell, u(x_{\ell,i} - h_\ell)\big) \qquad \text{and} \qquad \big(x_{\ell,i} + h_\ell, u(x_{\ell,i} + h_\ell)\big).$$

By construction we have $u(x_{\ell,i}) = g(x_{\ell,i}) + v_{\ell,i}$ and hence

$$v_{\ell,i} = u(x_{\ell,i}) - \frac{u(x_{\ell,i} + h_\ell) + u(x_{\ell,i} - h_\ell)}{2}. \tag{6.11}$$

On the other hand, integration by parts yields

$$\begin{aligned}
\int_\Omega \psi_{\ell,i}(x)\, u''(x)\, dx &= \int_{x_{\ell,i}-h_\ell}^{x_{\ell,i}+h_\ell} \psi_{\ell,i}(x)\, u''(x)\, dx \\
&= \underbrace{\Big[\psi_{\ell,i}(x)u'(x)\Big]_{x=x_{\ell,i}-h_\ell}^{x_{\ell,i}+h_\ell}}_{=0} - \int_{x_{\ell,i}-h_\ell}^{x_{\ell,i}+h_\ell} \psi'_{\ell,i}(x)\, u'(x)\, dx \\
&= -\int_{x_{\ell,i}-h_\ell}^{x_{\ell,i}} \underbrace{\psi'_{\ell,i}(x)}_{=-\frac{1}{2}}\, u'(x)\, dx - \int_{x_{\ell,i}}^{x_{\ell,i}+h_\ell} \underbrace{\psi'_{\ell,i}(x)}_{=+\frac{1}{2}}\, u'(x)\, dx \\
&= \frac{1}{2}\Big[u(x)\Big]_{x=x_{\ell,i}-h_\ell}^{x_{\ell,i}} - \frac{1}{2}\Big[u(x)\Big]_{x=x_{\ell,i}}^{x_{\ell,i}+h_\ell}
\end{aligned}$$

which coincides with (6.11). The case $d > 1$ can be shown by using the tensor structure of the basis functions. ∎

Lemma 6.2.3 (Decay of the coefficients) *The coefficients $v_{\ell,i}$ in* (6.8) *are bounded by*

$$|v_{\ell,i}| \le 2^{-d} \cdot 2^{-2|\ell|_1} \cdot |u|_{2,\infty}.$$

Proof: Lemma 6.2.2 and Hölder's inequality yield

$$\begin{aligned} |v_{\ell,i}| &= \Big| \int_\Omega \psi_{\ell,i}(x)\ \partial^{(2,\dots,2)} u(x)\ dx \Big| \\ &= \|\partial^{(2,\dots,2)} u\|_{L_\infty} \cdot \|\psi_{\ell,i}\|_{L_1} \\ &\leq |u|_{(2,\dots,2),\infty} \cdot 2^{-d} 2^{-|\ell|_1} \|\varphi_{\ell,i}\|_{L_1} \end{aligned}$$

due to definition (6.10). Since

$$\|\varphi_{\ell,i}\|_{L_1} = 2^{-|\ell|_1}$$

according to Lemma 6.2.1 (part 2.), the assertion follows. ∎

Theorem 6.2.4 (Bounds for u_ℓ)

$$\|u_\ell\|_{L_\infty(\Omega)} \leq 2^{-d}\ 2^{-2|\ell|_1}\ |u|_{2,\infty} \tag{6.12}$$

$$\|u_\ell\|_E \leq \frac{1}{2 \cdot 12^{(d-1)/2}} 2^{-2|\ell|_1} \left(\sum_{j=1}^{d} 2^{2\ell_j} \right)^{\frac{1}{2}} |u|_{2,\infty} \tag{6.13}$$

Proof. Let $\hat{x} \in \bar{\Omega}$ such that

$$\|u_\ell\|_{L_\infty(\Omega)} = |u_\ell(\hat{x})| = |\sum_{i \in I_\ell} v_{\ell,i} \varphi_{\ell,i}(\hat{x})|.$$

The support of the basis functions $\varphi_{\ell,i}$ is disjoint. Hence, there is a $j \in I_\ell$ such that $\varphi_{\ell,i}(\hat{x}) = 0$ for all $i \neq j$. Thus,

$$\|u_\ell\|_{L_\infty(\Omega)} = |v_{\ell,j}| \cdot \underbrace{|\varphi_{\ell,j}(\hat{x})|}_{\leq 1 \text{ (Lemma 6.2.1)}} \overset{\text{Lemma 6.2.3}}{\leq} 2^{-d}\ 2^{-2|\ell|_1}\ |u|_{2,\infty}$$

which proves (6.12). (6.13) follows from

$$\begin{aligned} \|u_\ell\|_E^2 &\leq \sum_{i \in I_\ell} \underbrace{|v_{\ell,i}|^2}_{\text{Lemma 6.2.3}} \underbrace{\|\varphi_{\ell,i}\|_E^2}_{\text{Lemma 6.2.1}} \\ &\leq \sum_{i \in I_\ell} 2^{-2d} \cdot 2^{-4|\ell|_1} \cdot |u|_{2,\infty}^2 \cdot 2 \left(\frac{2}{3} \right)^{d-1} 2^{-|\ell|_1} \sum_{j=1}^{d} 2^{2l_j} \end{aligned}$$

and the fact that $\sum_{i \in I_\ell} 1 = |W_\ell| = 2^{|\ell|_1} 2^{-d}$. ∎

Remark: The theorem states that

$$\|u_\ell\|_{L_\infty(\Omega)} \longrightarrow 0, \qquad \|u_\ell\|_E \longrightarrow 0$$

when $|\ell|_1 \longrightarrow \infty$. Hence, the multi-level expansion can be truncated if a small approximation error is acceptable.

6.3 Approximation on uniform and sparse grids

Goal: Approximate $u = \sum_{\ell \in \mathbb{N}^d} u_\ell$ in a finite-dimensional subspace.

First idea: Define the approximation

$$u_n^{(\infty)}(x) = \sum_{\substack{l \in \mathbb{N}^d \\ |\ell|_\infty \le n}} u_\ell(x) \in \bigoplus_{|\ell|_\infty \le n} W_\ell =: V_n^{(\infty)} \tag{6.14}$$

$V_n^{(\infty)}$ is the subspace of all piecewise d-linear functions on a **uniform** grid with equidistant mesh-width $h_n = 2^{-n}$ in each spatial dimension.

$$\text{Dimension of } V_n^{(\infty)} \quad = \quad \text{number of inner points} \quad = \quad (2^n - 1)^d$$

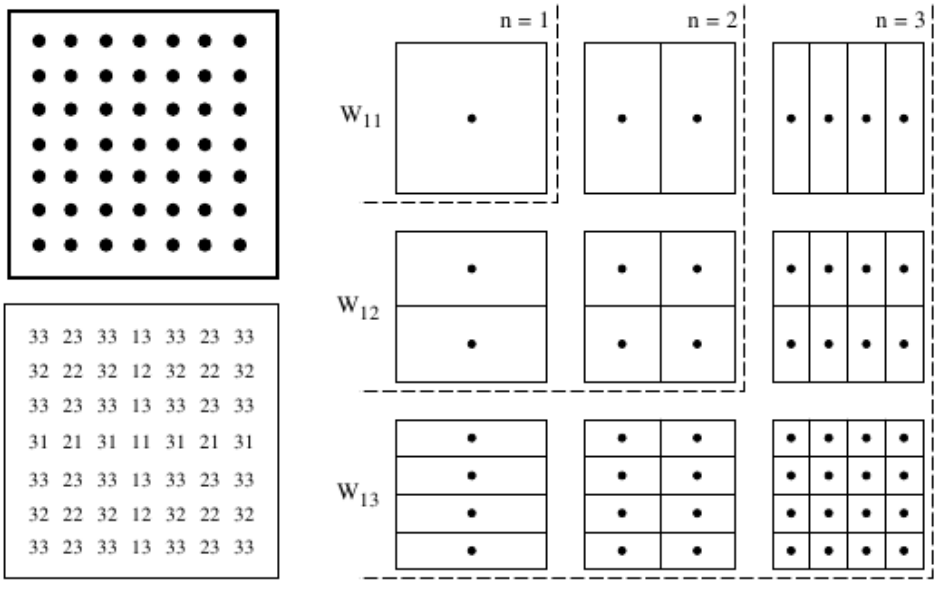


Figure 6.6: Illustration of the space $V_n^{(\infty)}$ for $d = 2$ and $n = 3$. Figure stolen from [BG04].

Theorem 6.3.1 (Accuracy of $u_n^{(\infty)}$) *Let $u(x) = \sum_{\ell \in \mathbb{N}^d} u_\ell(x) \in X_0^2(\bar{\Omega})$ with $u_\ell \in W_\ell$. There are constants $C_1(d)$ and $C_2(d)$ such that*

$$(a) \qquad \|u - u_n^{(\infty)}\|_{L_\infty(\Omega)} \leq C_1(d)\, 2^{-2n}\, |u|_{2,\infty} = \mathcal{O}\big(h_n^2\big) \qquad (2^{-n} = h_n) \tag{6.15}$$

$$(b) \qquad \|u - u_n^{(\infty)}\|_E \leq C_2(d)\, 2^{-n}\, |u|_{2,\infty} = \mathcal{O}(h_n) \tag{6.16}$$

Proof: [BG04].

Important consequence: The approximation $u_n^{(\infty)}$ is not suited for high-dimensional problems!

$$\|u - u_n^{(\infty)}\|_{L_\infty(\Omega)} \;\leq\; C h_n^2 \;=\; C \frac{1}{(2^n)^2} \quad \Longleftrightarrow \quad (2^n - 1)^d \text{ basis functions/coefficients} \tag{6.17}$$

$\Longrightarrow$ Curse of dimensionality!

Better approach? Observations:

$$\begin{aligned} \|u_\ell\|_{L_\infty(\Omega)} &\leq C \cdot 2^{-2\,|\ell|_1} \qquad \text{(Theorem 6.2.4)} \\ |W_\ell| = 2^{|\ell - \mathbb{1}|_1} &= 2^{|\ell|_1 - d} \end{aligned}$$

Hence, both $\|u_\ell\|_{L_\infty(\Omega)}$ and $|W_\ell|$ depend on $|\ell|_1$ and not on $|\ell|_\infty$!

Second idea: Define approximation in terms of $|\ell|_1$.

$$u_n^{(1)}(x) = \sum_{\substack{\ell \in \mathbb{N}^d \\ |\ell|_1 \leq n+d-1}} u_\ell(x) \in \bigoplus_{\substack{\ell \in \mathbb{N}^d \\ |\ell|_1 \leq n+d-1}} W_\ell \;=:\; V_n^{(1)}$$

The corresponding grids are called **Sparse Grids**.

How "costly" is the approximation on sparse grids?

Lemma 6.3.2 (Dimension of $V_n^{(1)}$) *The dimension (= number of basis elements) of the space $V_n^{(1)}$ is*

$$|V_n^{(1)}| = \sum_{i=0}^{n-1} 2^i \binom{d-1+i}{d-1} = 2^n \left(\frac{n^{d-1}}{(d-1)!} + \mathcal{O}\left(n^{d-2}\right) \right) \tag{6.18}$$

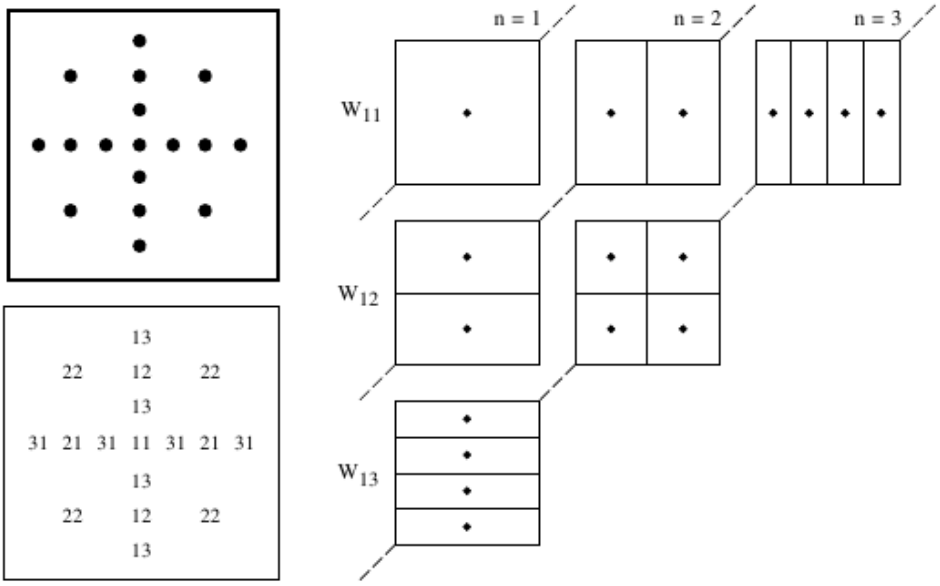


Figure 6.7: Illustration of the space $V_n^{(1)}$ for $d = 2$ and $n = 3$. Figure stolen from [BG04].

Remark: $|V_n^{(1)}| = \mathcal{O}\big(2^n \cdot n^{d-1}\big)$ *is much better than* $|V_n^{(\infty)}| = (2^n - 1)^d$.

The following identities will be helpful to prove Lemma 6.3.2: For all $N, n \in \mathbb{N}$ and smooth functions f,g, we have

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} \qquad \text{(Binomial theorem)} \tag{6.19a}$$

$$\frac{d^n}{dx^n}(fg) = \sum_{k=0}^{n} \binom{n}{k} \frac{d^k f}{dx^k} \cdot \frac{d^{n-k} g}{dx^{n-k}} \qquad \text{(Leibnitz rule)} \tag{6.19b}$$

$$\frac{d^n}{dx^n} x^N = \frac{N!}{(N-n)!} x^{N-n}, \qquad N \geq n \in \mathbb{N} \tag{6.19c}$$

$$\frac{d^n}{dx^n}(1-x)^{-1} = n! \cdot (1-x)^{-(n+1)} \tag{6.19d}$$

Proof of Lemma 6.3.2. First equality: Since $|W_\ell| = 2^{i-d}$ we obtain

$$\begin{aligned}
|V_n^{(1)}| &= \sum_{\substack{\ell\in\mathbb{N}^d \\ d\le|\ell|_1\le n+d-1}} |W_\ell| \\
&= \sum_{i=d}^{n+d-1} 2^{i-d} \underbrace{\sum_{|\ell|_1=i} 1}_{=\text{ number of spaces } W_\ell \text{ with } |\ell|_1=i} && (|\ell|_1 = i) \\
&= \sum_{i=d}^{n+d-1} 2^{i-d} \binom{i-1}{d-1} && (\text{induction over } d) \\
&= \sum_{i=0}^{n-1} 2^i \binom{d-1+i}{d-1}
\end{aligned}$$

Second equality:

$$\sum_{i=0}^{n-1} 2^i \binom{d-1+i}{d-1} \overset{(6.19c)}{=} \frac{1}{(d-1)!} \sum_{i=0}^{n-1} \frac{d^{d-1}}{dx^{d-1}} x^{i+d-1}\bigg|_{x=2} = \frac{1}{(d-1)!} \frac{d^{d-1}}{dx^{d-1}} \Big(x^{d-1} \sum_{i=0}^{n-1} x^i \Big) \tag{6.20}$$

Use that

$$\begin{aligned}
&\frac{d^{d-1}}{dx^{d-1}} \Big(x^{d-1} \sum_{i=0}^{n-1} x^i \Big) = \frac{d^{d-1}}{dx^{d-1}} \Big(x^{d-1} \frac{1-x^n}{1-x} \Big) && (\text{geom. sum}) \\
&= \frac{d^{d-1}}{dx^{d-1}} \Big((x^{d-1} - x^{d-1+n}) \cdot (1-x)^{-1} \Big) \\
&= \sum_{i=0}^{d-1} \binom{d-1}{i} \frac{d^i}{dx^i} (x^{d-1} - x^{d-1+n}) \cdot \frac{d^{d-1-i}}{dx^{d-1-i}} (1-x)^{-1} && (\text{by (6.19b)}) \\
&= \sum_{i=0}^{d-1} \binom{d-1}{i} \Big(\frac{(d-1)!}{(d-1-i)!} x^{d-1-i} - \frac{(d-1+n)!}{(d-1+n-i)!} x^{d-1+n-i} \Big) \cdot && (\text{by (6.19c), (6.19d)}) \\
&\qquad\qquad \cdot (d-1-i)! (1-x)^{-(d-1-i+1)}
\end{aligned}$$

Substituting into (6.20) and setting $x = 2$ yields

$$\begin{aligned}
&\sum_{i=0}^{d-1} \binom{d-1}{i} 2^{d-1-i} \cdot (-1)^{-(d-i)} - \sum_{i=0}^{d-1} \frac{1}{i!} \frac{(d-1+n)!}{(d-1+n-i)!} 2^{d-1+n-i} \cdot (-1)^{-(d-i)} \\
&= \underbrace{(-1)^{-d} (2-1)^{d-1}}_{=(-1)^d} - 2^n \sum_{i=0}^{d-1} \binom{n+d-1}{i} (-2)^{d-1-i} (-1) && (\text{by (6.19a)}).
\end{aligned}$$

The assertion about the leading order term follows from the fact that

$$\begin{aligned}\binom{n+d-1}{i} &= \frac{1}{i!}\underbrace{(n+d-1)(n+d-2)\dots(n+d-i)}_{i \text{ terms}} \\ &= \frac{1}{i!}n^i + \mathcal{O}\big(n^{i-1}\big)\end{aligned}$$

for $i = 0, \dots, d-1$. ∎

How accurate is the approximation on sparse grids?

Theorem 6.3.3 (Accuracy of $u_n^{(1)}$) *Let $u(x) = \sum_{\ell \in \mathbb{N}^d} u_\ell(x) \in X_0^2(\bar{\Omega})$ with $u_\ell \in W_\ell$. There are constants $C_1(d)$ and $C_2(d)$ such that*

$$\begin{aligned} &(a) \qquad \|u - u_n^{(1)}\|_{L_\infty(\Omega)} \le C_1(d) 2^{-2n} A(d,n) |u|_{2,\infty} \qquad = \mathcal{O}\big(h_n^2 \cdot n^{d-1}\big) \\ &(b) \qquad \|u - u_n^{(1)}\|_E \le C_2(d) 2^{-n} |u|_{2,\infty} \qquad = \mathcal{O}(h_n) \end{aligned}$$

where

$$A(d,n) = \sum_{k=0}^{d-1} \binom{n+d-1}{k} = \frac{n^{d-1}}{(d-1)!} + \mathcal{O}\left(n^{d-2}\right).$$

Remark: Comparing Theorem 6.3.3 with Theorem 6.3.1 shows:

(a) In $\|\cdot\|_{L_\infty(\Omega)}$ the approximation error on sparse grids is slightly larger than the approximation error on full grids due to the term $A(d,n)$. The advantage is that the approximation $u_n^{(1)}$ is obtained with much less basis elements/coefficients.

(b) In the energy norm, the order of the sparse grid approximation is the same as on full grids.

Proof. The bound (6.12) from Theorem 6.2.4 yields

$$\|u - u_n^{(1)}\|_{L_\infty(\Omega)} \le \sum_{|\ell|_1 \ge n+d} \|u_\ell\|_{L_\infty(\Omega)} \le \sum_{|\ell|_1 \ge n+d} 2^{-d} \cdot 2^{-2|\ell|_1} \cdot |u|_{2,\infty}$$

In order to prove (a), we have to show that

$$\sum_{|\ell|_1 \ge n+d} 2^{-2|\ell|_1} \le \tilde{C}_3(d)\, 2^{-2n} A(d,n). \tag{6.21}$$

First, we note that

$$\sum_{|\ell|_1 \geq n+d} 2^{-2|\ell|_1} = \sum_{i=n+d}^{\infty} 2^{-2i} \sum_{|\ell|_1 = i} 1 = \sum_{i=n+d}^{\infty} 2^{-2i} \binom{i-1}{d-1}$$
$$= 2^{-2(n+d)} \sum_{i=0}^{\infty} 2^{-2i} \binom{n+i+d-1}{d-1}$$

For every $x \in \mathbb{R}$ with $|x| < 1$ we have

$$\sum_{i=0}^{\infty} x^i \binom{n+i+d-1}{d-1} = \frac{x^{-n}}{(d-1)!} \frac{d^{d-1}}{dx^{d-1}} \sum_{i=0}^{\infty} x^{n+i+d-1} \qquad \text{(by (6.19c))}$$
$$= \frac{x^{-n}}{(d-1)!} \frac{d^{d-1}}{dx^{d-1}} \left(x^{n+d-1} \cdot \frac{1}{1-x} \right) \qquad \text{(geom. series)}$$
$$= \frac{x^{-n}}{(d-1)!} \sum_{k=0}^{d-1} \binom{d-1}{k} \frac{d^k}{dx^k} x^{n+d-1} \cdot \frac{d^{d-1-k}}{dx^{d-1-k}} \frac{1}{1-x} \qquad \text{(by (6.19b))}$$
$$= \frac{x^{-n}}{(d-1)!} \sum_{k=0}^{d-1} \binom{d-1}{k} \frac{(n+d-1)!}{(n+d-1-k)!} x^{n+d-1-k} \cdot \qquad \text{(by (6.19c),(6.19d))}$$
$$\cdot (d-1-k)!(1-x)^{-(d-k)}$$
$$= \sum_{k=0}^{d-1} \binom{n+d-1}{k} \left(\frac{x}{1-x} \right)^{d-1-k} (1-x)^{-1}.$$

With $x = 2^{-2}$ it follows that

$$\sum_{|\ell|_1 \geq n+d} 2^{-2|\ell|_1} = 2^{-2(n+d)} \sum_{k=0}^{d-1} \binom{n+d-1}{k} \underbrace{\left(\frac{1}{3}\right)^{d-1-k} \frac{4}{3}}_{\leq 4/3}$$
$$= \frac{4}{3} 2^{-2d} \cdot 2^{-2n} A(d,n)$$

which proves (6.21) and hence assertion (a). In order to prove (b), we apply (6.13) from Theorem 6.2.4 and obtain

$$\|u - u_n^{(1)}\|_E \leq \sum_{|\ell|_1 \geq n+d} \|u_\ell\|_E$$
$$\leq \underbrace{\frac{|u|_{2,\infty}}{2 \cdot 12^{(d-1)/2}}}_{=:C} \sum_{|\ell|_1 \geq n+d} 2^{-2\,|\ell|_1} \left(\sum_{j=1}^{d} 2^{2\ell_j} \right)^{\frac{1}{2}}$$
$$= C \sum_{i=n+d}^{\infty} 2^{-2i} \sum_{|\ell|_1 = i} \left(\sum_{j=1}^{d} 4^{\ell_j} \right)^{\frac{1}{2}}$$

By induction with respect to d it can be shown that

$$\sum_{|\ell|_1=i} \left(\sum_{j=1}^{d} 4^{\ell_j} \right)^{\frac{1}{2}} \leq d \cdot 2^i.$$

This yields

$$\begin{aligned} \|u - u_n^{(1)}\|_E &\leq C \sum_{i=n+d}^{\infty} 2^{-2i} \cdot d \cdot 2^i \\ &= C \cdot d \left(\sum_{i=0}^{\infty} 2^{-i} - \sum_{i=0}^{n+d-1} 2^{-i} \right) \\ &= C \cdot d \left(\frac{1}{1-\frac{1}{2}} - \frac{1-2^{-(n+d)}}{1-\frac{1}{2}} \right) \qquad \text{(geometric sum/series)} \\ &= C \cdot d \cdot 2^{-(n+d)} \cdot 2 \end{aligned}$$

and the assertion (b) follows. ∎

Remark. Similar error bounds can be shown for $\|\cdot\|_{L_2}$; cf. [BG04, Theorem 3.8].

Extensions and modifications

- Other boundary conditions
- Other basis functions: Wavelets, interpolets, ...
- Refined truncation strategies. It can be shown, however, that the cost-benefit ratio of $V_n^{(1)}$ is optimal if the error is measured in $\|\cdot\|_{L_\infty}$ or $\|\cdot\|_{L_2}$; cf. [BG04, p. 26].
- Adaptive sparse grids
- ...

6.4 Differential operators on sparse grids

In order to solve PDEs on sparse grids, we need suitable approximations of differential operators such as, e.g., $\Delta = \sum_{i=1}^{d} \partial_{x_i}^2$.

Finite differences. The problem is finite differences in the nodes of the sparse grid are not consistent. If we want to approximate, e.g., $\partial_{x_i}^2$ by the central difference quotient with respect to the $i-$th direction, then for some nodes of the sparse grid the step-size must be chosen to be $1/2$ no matter how large n and how small $h_n = 2^{-n}$ is; cf. Figure 6.7. The solution is to change from the hierarchical to the nodal representation (in direction i only), apply the difference quotient, and change back.

Galerkin approximation. This works exactly as for standard finite elements: Derive weak formulation, define bilinear form $a(\cdot,\cdot)$, restrict the problem to the finite-dimensional approximation space $V_n^{(1)}$. Error bounds can be obtained by combining Céa's Lemma (Lemma 4.8.1) with the bound for the interpolation error (Theorem 6.3.3) as in Section 4.8(c).

— FIN —

Bibliography

[BG04] Hans-Joachim Bungartz and Michael Griebel. Sparse grids. *Acta Numerica*, 13:147–269, 2004.

[Bra07] Dietrich Braess. *Finite Elemente. Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie.* Springer, Berlin, 4nd ed. edition, 2007.

[BS08] Susanne C. Brenner and L.Ridgway Scott. *The mathematical theory of finite element methods.* Texts in Applied Mathematics 15. Springer, New York, 3rd edition, 2008.

[CT04] Rama Cont and Peter Tankov. *Financial modelling with jump processes.* Chapman & Hall/CRC Financial Mathematics Series. Chapman & Hall, Boca Raton, 2004.

[CV05] Rama Cont and Ekaterina Voltchkova. A finite difference scheme for option pricing in jump diffusion and exponential Lévy models. *SIAM J. Numer. Anal.*, 43(4):1596–1626, 2005.

[GHM09] Michael B. Giles, Desmond J. Higham, and Xuerong Mao. Analyzing multi-level Monte Carlo for options with non-globally Lipschitz payoff. *Finance Stoch.*, 13(3):403–413, 2009.

[Gil08a] Michael B. Giles. *Improved multilevel Monte Carlo convergence using the Milstein scheme*, pages 343–358. In: Monte Carlo and quasi-Monte Carlo methods 2006. Selected papers based on the presentations at the 7th international conference 'Monte Carlo and quasi-Monte Carlo methods in scientific computing', Ulm, Germany, August 14–18, 2006. Springer, Berlin, 2008.

[Gil08b] Michael B. Giles. Multilevel Monte Carlo path simulation. *Operations Research*, 56(3):607–617, 2008.

[Goc06] Mark S. Gockenbach. *Understanding and implementing the finite element method.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2006.

[HB09] Martin Hanke-Bourgeois. *Grundlagen der numerischen Mathematik und des wissenschaftlichen Rechnens.* Vieweg+Teubner, Wiesbaden, 3rd revised ed. edition, 2009.

[HV03] Willem Hundsdorfer and Jan Verwer. *Numerical solution of time-dependent advection-diffusion-reaction equations.* Number 33 in Springer Series in Computational Mathematics. Springer, Berlin, 2003.

[KKK10] Ralf Korn, Elke Korn, and Gerald Kroisandt. *Monte Carlo methods and models in finance and insurance.* CRC Financial Mathematics Series. Chapman & Hall, Boca Raton, FL, 2010.

[LT09] Stig Larsson and Vidar Thomée. *Partial differential equations with numerical methods (paperback reprint).* Number 45 in Texts in Applied Mathematics. Springer, Berlin, 2009.