Computational Methods in Optimal Control Lecture 4. Convergence Analysis

William W. Hager

July 24, 2018

Heat Equation

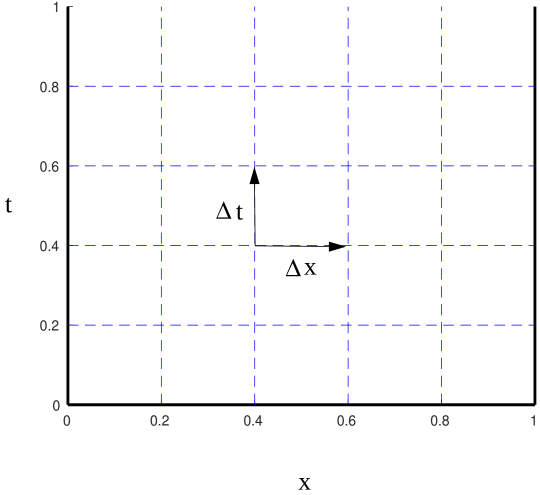
Let us consider the heat equation for a rod with unit conductivity:

$$\frac{\partial u(t,x)}{\partial t} = \frac{\partial^2 u(t,x)}{\partial x^2}$$

$$u(0,x) = u_0(x), \quad 0 \leq x \leq 1$$

$$u(t,0) = u(t,1) = 0, \quad t \geq 0$$

Mesh



Finite Difference Approximation

$$
\begin{aligned}
t_j &= j\Delta t \\
x_k &= k\Delta x, \quad \Delta x = 1/N, \quad 0 \le k \le N \\
u_{jk} &\approx u(t_j, x_k) \\
\frac{\partial u^*(t_j, x_k)}{\partial t} &= \frac{u^*_{j+1,k} - u^*_{jk}}{\Delta t} + O(\Delta t) \\
\frac{\partial^2 u(t_j, x_k)}{\partial x^2} &= \frac{u^*_{j,k+1} - 2u^*_{jk} + u^*_{j,k-1}}{(\Delta x)^2} + O((\Delta x)^2)
\end{aligned}
$$

$$
\begin{aligned}
\frac{u_{j+1,k} - u_{jk}}{\Delta t} &= \frac{u_{j,k+1} - 2u_{jk} + u_{j,k-1}}{(\Delta x)^2} \\
u_{j+1,k} &= u_{jk} + \rho(u_{j,k+1} - 2u_{jk} + u_{j,k-1}), \quad \rho = \Delta t/(\Delta x)^2 \\
u_{0k} &= u_0(x_k)
\end{aligned}
$$

Error Analysis

Rewrite the finite difference system using matrix and vector notation. Let $\mathbf{u}_j = [u_{j0}, u_{j1}, \ldots, u_{jN}]^{\mathsf{T}}$ and let $\mathbf{A}$ denote the $(N-1)$ by $(N-1)$ coefficient matrix for the finite difference system:

$$\mathbf{A} = \mathbf{I} + \rho \begin{pmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{pmatrix}$$

$$\begin{aligned}
\mathbf{u}_{j+1} &= \mathbf{A}\mathbf{u}_j \\
\mathbf{u}^*_{j+1} &= \mathbf{A}\mathbf{u}^*_j + \Delta t\mathbf{r}_j, \quad r_{jk} = O(\Delta t) + O((\Delta x)^2) \\
\mathbf{e}_{j+1} &= \mathbf{A}\mathbf{e}_j + (\Delta t)\mathbf{r}_j, \quad \mathbf{e}_j = \mathbf{u}^*_j - \mathbf{u}_j \\
\mathbf{e}_{j+1} &= \mathbf{r}_j + \mathbf{A}(\mathbf{A}\mathbf{e}_{j-1} + (\Delta t)\mathbf{r}_{j-1}) \\
\mathbf{e}_{j+1} &= (\Delta t)(\mathbf{r}_j + \mathbf{A}\mathbf{r}_{j-1}) + \mathbf{A}^2\mathbf{e}_{j-1}
\end{aligned}$$

Error Bound

Since $\mathbf{e}_0 = 0$, we have

$$\mathbf{e}_j = \Delta t \sum_{i=0}^{j-1} \mathbf{A}^i \mathbf{r}_{j-i-1}.$$

Let us measure the error using a norm similar to the $\mathcal{L}^2$ norm:

$$\|\mathbf{e}\|_2^2 = \sum_{k=1}^{N-1} (\Delta x) e_k^2 \Longleftrightarrow \|\mathbf{e}\|_2 = \sqrt{\Delta x} \langle \mathbf{e}, \mathbf{e} \rangle^{1/2}$$

$$\|\mathbf{e}_j\|_2 \leq \Delta t \sum_{i=0}^{j-1} \|\mathbf{A}\|_2^i \|\mathbf{r}_{j-i-1}\|_2.$$

If $\|\mathbf{A}\| \leq 1$, then

$$\|\mathbf{e}_j\|_2 \leq \Delta t \sum_{i=0}^{j-1} \|\mathbf{r}_i\|_2 \leq j(\Delta t)(O(\Delta t) + O((\Delta x)^2).$$

If $\|\mathbf{A}\|_2 \leq 1$, then $\mathbf{A}$ is stable and if $\|\mathbf{r}_j\|_2$ tends to zero as $\Delta t \to 0$ and $\Delta x \to 0$, then the scheme is consistent. Hence, a stable, consistent scheme is convergent in the sense that $\mathbf{e}_j \to 0$ as $\Delta t \to 0$, $\Delta x \to 0$, and $j\Delta t \to t$.

In our example, $\|\mathbf{A}\| \leq 1$ for certain choices of Δt and Δx. Since $\mathbf{A}$ is a tridiagonal Toeplitz matrix, there is an explicit formula for its eigenvalues:

$$1 - 4\rho \sin^2\left(\frac{k\pi}{2N}\right), \quad 1 \leq k \leq N - 1.$$

Since the eigenvalues are ≤ 1 for any k, we only need to worry about the smallest eigenvalue. It is greater than -1 when

$$1 - 4\rho \sin^2\left(\frac{k\pi}{2N}\right) \geq -1 \implies \rho \sin^2\left(\frac{k\pi}{2N}\right) \leq \frac{1}{2}.$$

Since the sin term is at most one, this inequality holds when $\rho \leq 1/2$, or equivalently, when $\Delta t \leq (\Delta x)^2/2$.

Lax Equivalence Theorem

This example illustrates what is known as the Lax Equivalence Theorem. Usually, it is presented in the context of a semigroup of operators, the matrix $\mathbf{A}$ corresponding to the discretization is replaced by a linear operator $\mathbf{A}(\Delta t)$, and stability is defined to mean that the operators $\mathbf{A}(\Delta t)^j$ are uniformly bounded for all Δt and j such that $0 \leq j(\Delta t) \leq T$. In this more general setting, consistency and stability are equivalent to convergence.

Limitations to this theory are the following:

- The theory is developed in the context of linear variational problems.
- The theory does not take into account variational problems with constraints, or equivalently, the theory is developed for an equation, not an inclusion.

To treat nonlinear variational problems with constraints, we need to address both of these limitations.

Abstract Convergence Theorem

Theorem. Let $\mathcal{X}$ be a Banach space and let $\mathcal{Y}$ be a linear normed space with the norms in both spaces denoted $\|\cdot\|$. Let $\mathcal{F} : \mathcal{X} \mapsto 2^{\mathcal{Y}}$, let $\mathcal{L} : \mathcal{X} \mapsto \mathcal{Y}$ be a bounded linear operator, and let $\mathcal{T} : \mathcal{X} \mapsto \mathcal{Y}$ with $\mathcal{T}$ continuously Fréchet differentiable in $B_r(\boldsymbol{\theta}^*)$ for some $\boldsymbol{\theta}^* \in \mathcal{X}$ and $r > 0$. Suppose that the following conditions hold for some $\boldsymbol{\delta} \in \mathcal{Y}$ and scalars ϵ and $\gamma > 0$:

(C1) $\mathcal{T}(\boldsymbol{\theta}^*) + \boldsymbol{\delta} \in \mathcal{F}(\boldsymbol{\theta}^*)$.

(C2) $\|\nabla \mathcal{T}(\boldsymbol{\theta}) - \mathcal{L}\| \leq \epsilon$ for all $\boldsymbol{\theta} \in B_r(\boldsymbol{\theta}^*)$.

(C3) The map $(\mathcal{F} - \mathcal{L})^{-1}$ is single-valued and Lipschitz continuous with Lipschitz constant γ.

If $\epsilon\gamma < 1$ and $\|\boldsymbol{\delta}\| \leq (1 - \gamma\epsilon)r/\gamma$, then there exists a unique $\boldsymbol{\theta} \in B_r(\boldsymbol{\theta}^*)$ such that $\mathcal{T}(\boldsymbol{\theta}) \in \mathcal{F}(\boldsymbol{\theta})$. Moreover, we have the estimate

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \leq \frac{\gamma}{1 - \gamma\epsilon}\|\boldsymbol{\delta}\|. \qquad \text{(Error Bound)}$$

Inverse of a Set-Valued Map and the Interpretation of (C3)

If $\mathcal{S} : \mathcal{X} \mapsto 2^{\mathcal{Y}}$, then for any $\mathbf{y} \in \mathcal{Y}$,

$$\mathcal{S}^{-1}(\mathbf{y}) = \{\mathbf{x} \in \mathcal{X} : \mathbf{y} \in \mathcal{S}(\mathbf{x}).$$

Let us consider the optimization problem

$$\min \quad f(\mathbf{x}) \text{ subject to } \mathbf{x} \in \mathcal{U},$$

where $\mathcal{U} \subset \mathbb{R}^n$ is a closed, convex set and f is twice continuously differentiable. Suppose $\mathbf{x}^*$ is a local minimizer and $\mathbf{A} := \nabla^2 f(\mathbf{x}^*)$ is positive definite. The first-order optimality condition for $\mathbf{x}^*$ is that $-\nabla f(\mathbf{x}^*) \in N_{\mathcal{U}}(\mathbf{x}^*)$. If $\mathcal{F}(\mathbf{x}) = -N_{\mathcal{U}}(\mathbf{x})$, then the first-order condition is $\nabla f(\mathbf{x}^*) \in \mathcal{F}(\mathbf{x}^*)$. Focusing on (C3), observe that

$$\begin{aligned}
\mathbf{x} \in (\mathcal{F} - \mathbf{A})^{-1}\,\mathbf{y} \quad &\Leftrightarrow \quad \mathbf{y} \in \mathcal{F}(\mathbf{x}) - \mathbf{A}\mathbf{x} \\
&\Leftrightarrow \quad \mathbf{A}\mathbf{x} + \mathbf{y} \in \mathcal{F}(\mathbf{x}) \\
&\Leftrightarrow \quad \langle \mathbf{A}\mathbf{x} + \mathbf{y}, \mathbf{z} - \mathbf{x} \rangle \geq 0 \text{ for all } \mathbf{z} \in \mathcal{U} \\
&\Leftrightarrow \quad \mathbf{x} = \arg\min \{\frac{1}{2}\mathbf{z}^{\mathsf{T}}\mathbf{A}\mathbf{z} + \mathbf{y}^{\mathsf{T}}\mathbf{z} : \mathbf{z} \in \mathcal{U}\}
\end{aligned}$$

Equivalence Between Variational Inequality and Minimizer

Earlier we showed that if $\mathcal{U}$ is convex, $F : \mathcal{U} \mapsto \mathbb{R}$ is differentiable, and

$$\mathbf{x}^* = \arg\min\{F(\mathbf{x}) : \mathbf{x} \in \mathcal{U}\}, \tag{V1}$$

then

$$\nabla F(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}^*) \geq 0 \quad \text{for all } \mathbf{x} \in \mathcal{U}. \tag{V2}$$

If F is convex, then (V1) and (V2) are equivalent since a convex differentiable function has the property that

$$F(\mathbf{y}) \geq F(\mathbf{x}) + \nabla F(\mathbf{x})(\mathbf{y} - \mathbf{x})$$

for all $\mathbf{x}$ and $\mathbf{y} \in \mathcal{U}$. In particular,

$$F(\mathbf{x}) \geq F(\mathbf{x}^*) + \nabla F(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}^*).$$

Hence, if (V2) holds, then $F(\mathbf{x}) \geq F(\mathbf{x}^*)$ for all $\mathbf{x} \in \mathcal{U}$.

Proof of the Theorem

Define

$$\Phi(\boldsymbol{\theta}) = (\mathcal{F} - \mathcal{L})^{-1}(\mathcal{T}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}))$$

For all $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2 \in B_r(\boldsymbol{\theta}^*)$, a Taylor expansion with integral remainder term yields

$$(\mathcal{T}-\mathcal{L})(\boldsymbol{\theta}_2) = (\mathcal{T}-\mathcal{L})(\boldsymbol{\theta}_1)+\int_0^1 [\nabla\mathcal{T}(\boldsymbol{\theta}_1+s(\boldsymbol{\theta}_2-\boldsymbol{\theta}_1))-\mathcal{L}]\, ds\, (\boldsymbol{\theta}_2-\boldsymbol{\theta}_1).$$

By (C2), it follows that

$$\|(\mathcal{T} - \mathcal{L})(\boldsymbol{\theta}_2) - (\mathcal{T} - \mathcal{L})(\boldsymbol{\theta}_1)\| \leq \epsilon\|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\|. \qquad \text{(L1)}$$

By (C3) and (L1), we have

$$\begin{aligned} & \|\Phi(\boldsymbol{\theta}_1) - \Phi(\boldsymbol{\theta}_2)\| \\ = \; & \|(\mathcal{F} - \mathcal{L})^{-1}(\mathcal{T} - \mathcal{L})(\boldsymbol{\theta}_1) - (\mathcal{F} - \mathcal{L})^{-1}(\mathcal{T} - \mathcal{L})(\boldsymbol{\theta}_2)\| \\ \leq \; & \gamma\|(\mathcal{T} - \mathcal{L})(\boldsymbol{\theta}_1) - (\mathcal{T} - \mathcal{L})(\boldsymbol{\theta}_2)\| \\ \leq \; & \epsilon\gamma\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| \end{aligned}$$

Since $\epsilon\gamma < 1$, Φ is a contraction on $B_r(\boldsymbol{\theta}^*)$.

Proof Continued . . .

Subtracting $\mathcal{L}(\boldsymbol{\theta}^*)$ from each side of (C1) and utilizing the uniqueness in (C3) gives

$$(\mathcal{T}-\mathcal{L})\boldsymbol{\theta}^*+\boldsymbol{\delta}\in(\mathcal{F}-\mathcal{L})(\boldsymbol{\theta}^*)\Longrightarrow\boldsymbol{\theta}^*=(\mathcal{F}-\mathcal{L})^{-1}[(\mathcal{T}-\mathcal{L})\boldsymbol{\theta}^*+\boldsymbol{\delta}].$$

With this substitution, it follows from (L1), (C3) and (C2) that

$$\begin{aligned}
& \|\Phi(\boldsymbol{\theta})-\boldsymbol{\theta}^*\| \\
= \quad & \|(\mathcal{F}-\mathcal{L})^{-1}(\mathcal{T}-\mathcal{L})(\boldsymbol{\theta})-(\mathcal{F}-\mathcal{L})^{-1}[(\mathcal{T}-\mathcal{L})(\boldsymbol{\theta}^*)+\boldsymbol{\delta}]\| \\
\leq \quad & \gamma\|(\mathcal{T}-\mathcal{L})(\boldsymbol{\theta})-[(\mathcal{T}-\mathcal{L})(\boldsymbol{\theta}^*)+\boldsymbol{\delta}]\| \\
\leq \quad & \gamma(\epsilon\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|+\|\boldsymbol{\delta}\|)\leq\gamma(\epsilon r+\|\boldsymbol{\delta}\|)
\end{aligned}$$

for all $\boldsymbol{\theta}\in B_r(\boldsymbol{\theta}^*)$. The assumption that $\|\boldsymbol{\delta}\|\leq(1-\gamma\epsilon)r/\gamma$ can be rearranged to obtain $\gamma(\epsilon r+\|\boldsymbol{\delta}\|)\leq r$, which implies that $\|\Phi(\boldsymbol{\theta})-\boldsymbol{\theta}^*\|\leq r$. Since Φ maps $B_r(\boldsymbol{\theta}^*)$ into itself and Φ is a contraction on $B_r(\boldsymbol{\theta}^*)$, the contraction mapping principle yields the existence of a unique fixed point $\boldsymbol{\theta}\in B_r(\boldsymbol{\theta}^*)$. Since $\|\Phi(\boldsymbol{\theta})-\boldsymbol{\theta}^*\|=\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|$ for this fixed point, (Error Bound) is a consequence of the bound above for $\|\Phi(\boldsymbol{\theta})-\boldsymbol{\theta}^*\|$.

Consistency + Stability ⇒ Convergence

The Theorem represents a generalization of the principle that consistency and stability imply convergence. Stability corresponds to (C3) and the Lipschitz continuity of $(\mathcal{F} - \nabla\mathcal{T}(\boldsymbol{\theta}^*))^{-1}$, while consistency corresponds to (C1) $\mathcal{T}(\boldsymbol{\theta}^*) + \boldsymbol{\delta} \in \mathcal{F}(\boldsymbol{\theta}^*)$. Finally, as $\boldsymbol{\delta} \to \mathbf{0}$, (Error Bound) shows that the solution $\boldsymbol{\theta}$ approaches $\boldsymbol{\theta}^*$.

NOTES:

- In practice, we often take $\mathcal{L} = \nabla\mathcal{T}(\boldsymbol{\theta}^*)$, or something close to $\nabla\mathcal{T}(\boldsymbol{\theta}^*)$.
- A further generalization of the theorem can be achieved in the following way: Lipschitz continuity does not need to hold everywhere, but only a ball of radius σ centered at $(\mathcal{T} - \nabla\mathcal{T})(\boldsymbol{\theta}^*)$, where σ is large enough that $\sigma \geq \epsilon r$ and $\sigma \geq \|\boldsymbol{\delta}\|$.