Aspects of numerical analysis in the optimal control of nonlinear PDEs II: state constraints and problems with quasilinear equations

Fredi Tröltzsch

Technische Universität Berlin

Inverse Problems and Optimal Control for PDEs

Warwick, 23-27 May 2011
Outline

- Motivating industrial applications
- Elliptic problems with linear state equation
- Semilinear elliptic state equation
- State-constrained control problems
- The case of quasilinear elliptic equations
- Error estimates
Outline

Pointwise state constraints
- The control problem and necessary conditions
- A test example
- An open problem for SSC

Quasilinear elliptic control problems
- The problem and well-posedness of the state equation
- Optimality conditions
- Approximation by finite elements
The optimal control problem

Let real bounds $\alpha < \beta$, $y_a < 0 < y_b$ be given.

Problem with control and state constraints:

$$(P) \quad \min J(y, u) := \frac{1}{2} \int_{\Omega} (y(x) - y_d(x))^2 \, dx + \frac{\lambda}{2} \int_{\Omega} (u(x))^2 \, dx$$

$$-\Delta y(x) + d(y(x)) = u(x) \quad \text{in } \Omega$$

$$y(x) = 0 \quad \text{on } \Gamma,$$

$$\alpha \leq u(x) \leq \beta, \quad \text{a.e. in } \Omega,$$

$$y_a \leq y(x) \leq y_b \quad \text{for all } x \in \bar{\Omega}.$$
Lagrangian function

It holds \( y_u = G(u), \ G : L^2(\Omega) \to H^1_0(\Omega) \cap C(\bar{\Omega}), \ n \leq 3 \). Therefore, the state-constrained problem can be written as follows:

\[
(P) \quad \min f(u), \quad \alpha \leq u(x) \leq \beta, \quad y_a \leq G(u) \leq y_b.
\]
Lagrangian function

It holds $y_u = G(u)$, $G : L^2(\Omega) \rightarrow H^1_0(\Omega) \cap C(\bar{\Omega})$, $n \leq 3$. Therefore, the state-constrained problem can be written as follows:

$$(P) \quad \min f(u), \quad \alpha \leq u(x) \leq \beta, \quad y_a \leq G(u) \leq y_b.$$ 

For several reasons, we need $G : L^2(\Omega) \rightarrow C(\bar{\Omega})$ or (if $n > 3$), $G : L^p(\Omega) \rightarrow C(\bar{\Omega})$, $p > n/2$. 
It holds $y_u = G(u)$, $G : L^2(\Omega) \to H^1_0(\Omega) \cap C(\bar{\Omega})$, $n \leq 3$. Therefore, the state-constrained problem can be written as follows:

\[
(P) \quad \min f(u), \quad \alpha \leq u(x) \leq \beta, \quad y_a \leq G(u) \leq y_b.
\]

For several reasons, we need $G : L^2(\Omega) \to C(\bar{\Omega})$ or (if $n > 3$), $G : L^p(\Omega) \to C(\bar{\Omega})$, $p > n/2$.

Following the Lagrange formalism, we (formally) remove the state constraints by Lagrange multipliers.
It holds $y_u = G(u)$, $G : L^2(\Omega) \rightarrow H^1_0(\Omega) \cap C(\bar{\Omega})$, $n \leq 3$. Therefore, the state-constrained problem can be written as follows:

\[(P) \quad \min f(u), \quad \alpha \leq u(x) \leq \beta, \quad y_a \leq G(u) \leq y_b.\]

For several reasons, we need $G : L^2(\Omega) \rightarrow C(\bar{\Omega})$ or (if $n > 3$), $G : L^p(\Omega) \rightarrow C(\bar{\Omega})$, $p > n/2$.

Following the Lagrange formalism, we (formally) remove the state constraints by Lagrange multipliers.

Lagrangian function

$$\mathcal{L}(u, \mu_a, \mu_b) := f(u) + \int_{\bar{\Omega}} (y_a - G(u))d\mu_a + \int_{\bar{\Omega}} (G(u) - y_b)d\mu_b.$$
Lagrange multipliers

In $\mathcal{L}$, regular Borel measures $\mu_a, \mu_b$ are Lagrange multipliers associated with the state constraints.

Definition: $\mu_a, \mu_b$ are said to be Lagrange multipliers associated with $\bar{u}$, if
Lagrange multipliers

In $\mathcal{L}$, regular Borel measures $\mu_a, \mu_b$ are Lagrange multipliers associated with the state constraints.

**Definition:** $\mu_a, \mu_b$ are said to be Lagrange multipliers associated with $\bar{u}$, if

- The variational inequality

$$\frac{\partial \mathcal{L}}{\partial u}(\bar{u}, \mu_a, \mu_b)(u - \bar{u}) \geq 0 \quad \forall u \in U_{ad}$$

is satisfied
Lagrange multipliers

In $\mathcal{L}$, regular Borel measures $\mu_a$, $\mu_b$ are Lagrange multipliers associated with the state constraints.

**Definition:** $\mu_a$, $\mu_b$ are said to be Lagrange multipliers associated with $\bar{u}$, if

- The variational inequality

$$\frac{\partial \mathcal{L}}{\partial u}(\bar{u}, \mu_a, \mu_b)(u - \bar{u}) \geq 0 \quad \forall u \in U_{ad}$$

is satisfied (i.e. $\bar{u}$ satisfies the necessary conditions for the problem of minimizing $\mathcal{L}$ subject to $u \in U_{ad}$),
Lagrange multipliers

In $\mathcal{L}$, regular Borel measures $\mu_a, \mu_b$ are Lagrange multipliers associated with the state constraints.

Definition: $\mu_a, \mu_b$ are said to be Lagrange multipliers associated with $\bar{u}$, if

- The variational inequality
  \[
  \frac{\partial \mathcal{L}}{\partial u}(\bar{u}, \mu_a, \mu_b)(u - \bar{u}) \geq 0 \quad \forall u \in U_{ad}
  \]
  is satisfied (i.e. $\bar{u}$ satisfies the necessary conditions for the problem of minimizing $\mathcal{L}$ subject to $u \in U_{ad}$),

- $\mu_a \geq 0, \mu_b \geq 0$ in the sense of $C(\bar{\Omega})^*$,
Lagrange multipliers

In \( L \), regular Borel measures \( \mu_a, \mu_b \) are Lagrange multipliers associated with the state constraints.

**Definition:** \( \mu_a, \mu_b \) are said to be **Lagrange multipliers** associated with \( \bar{u} \), if

- The variational inequality
  \[
  \frac{\partial L}{\partial u}(\bar{u}, \mu_a, \mu_b)(u - \bar{u}) \geq 0 \quad \forall u \in U_{ad}
  \]
  is satisfied (i.e. \( \bar{u} \) satisfies the necessary conditions for the problem of minimizing \( L \) subject to \( u \in U_{ad} \)),

- \( \mu_a \geq 0, \mu_b \geq 0 \) in the sense of \( C(\bar{\Omega})^* \),

- and the following **complementarity conditions** are satisfied:
  \[
  \int_{\bar{\Omega}} (y_a - G(\bar{u})) d\mu_a = 0 = \int_{\bar{\Omega}} (G(\bar{u}) - y_b) d\mu_b.
  \]
Adjoint equation with measures

\[ \mathcal{L}(u, \mu_a, \mu_b) = f(u) + \int_{\tilde{\Omega}} (y_a - G(u)) \, d\mu_a + \int_{\tilde{\Omega}} (G(u) - y_b) \, d\mu_b \]
Adjoint equation with measures

\[ \mathcal{L}(u, \mu_a, \mu_b) = f(u) + \int_{\tilde{\Omega}} (y_a - G(u))d\mu_a + \int_{\tilde{\Omega}} (G(u) - y_b)d\mu_b \]

\[ \frac{\partial \mathcal{L}}{\partial u}(\bar{u}, \mu_a, \mu_b) v = \]
Adjoint equation with measures

\[ \mathcal{L}(u, \mu_a, \mu_b) = f(u) + \int_{\bar{\Omega}} (y_a - G(u)) d\mu_a + \int_{\bar{\Omega}} (G(u) - y_b) d\mu_b \]

\[ \frac{\partial \mathcal{L}}{\partial u}(\bar{u}, \mu_a, \mu_b) v = f'(\bar{u}) v + \int_{\bar{\Omega}} (G'(\bar{u}) v) d(\mu_b - \mu_a) \]

This new adjoint state \( \bar{\phi} \) is the weak solution of an adjoint elliptic equation. The first rigorous mathematical explanation of this fact was given by E. Casas.

Adjoints equation with measures

\[
\mathcal{L}(u, \mu_a, \mu_b) = f(u) + \int_{\Omega} (y_a - G(u)) d\mu_a + \int_{\Omega} (G(u) - y_b) d\mu_b
\]

\[
\frac{\partial \mathcal{L}}{\partial u}(\bar{u}, \mu_a, \mu_b) v = f'(\bar{u}) v + \int_{\Omega} (G'(\bar{u}) v) d(\mu_b - \mu_a)
\]

\[
= \int_{\Omega} (\varphi \bar{u} + \lambda \bar{u}) v \, dx + \int_{\Omega} (G'(\bar{u})^* (\mu_b - \mu_a)) \varphi_{\mu} v \, dx
\]

This new adjoint state \(\bar{\varphi}\) is the weak solution of an adjoint elliptic equation. The first rigorous mathematical explanation of this fact was given by E. Casas.

Adjoint equation with measures

\[ \mathcal{L}(u, \mu_a, \mu_b) = f(u) + \int_{\tilde{\Omega}} (y_a - G(u)) d\mu_a + \int_{\tilde{\Omega}} (G(u) - y_b) d\mu_b \]

\[ \frac{\partial \mathcal{L}}{\partial u}(\tilde{u}, \mu_a, \mu_b) v = f'(\tilde{u}) v + \int_{\tilde{\Omega}} (G'(\tilde{u}) v) d(\mu_b - \mu_a) \]

\[ = \int_{\Omega} (\varphi \tilde{u} + \lambda \tilde{u}) v \, dx + \int_{\Omega} \left( G'(\tilde{u})^{*} (\mu_b - \mu_a) \right) \underbrace{v \, dx}_{\varphi_\mu} \]

\[ = \int_{\Omega} (\varphi \tilde{u} + \varphi_\mu + \lambda \tilde{u}) v \, dx = \int_{\Omega} (\tilde{\varphi} + \lambda \tilde{u}) v \, dx \]
Adjoint equation with measures

\[
\begin{align*}
\mathcal{L}(u, \mu_a, \mu_b) &= f(u) + \int_{\Omega} (y_a - G(u)) d\mu_a + \int_{\Omega} (G(u) - y_b) d\mu_b \\
\frac{\partial \mathcal{L}}{\partial u}(\bar{u}, \mu_a, \mu_b) v &= f'(\bar{u}) v + \int_{\Omega} (G'(\bar{u}) v) d(\mu_b - \mu_a) \\
&= \int_{\Omega} (\varphi \bar{u} + \lambda \bar{u}) v \, dx + \int_{\Omega} \left( G'(\bar{u})^* (\mu_b - \mu_a) \right) v \, dx \\
&= \int_{\Omega} \left( \varphi \bar{u} + \varphi_\mu + \lambda \bar{u} \right) v \, dx = \int_{\Omega} (\tilde{\varphi} + \lambda \bar{u}) v \, dx
\end{align*}
\]

This new adjoint state \( \tilde{\varphi} \) is the weak solution of an adjoint elliptic equation. The first rigorous mathematical explanation of this fact was given by E. Casas.

Theorem (Karush-Kuhn-Tucker conditions)

Let \( \bar{u} \) be locally optimal for (P) and let \( \bar{y} \) the associated state. Assume that a linearized Slater condition is satisfied: \( \exists \tilde{u} \in U_{ad} \) such that

\[
y_a < \left( G(\bar{u}) + G'(\bar{u})(\bar{u} - \tilde{u}) \right)(x) < y_b \quad \forall x \in \bar{\Omega}.
\]
Theorem (Karush-Kuhn-Tucker conditions)

Let $\bar{u}$ be locally optimal for $(P)$ and let $\bar{y}$ the associated state. Assume that a linearized Slater condition is satisfied: $\exists \tilde{u} \in U_{ad}$ such that

$$y_a < \left( G(\bar{u}) + G'(\bar{u})(\bar{u} - \tilde{u}) \right)(x) < y_b \quad \forall x \in \bar{\Omega}.$$ 

Then there exist nonnegative regular Borel measures $\mu_a$, $\mu_b$ on $\bar{\Omega}$ and an adjoint state $\bar{\varphi} \in W^{1,s}(\Omega) \quad \forall s < n/(n-1)$ such that

$$-\Delta \bar{\varphi} + d'(\bar{y})\bar{\varphi} = \bar{y} - y_d + \mu_b - \mu_a$$

$$\bar{\varphi}|_{\Gamma} = 0,$$

$$\int_{\Omega} (\bar{\varphi} + \lambda \bar{u})(u - \bar{u}) \, dx \geq 0 \quad \forall u \in U_{ad},$$

$$\int_{\bar{\Omega}} (\bar{y} - y_b) \, d\mu_b = \int_{\bar{\Omega}} (\bar{y} - y_a) \, d\mu_a = 0.$$
Two main numerical approaches

To solve state-constrained problems numerically, the following options are useful:

\[
\begin{align*}
\min_{u \in U_{\text{ad}}} & \quad f(u) + \rho \int_\Omega \left( (y_a - y) + (y - y_b) \right)^2 \, dx, \quad \rho \gg 0 \\
\to & \quad \text{Moreau-Yosida type regularization.}
\end{align*}
\]

If no control constraints are given, you may also regularize as follows:

\[
y_a \leq y(x) \leq y_b \to y_a \leq \epsilon u(x) + y(x) \leq y_b, \quad \epsilon > 0 \text{ small}
\]

\[
\to \quad \text{Lavrentiev type regularization.}
\]
Two main numerical approaches

To solve state-constrained problems numerically, the following options are useful:

- Discretize and solve the resulting large scale optimization problem by available software.

\[
\min_{u \in U_{ad}} f(u) + \rho \int_\Omega \left( (y_a - y)^2 + (y - y_b)^2 \right) dx, \quad \rho \gg 1
\]

→ Moreau-Yosida type regularization.

If no control constraints are given, you may also regularize as follows:

\[
y_a \leq y(x) \leq y_b \quad \rightarrow \quad y_a \leq \varepsilon u(x) + y(x) \leq y_b, \quad \varepsilon > 0 \text{ small}
\]

→ Lavrentiev type regularization.
Two main numerical approaches

To solve state-constrained problems numerically, the following options are useful:

- Discretize and solve the resulting large scale optimization problem by available software.

- Reduce the problem to a control-constrained one by penalization:

\[
\min_{u \in U_{ad}} f(u) + \rho \int_{\Omega} \left\{ (y_a - y)^2 + (y - y_b)^2 \right\} dx, \quad \rho >> 0
\]

→ Moreau-Yosida type regularization.
Two main numerical approaches

To solve state-constrained problems numerically, the following options are useful:

- Discretize and solve the resulting large scale optimization problem by available software.
- Reduce the problem to a control-constrained one by penalization:

\[
\min_{u \in U_{ad}} f(u) + \rho \int_{\Omega} \left\{ ((y_a - y)_+)^2 + ((y - y_b)_+)^2 \right\} \, dx, \quad \rho \gg 0
\]

→ Moreau-Yosida type regularization.

- If no control constraints are given, you may also regularize as follows:

\[
y_a \leq y(x) \leq y_b \quad \rightarrow \quad y_a \leq \varepsilon u(x) + y(x) \leq y_b, \quad \varepsilon > 0 \text{ small}
\]

→ Lavrentiev type regularization.
Problem with semilinear equation

\[
\begin{align*}
\min & \quad \frac{1}{2} \| y - y_d \|^2 + \frac{\lambda}{2} \| u \|^2 \\
- \Delta y + y + y^3 &= u \quad \text{in } \Omega \\
\partial_\nu y &= 0 \quad \text{on } \Gamma \\
-1 &\leq y(x) \leq 1 \quad \text{in } \Omega
\end{align*}
\]

\[
\text{in } \Omega = (0, 1)^2, \quad y_d = 8 \sin(\pi x_1) \sin(\pi x_2) - 4
\]
Problem with semilinear equation

\[
\begin{align*}
\min \ & \frac{1}{2} \| y - y_d \|^2 + \frac{\lambda}{2} \| u \|^2 \\
- \Delta y + y + y^3 &= u \quad \text{in } \Omega \\
\partial_\nu y &= 0 \quad \text{on } \Gamma \\
-1 &\leq y(x) \leq 1 \quad \text{in } \Omega
\end{align*}
\]

\[\Omega = (0, 1)^2, \quad y_d = 8 \sin(\pi x_1) \sin(\pi x_2) - 4\]

Computations: Christian Meyer, by regularization \(-1 \leq \varepsilon u + y \leq 1\)

Numerical Technique: SQP + primal dual active set strategy
Data: \( \lambda = 10^{-5}, \varepsilon = 10^{-4} \)
Lagrange multipliers $\mu_a, \mu_b$

Data: $\lambda = 10^{-5}, \varepsilon = 10^{-4}$
Sufficient second-order conditions

For non-convex problems, the KKT-conditions are not sufficient for optimality, hence higher-order conditions are needed to check for optimality.

For state-constraints, the difficulty is to show that such SSC are really sufficient for local optimality.
Sufficient second-order conditions

For non-convex problems, the KKT-conditions are not sufficient for optimality, hence higher-order conditions are needed to check for optimality.

General form of second-order sufficient conditions (SSC):

The pair \((\bar{y}, \bar{u})\) satisfies the KKT conditions and there exists \(\delta > 0\) such that

\[
\mathcal{L}''(y, u)(\bar{y}, \bar{u}, \bar{p}, \mu_a, \mu_b)(y, u)^2 \geq \delta \|u\|_{L^2}^2
\]

for all \((y, u)\) belonging to the so-called critical cone (accounts for linearization and active state and control constraints).
Sufficient second-order conditions

For non-convex problems, the KKT-conditions are not sufficient for optimality, hence higher-order conditions are needed to check for optimality.

General form of second-order sufficient conditions (SSC):

The pair $(\bar{y}, \bar{u})$ satisfies the KKT conditions and there exists $\delta > 0$ such that

$$L''_{(y,u)}(\bar{y}, \bar{u}, \bar{p}, \mu_a, \mu_b)(y, u)^2 \geq \delta \|u\|_{L^2}^2$$

for all $(y, u)$ belonging to the so-called critical cone (accounts for linearization and active state and control constraints).

For state-constraints, the difficulty is to show that such SSC are really sufficient for local optimality.
On open problem

We are not able to set up second-order sufficient optimality conditions for important cases of elliptic and parabolic control problems.

Where is the obstacle?
On open problem

We are not able to set up second-order sufficient optimality conditions for important cases of elliptic and parabolic control problems.

Where is the obstacle?

Consider first (P) for the (not that important) case: $n = 4$. 
On open problem

We are not able to set up second-order sufficient optimality conditions for important cases of elliptic and parabolic control problems.

Where is the obstacle?

Consider first (P) for the (not that important) case: \( n = 4 \).

\[
\mathcal{L}(u, \mu_a, \mu_b) = f(u) + \int_{\tilde{\Omega}} (y_a - G(u))d\mu_a + \int_{\tilde{\Omega}} (G(u) - y_b)d\mu_b.
\]
On open problem

We are not able to set up second-order sufficient optimality conditions for important cases of elliptic and parabolic control problems.

Where is the obstacle?

Consider first (P) for the (not that important) case: $n = 4$.

\[
\mathcal{L}(u, \mu_a, \mu_b) = f(u) + \int_{\tilde{\Omega}} (y_a - G(u)) d\mu_a + \int_{\tilde{\Omega}} (G(u) - y_b) d\mu_b.
\]

\[
\frac{\partial \mathcal{L}}{\partial u}(u, \mu_a, \mu_b) \, v = f'(u) \, v + \int_{\tilde{\Omega}} G'(u) \, v \, d(\mu_b - \mu_a).
\]
On open problem

We are not able to set up second-order sufficient optimality conditions for important cases of elliptic and parabolic control problems.

Where is the obstacle?

Consider first (P) for the (not that important) case: $n = 4$.

\[
\mathcal{L}(u, \mu_a, \mu_b) = f(u) + \int_{\tilde{\Omega}} (y_a - G(u)) d\mu_a + \int_{\tilde{\Omega}} (G(u) - y_b) d\mu_b.
\]

\[
\frac{\partial \mathcal{L}}{\partial u}(u, \mu_a, \mu_b) \nu = f'(u) \nu + \int_{\tilde{\Omega}} G'(u) \nu d(\mu_b - \mu_a).
\]

We need the continuity of $\mathcal{L}''$ with respect to $\nu$ in the $L^2$-norm, in particular for the second part.
\[ \left| \int_{\tilde{\Omega}} G'(u) v \, d(\mu_b - \mu_a) \right| \leq c \| v \|_{L^2(\Omega)}. \]
We have

\[
\left| \int_{\tilde{\Omega}} G'(u) v \, d(\mu_b - \mu_a) \right| \leq c \| v \|_{L^2(\Omega)}.
\]

Hence, we need

\[
\| z \|_{C(\tilde{\Omega})} \leq c \| v \|_{L^2(\Omega)},
\]

where

\[
- \Delta z + d'(\tilde{y}) z = v.
\]
\[
\left| \int_{\Omega} G'(u) v \ d(\mu_b - \mu_a) \right| \leq c \|v\|_{L^2(\Omega)}.
\]

We have
\[
\left| \int_{\Omega} z \ d(\mu_b - \mu_a) \right| \leq \|z\|_{C(\bar{\Omega})} \|\mu_b - \mu_a\|_{C(\bar{\Omega})},
\]

hence we need \( \|z\|_{C(\bar{\Omega})} \leq c \|v\|_{L^2(\Omega)}, \) where
\[
-\Delta z + d'(\bar{y})z = v.
\]

However, the mapping \( v \mapsto z \) is not continuous from \( L^2(\Omega) \) to \( C(\bar{\Omega}) \) for \( n > 3 \).
We cannot establish the standard SSC for elliptic distributed control problems with pointwise state constraints, if \( n = \dim \Omega > 3 \). Even with stronger requirements, this problem cannot be fully resolved.

This happens already for \( n > 2 \) in elliptic boundary control, if the state constraints are imposed in the whole domain.

In parabolic distributed control we cannot have more than \( n = 1 \).

There are no SSC for parabolic boundary control problems with state constraints in the whole domain.
Outline

1. Pointwise state constraints
   - The control problem and necessary conditions
   - A test example
   - An open problem for SSC

2. Quasilinear elliptic control problems
   - The problem and well-posedness of the state equation
   - Optimality conditions
   - Approximation by finite elements
Quasilinear control problem

We substitute $\Delta y(x)$ by $\text{div} [a(x, y(x)) \nabla y(x)]$.

\[
(P) \quad \min J(y, u) := \frac{1}{2} \int_{\Omega} (y(x) - y_d(x))^2 dx + \frac{\lambda}{2} \int_{\Omega} u(x)^2 dx \\
- \text{div} [a(x, y(x)) \nabla y(x)] + d(y(x)) = u(x) \quad \text{in} \quad \Omega \\
y(x) = 0 \quad \text{on} \quad \Gamma
\]
Quasilinear control problem

We substitute $\Delta y(x)$ by $\text{div}[a(x, y(x)) \nabla y(x)]$.

\[(P) \quad \min J(y, u) := \frac{1}{2} \int_{\Omega} (y(x) - y_d(x))^2 \, dx + \frac{\lambda}{2} \int_{\Omega} u(x)^2 \, dx \]

\[-\text{div}[a(x, y(x)) \nabla y(x)] + d(y(x)) = u(x) \quad \text{in} \quad \Omega \]

\[y(x) = 0 \quad \text{on} \quad \Gamma \]

\[\alpha \leq u(x) \leq \beta \quad \text{a.e. in} \ \Omega, \quad u \in L^2(\Omega).\]
Quasilinear control problem

We substitute $\Delta y(x)$ by $\text{div} [a(x, y(x)) \nabla y(x)]$.

\[
(P) \quad \min J(y, u) := \frac{1}{2} \int_{\Omega} (y(x) - y_d(x))^2 \, dx + \frac{\lambda}{2} \int_{\Omega} u(x)^2 \, dx \\
- \text{div} [a(x, y(x)) \nabla y(x)] + d(y(x)) = u(x) \quad \text{in} \quad \Omega \\
y(x) = 0 \quad \text{on} \quad \Gamma
\]

\[\alpha \leq u(x) \leq \beta \quad \text{a.e. in} \quad \Omega, \quad u \in L^2(\Omega).\]

Remark:

Even if $y \mapsto a(x, y)$ is monotone, the state equation is not of monotone type!
Assumptions on $a$

The function $a : \Omega \times \mathbb{R} \to \mathbb{R}$ is a Carathéodory function,

$$\exists \alpha_0 > 0 \text{ such that } a(x, y) \geq \alpha_0 \text{ for a.e. } x \in \Omega \text{ and all } y \in \mathbb{R}$$
Assumptions on \( a \)

The function \( a : \Omega \times \mathbb{R} \to \mathbb{R} \) is a Carathéodory function,

\[
\exists \alpha_0 > 0 \text{ such that } a(x, y) \geq \alpha_0 \text{ for a.e. } x \in \Omega \text{ and all } y \in \mathbb{R}
\]

The function \( a(\cdot, 0) \) belongs to \( L^\infty(\Omega) \) and for any \( M > 0 \) there exist a constant \( C_M > 0 \) such that for all \( |y_1|, |y_2| \leq M \)

\[
|a(x, y_2) - a(x, y_1)| \leq C_M |y_2 - y_1| \text{ for a.e. } x \in \Omega.
\]
Assumptions on $a$

The function $a : \Omega \times \mathbb{R} \to \mathbb{R}$ is a Carathéodory function,

$$\exists \alpha_0 > 0 \text{ such that } a(x, y) \geq \alpha_0 \text{ for a.e. } x \in \Omega \text{ and all } y \in \mathbb{R}$$

The function $a(\cdot, 0)$ belongs to $L^\infty(\Omega)$ and for any $M > 0$ there exist a constant $C_M > 0$ such that for all $|y_1|, |y_2| \leq M$

$$|a(x, y_2) - a(x, y_1)| \leq C_M |y_2 - y_1| \text{ for a.e. } x \in \Omega.$$

Remarks:
Assumptions on $a$

The function $a : \Omega \times \mathbb{R} \to \mathbb{R}$ is a Carathéodory function,

$$\exists \alpha_0 > 0 \text{ such that } a(x, y) \geq \alpha_0 \text{ for a.e. } x \in \Omega \text{ and all } y \in \mathbb{R}$$

The function $a(\cdot, 0)$ belongs to $L^\infty(\Omega)$ and for any $M > 0$ there exist a constant $C_M > 0$ such that for all $|y_1|, |y_2| \leq M$

$$|a(x, y_2) - a(x, y_1)| \leq C_M |y_2 - y_1| \text{ for a.e. } x \in \Omega.$$ 

Remarks:

- Instead of $d(y)$, a more general function $d(x, y)$ can be considered under associated assumptions.
Assumptions on $a$

The function $a : \Omega \times \mathbb{R} \to \mathbb{R}$ is a Carathéodory function,

$$\exists \alpha_0 > 0 \text{ such that } a(x, y) \geq \alpha_0 \text{ for a.e. } x \in \Omega \text{ and all } y \in \mathbb{R}$$

The function $a(\cdot, 0)$ belongs to $L^\infty(\Omega)$ and for any $M > 0$ there exist a constant $C_M > 0$ such that for all $|y_1|, |y_2| \leq M$

$$|a(x, y_2) - a(x, y_1)| \leq C_M |y_2 - y_1| \text{ for a.e. } x \in \Omega.$$ 

Remarks:

- Instead of $d(y)$, a more general function $d(x, y)$ can be considered under associated assumptions.

- We shall also need the derivatives $\frac{\partial a}{\partial y}(x, y)$ and $\frac{\partial^2 a}{\partial y^2}(x, y)$. 
Well-posedness of the state equation

Define: \( p > n \) and \( q > n/2 \).

**Theorem**

Under our assumptions, for any element \( u \in W^{-1,p}(\Omega) \), the quasilinear state equation has a unique solution \( y_u \in H^1_0(\Omega) \cap L^\infty(\Omega) \). Moreover there exists \( \mu \in (0,1) \) independent of \( u \) such that \( y_u \in C^\mu(\bar{\Omega}) \) and for any bounded set \( U \subset W^{-1,p}(\Omega) \)

\[
\|y_u\|_{H^1_0(\Omega)} + \|y_u\|_{C^\mu(\bar{\Omega})} \leq C_U \quad \forall u \in U
\]

for some constant \( C_U > 0 \). Finally, if \( u_k \to u \) in \( W^{-1,p}(\Omega) \), then \( y_{u_k} \to y_u \) in \( H^1_0(\Omega) \cap C^\mu(\bar{\Omega}) \).
Idea of proof:

Depending on $M > 0$, we introduce the truncated function $a_M$ by

$$a_M(x, y) = \begin{cases} a(x, y), & |y| \leq M \\ a(x, y + M), & y > M \\ a(x, y - M), & y < -M. \end{cases}$$

Analogously, the truncation $d_M$ of $d$ is defined.

We prove that $-\text{div} \left[a_M(x, y) \nabla y\right] + d_M(y) = u$ in $\Omega$ with $y = 0$ on $\Gamma$ has at least one solution $y \in H^{1,0}(\Omega)$.

For fixed $u$, consider the linear equation $-\text{div} \left[a_M(x, z) \nabla y\right] + d_M(z) = u$ in $\Omega$ with $y = 0$ on $\Gamma$.

Define $F : L^2(\Omega) \to L^2(\Omega)$ by $F : z \mapsto y$.

Compact embedding of $H^1(\Omega)$ in $L^2(\Omega)$, Schauder fixed point theorem $\Rightarrow F$ has a fixed point $y_M$. 

Fredi Tröltzsch (TU Berlin)
Idea of proof:

a) **Existence:** Depending on $M > 0$, we introduce the truncated function $a_M$ by

$$a_M(x, y) = \begin{cases} 
  a(x, y), & |y| \leq M \\
  a(x, +M), & y > +M \\
  a(x, -M), & y < -M.
\end{cases}$$
Idea of proof:

a) **Existence:** Depending on $M > 0$, we introduce the truncated function $a_M$ by

$$a_M(x, y) = \begin{cases} 
  a(x, y), & |y| \leq M \\
  a(x, +M), & y > +M \\
  a(x, -M), & y < -M.
\end{cases}$$

Analogously, the truncation $d_M$ of $d$ is defined.
Idea of proof:

a) **Existence:** Depending on $M > 0$, we introduce the truncated function $a_M$ by

$$a_M(x, y) = \begin{cases} 
  a(x, y), & |y| \leq M \\
  a(x, +M), & y > +M \\
  a(x, -M), & y < -M.
\end{cases}$$

Analogously, the truncation $d_M$ of $d$ is defined. We prove that

$$-\text{div} \left[ a_M(x, y) \nabla y \right] + d_M(y) = u \quad \text{in } \Omega$$
$$y = 0 \quad \text{on } \Gamma$$

has at least one solution $y \in H^1_0(\Omega)$. 
Idea of proof:

a) **Existence:** Depending on $M > 0$, we introduce the truncated function $a_M$ by

$$a_M(x, y) = \begin{cases} 
a(x, y), & |y| \leq M \\
a(x, +M), & y > +M \\
a(x, -M), & y < -M. 
\end{cases}$$

Analogously, the truncation $d_M$ of $d$ is defined. We prove that

$$- \text{div} [a_M(x, y) \nabla y] + d_M(y) = u \quad \text{in } \Omega$$

$$y = 0 \quad \text{on } \Gamma$$

has at least one solution $y \in H^1_0(\Omega)$. For fixed $u$, consider the linear equation

$$- \text{div} [a_M(x, z) \nabla y] + d_M(z) = u \quad \text{in } \Omega$$

$$y = 0 \quad \text{on } \Gamma.$$
Idea of proof:

a) **Existence:** Depending on $M > 0$, we introduce the truncated function $a_M$ by

$$a_M(x, y) = \begin{cases} 
    a(x, y), & |y| \leq M \\
    a(x, +M), & y > +M \\
    a(x, -M), & y < -M.
\end{cases}$$

Analogously, the truncation $d_M$ of $d$ is defined. We prove that

$$-\text{div} [a_M(x, y) \nabla y] + d_M(y) = u \quad \text{in } \Omega \quad y = 0 \quad \text{on } \Gamma$$

has at least one solution $y \in H_0^1(\Omega)$. For fixed $u$, consider the linear equation

$$-\text{div} [a_M(x, z) \nabla y] + d_M(z) = u \quad \text{in } \Omega \quad y = 0 \quad \text{on } \Gamma.$$ 

Define $F : L^2(\Omega) \to L^2(\Omega)$ by $F : z \mapsto y$. 
Idea of proof:

a) **Existence:** Depending on $M > 0$, we introduce the truncated function $a_M$ by

\[
    a_M(x, y) = \begin{cases} 
    a(x, y), & |y| \leq M \\
    a(x, +M), & y > +M \\
    a(x, -M), & y < -M.
    \end{cases}
\]

Analogously, the truncation $d_M$ of $d$ is defined. We prove that

\[
    -\text{div} \left[ a_M(x, y) \nabla y \right] + d_M(y) = u \quad \text{in } \Omega \\
    y = 0 \quad \text{on } \Gamma
\]

has at least one solution $y \in H^1_0(\Omega)$. For fixed $u$, consider the linear equation

\[
    -\text{div} \left[ a_M(x, z) \nabla y \right] + d_M(z) = u \quad \text{in } \Omega \\
    y = 0 \quad \text{on } \Gamma.
\]

Define $F : L^2(\Omega) \to L^2(\Omega)$ by $F : z \mapsto y$. Compact embedding of $H^1(\Omega)$ in $L^2(\Omega)$, Schauder fixed point theorem.
Idea of proof:

a) **Existence:** Depending on \( M > 0 \), we introduce the truncated function \( a_M \) by

\[
a_M(x, y) = \begin{cases} 
  a(x, y), & |y| \leq M \\
  a(x, +M), & y > +M \\
  a(x, -M), & y < -M.
\end{cases}
\]

Analogously, the truncation \( d_M \) of \( d \) is defined. We prove that

\[
-\text{div} \left[ a_M(x, y) \nabla y \right] + d_M(y) = u \quad \text{in} \quad \Omega \\
y = 0 \quad \text{on} \quad \Gamma
\]

has at least one solution \( y \in H^1_0(\Omega) \). For fixed \( u \), consider the linear equation

\[
-\text{div} \left[ a_M(x, z) \nabla y \right] + d_M(z) = u \quad \text{in} \quad \Omega \\
y = 0 \quad \text{on} \quad \Gamma.
\]

Define \( F : L^2(\Omega) \to L^2(\Omega) \) by \( F : z \mapsto y \). Compact embedding of \( H^1(\Omega) \) in \( L^2(\Omega) \), Schauder fixed point theorem \( \Rightarrow F \) has a fixed point \( y_M \).
Stampacchia truncation method ⇒

$$\| y_M \|_{L^\infty(\Omega)} \leq c_\infty,$$

where $c_\infty$ does not depend on $M$. 

Taking $M$ sufficiently large, the solution $y_M$ is shown to be a solution of the state equation.

Hölder regularity of $y$: results of Gilbarg and Trudinger.

Stampacchia truncation method ⇒

\[ \|y_M\|_{L^\infty(\Omega)} \leq c_\infty, \]

where \( c_\infty \) does not depend on \( M \). Taking \( M \) sufficiently large, the solution \( y_M \) is shown to be a solution of the state equation.
Stampacchia truncation method ⇒

\[ \| y_M \|_{L^\infty(\Omega)} \leq c_\infty, \]

where \( c_\infty \) does not depend on \( M \). Taking \( M \) sufficiently large, the solution \( y_M \) is shown to be a solution of the state equation.

Hölder regularity of \( y \): results of Gilbarg and Trudinger.
Stampacchia truncation method ⇒

$$\|y_M\|_{L^\infty(\Omega)} \leq c_\infty,$$

where $c_\infty$ does not depend on $M$. Taking $M$ sufficiently large, the solution $y_M$ is shown to be a solution of the state equation.

Hölder regularity of $y$: results of Gilbarg and Trudinger.

b) **Uniqueness:** First surprise: Very delicate!

Application of a comparison principle; we use ideas of Douglas/Dupont/Serrin (1971) and Křížek/Liu (2003).
W^{1,p}-regularity

Assume slightly higher regularity of \( a, \Gamma \) and \( u \):

**Theorem**

Assume in addition that \( a : \bar{\Omega} \times \mathbb{R} \rightarrow \mathbb{R} \) is continuous and \( \Gamma \) is of class \( C^1 \). Then the state equation has a unique solution \( y_u \in W^{1,p}_0(\Omega) \). Moreover, for any bounded set \( U \subset W^{-1,p}(\Omega) \), there exists a constant \( C_U > 0 \) such that

\[
\| y_u \|_{W^{1,p}_0(\Omega)} \leq C_U \quad \forall u \in U.
\]

If \( u_k \rightarrow u \) in \( W^{-1,p}(\Omega) \) then \( y_{u_k} \rightarrow y_u \) strongly in \( W^{1,p}_0(\Omega) \).
Assume slightly higher regularity of $a$, $\Gamma$ and $u$:

**Theorem**

Assume in addition that $a : \bar{\Omega} \times \mathbb{R} \rightarrow \mathbb{R}$ is continuous and $\Gamma$ is of class $C^1$. Then the state equation has a unique solution $y_u \in W^{1,p}_0(\Omega)$. Moreover, for any bounded set $U \subset W^{-1,p}(\Omega)$, there exists a constant $C_U > 0$ such that

$$\|y_u\|_{W^{1,p}_0(\Omega)} \leq C_U \quad \forall u \in U.$$  

If $u_k \rightarrow u$ in $W^{-1,p}(\Omega)$ then $y_{u_k} \rightarrow y_u$ strongly in $W^{1,p}_0(\Omega)$.

Follows from $W^{1,p}(\Omega)$-results for linear elliptic equations; Giaquinta (1993) and Morrey (1966).

Notice that $\hat{a}(x) = a(x, y_u(x))$ is continuous in $\bar{\Omega}$ and $u - d(y_u) \in W^{-1,p}(\Omega)$. 

$W^{1,p}$-regularity
Assume more smoothness of $a$:

$$|a(x_1, y_1) - a(x_2, y_2)| \leq c_M \left( |x_1 - x_2| + |y_1 - y_2| \right)$$

for all $x_i \in \bar{\Omega}$, $y_i \in [-M, M]$, $i = 1, 2$. 
Assume more smoothness of $a$:

$$|a(x_1, y_1) - a(x_2, y_2)| \leq c_M \left\{ |x_1 - x_2| + |y_1 - y_2| \right\}$$

for all $x_i \in \bar{\Omega}$, $y_i \in [-M, M]$, $i = 1, 2$.

**Theorem**

Let this additional assumption be satisfied and $\Gamma$ be of class $C^{1,1}$. Then for any $u \in L^q(\Omega)$, the quasilinear equation has one solution $y_u \in W^{2,q}(\Omega)$. Moreover, for any bounded set $U \subset L^q(\Omega)$, there exists a constant $C_U > 0$ such that

$$\|y_u\|_{W^{2,q}(\Omega)} \leq C_U \quad \forall u \in U.$$
Main trick of the proof: Expand the divergence term $a(x, y)$ and divide by $a$: 

\[
\nabla y = \frac{1}{a} \left\{ u - d(y) + \sum_{j=1}^{n} \partial_j a(x, y) \right\} + \partial_a \partial_y \left| \nabla y \right|^2
\]

$\Rightarrow$ right-hand side in $L^q(\Omega)$. By the Lipschitz property and $y \in L^\infty(\Omega)$. The $C^1, 1$-smoothness of $\Gamma$ permits to apply a result by Grisvard (1985) to get $y \in W^{2, q}(\Omega)$. The case $n/2 < q < n$ follows by some embedding results. □
Main trick of the proof: Expand the divergence term $a(x,y)$ and divide by $a$: We have $y \in W^{1,p}(\Omega)$ for all $p < \infty$, in particular in $W^{1,2q}(\Omega)$.

Consider the case $q \geq n$. 
Main trick of the proof: Expand the divergence term $a(x, y)$ and divide by $a$.

We have $y \in W^{1,p}(\Omega)$ for all $p < \infty$, in particular in $W^{1,2q}(\Omega)$.

Consider the case $q \geq n$.

$$-\Delta y = \frac{1}{a} \left\{ u - d(y) + \sum_{j=1}^{n} \frac{\partial_j a(x, y)}{L^\infty} \frac{\partial_j y}{L^q} + \frac{\partial a}{\partial y} \frac{|\nabla y|^2}{L^q} \right\},$$

$\Rightarrow$ right-hand side in $L^q(\Omega)$.

$\frac{\partial a}{\partial y} \in L^\infty$: By the Lipschitz property and $y \in L^\infty(\Omega)$.
Main trick of the proof: Expand the divergence term \( a(x, y) \) and divide by \( a \): We have \( y \in W^{1,p}(\Omega) \) for all \( p < \infty \), in particular in \( W^{1,2q}(\Omega) \).

Consider the case \( q \geq n \).

\[
-\Delta y = \frac{1}{a} \left\{ u - d(y) + \sum_{j=1}^{n} \partial_j a(x, y) \partial_j y + \frac{\partial a}{\partial y} |\nabla y|^2 \right\},
\]

\( \Rightarrow \) right-hand side in \( L^q(\Omega) \).

\( \frac{\partial a}{\partial y} \in L^\infty \): By the Lipschitz property and \( y \in L^\infty(\Omega) \).

The \( C^{1,1} \)-smoothness of \( \Gamma \) permits to apply a result by Grisvard (1985) to get \( y \in W^{2,q}(\Omega) \). The case \( n/2 < q < n \) follows by some embedding results. \( \square \)
Since $n \leq 3$, $q = 2 > n/2$ is satisfied.

Therefore, $G : u \mapsto y_u$ is continuous from $L^2(\Omega)$ to $H^2(\Omega) \cap H^1_0(\Omega)$.

The choice $q = 2$ is possible in the theorems below.
Differentiability of $G$

Since $n \leq 3$, $q = 2 > n/2$ is satisfied.

Therefore, $G : u \mapsto y_u$ is continuous from $L^2(\Omega)$ to $H^2(\Omega) \cap H^1_0(\Omega)$.

The choice $q = 2$ is possible in the theorems below.

Additional assumption:

The function $a$ is of class $C^2$ with respect to the second variable and, $\forall \, M > 0$ 
$\exists \, D_M > 0$ such that

$$\left| \frac{\partial a}{\partial y} (x, y) \right| + \left| \frac{\partial^2 a}{\partial y^2} (x, y) \right| \leq D_M \text{ for a.e. } x \in \Omega \text{ and all } |y| \leq M.$$
Differentiability of $G$

Since $n \leq 3$, $q = 2 > n/2$ is satisfied.

Therefore, $G : u \mapsto y_u$ is continuous from $L^2(\Omega)$ to $H^2(\Omega) \cap H^1_0(\Omega)$. The choice $q = 2$ is possible in the theorems below.

Additional assumption:

The function $a$ is of class $C^2$ with respect to the second variable and, $\forall M > 0$ $\exists D_M > 0$ such that

$$\left| \frac{\partial a}{\partial y}(x, y) \right| + \left| \frac{\partial^2 a}{\partial y^2}(x, y) \right| \leq D_M \text{ for a.e. } x \in \Omega \text{ and all } |y| \leq M.$$ 

Next surprise: The differentiability of $G$ is very delicate, too.
Differentiability of $G$

Since $n \leq 3$, $q = 2 > n/2$ is satisfied.

Therefore, $G : u \mapsto y_u$ is continuous from $L^2(\Omega)$ to $H^2(\Omega) \cap H^1_0(\Omega)$.

The choice $q = 2$ is possible in the theorems below.

Additional assumption:

The function $a$ is of class $C^2$ with respect to the second variable and, $\forall M > 0$ $\exists D_M > 0$ such that

$$
\left| \frac{\partial a}{\partial y}(x, y) \right| + \left| \frac{\partial^2 a}{\partial y^2}(x, y) \right| \leq D_M \text{ for a.e. } x \in \Omega \text{ and all } |y| \leq M.
$$

Next surprise: The differentiability of $G$ is very delicate, too.

Differentiability will hold, if the linearized equation defines an isomorphism in the associated spaces.
Theorem

Given $y \in W^{1,p}(\Omega)$, for any $v \in H^{-1}(\Omega)$ the linearized equation

$$-\text{div} \left[ a(x, y) \nabla z + \frac{\partial a}{\partial y}(x, y) z \nabla y \right] + d'(y) z = v \text{ in } \Omega$$

$$z = 0 \text{ on } \Gamma$$

has a unique solution $z_v \in H^1_0(\Omega)$.
Theorem

Given $y \in W^{1,p}(\Omega)$, for any $v \in H^{-1}(\Omega)$ the linearized equation

$$-\text{div} \left[ a(x,y) \nabla z + \frac{\partial a}{\partial y}(x,y)z \nabla y \right] + d'(y) z = v \text{ in } \Omega$$

$$z = 0 \text{ on } \Gamma$$

has a unique solution $z_v \in H^1_0(\Omega)$.

Steps of the proof:

a) The uniqueness is shown by a comparison principle as for the state equation.
Idea of proof

b) A homotopy with respect to $t \in [0, 1]$ is considered:

$$-\text{div} \left[ a(x, y) \nabla z + t \frac{\partial a}{\partial y}(x, y) z \nabla y u \right] + d'(y) z = v \text{ in } \Omega$$

$$z = 0 \text{ on } \Gamma.$$
Idea of proof

b) A homotopy with respect to $t \in [0, 1]$ is considered:

$$\begin{align*}
-\text{div} \left[ a(x, y) \nabla z + t \frac{\partial a}{\partial y}(x, y)z \nabla y u \right] + d'(y) z &= v \quad \text{in } \Omega \\
z &= 0 \quad \text{on } \Gamma.
\end{align*}$$

For $t = 0$: Apply the Lax-Milgram Theorem.
There exists a unique solution $z_0 \in H^1_0(\Omega)$ for every $v \in H^{-1}(\Omega)$. 
b) A homotopy with respect to \( t \in [0, 1] \) is considered:

\[
- \text{div} \left[ a(x, y) \nabla z + t \frac{\partial a}{\partial y}(x, y)z \nabla yu \right] + d'(y) z = v \quad \text{in } \Omega
\]

\[
z = 0 \quad \text{on } \Gamma.
\]

- For \( t = 0 \): Apply the Lax-Milgram Theorem.
  There exists a unique solution \( z_0 \in H^1_0(\Omega) \) for every \( v \in H^{-1}(\Omega) \).

- Let \( S \) be the set of points \( t \in [0, 1] \) for which the equation above defines an isomorphism between \( H^1_0(\Omega) \) and \( H^{-1}(\Omega) \); \( 0 \in S \).
Idea of proof

b) A homotopy with respect to \( t \in [0, 1] \) is considered:

\[
-\text{div} \left[ a(x, y) \nabla z + t \frac{\partial a}{\partial y}(x, y)z \nabla y u \right] + d'(y) z = v \quad \text{in} \ \Omega
\]

\[
z = 0 \quad \text{on} \ \Gamma.
\]

- For \( t = 0 \): Apply the Lax-Milgram Theorem.
  There exists a unique solution \( z_0 \in H^1_0(\Omega) \) for every \( v \in H^{-1}(\Omega) \).

- Let \( S \) be the set of points \( t \in [0, 1] \) for which the equation above defines an isomorphism between \( H^1_0(\Omega) \) and \( H^{-1}(\Omega) \); \( 0 \in S \).

- \( t_{max} := \sup S \). First, it is shown \( t_{max} \in S \) and second \( t_{max} = 1 \). \( \square \)
Let all previous assumptions be satisfied. Then $G : W^{-1,p}(\Omega) \rightarrow W^{1,p}_0(\Omega)$, $G : u \mapsto y_u$, is of class $C^2$. 

For any $v \in W^{-1,p}(\Omega)$ the function $z_v = G'(u)v$ is the unique solution in $W^{1,p}_0(\Omega)$ of

$$-\text{div} \left[ a(x,y_u) \nabla z + \frac{\partial a}{\partial y}(x,y_u) z \nabla y_u \right] + d'(y_u)z = v \text{ in } \Omega,$$

$$z = 0 \text{ on } \Gamma.$$ 

For all $v_1, v_2 \in W^{-1,p}(\Omega)$ the function $z_{v_1}, v_2 = G''(u)[v_1, v_2]$ is the unique solution in $W^{1,p}_0(\Omega)$ of

$$-\text{div} \left[ a(x,y_u) \nabla z + \frac{\partial a}{\partial y}(x,y_u) z \nabla y_u \right] + d'(y_u)z = -d''(y_u)v_1z + \text{div} \left[ \frac{\partial^2 a}{\partial y^2}(x,y_u)z \nabla y_u \right] \text{ in } \Omega,$$

$$z = 0 \text{ on } \Gamma.$$ 

respectively, where $z_{v_i} = G'(u)v_i$, $i = 1, 2$. 

Fredi Tröltzsch (TU Berlin)
Theorem

Let all previous assumptions be satisfied. Then $G : W^{-1,p}(\Omega) \to W^{1,p}_0(\Omega)$, $G : u \mapsto y_u$, is of class $C^2$. For any $v \in W^{-1,p}(\Omega)$ the function $z_v = G'(u)v$ is the unique solution in $W^{1,p}_0(\Omega)$ of

$$ -\text{div} \left[ a(x, y_u) \nabla z + \frac{\partial a}{\partial y}(x, y_u) z \nabla y_u \right] + d'(y) z = v \text{ in } \Omega $$

$$ z = 0 \text{ on } \Gamma. $$
Theorem

Let all previous assumptions be satisfied. Then $G : W^{-1,p}(\Omega) \rightarrow W^{1,p}_0(\Omega)$, $G : u \mapsto y_u$, is of class $C^2$. For any $v \in W^{-1,p}(\Omega)$ the function $z_v = G'(u)v$ is the unique solution in $W^{1,p}_0(\Omega)$ of

$$-
\text{div} \left[ a(x, y_u) \nabla z + \frac{\partial a}{\partial y}(x, y_u) z \nabla y_u \right] + d'(y) z = v \text{ in } \Omega$$

$$z = 0 \text{ on } \Gamma.$$ 

For all $v_1, v_2 \in W^{-1,p}(\Omega)$ the function $z_{v_1, v_2} = G''(u)[v_1, v_2]$ is the unique solution in $W^{1,p}_0(\Omega)$ of

$$-
\text{div} \left[ a(x, y_u) \nabla z + \frac{\partial a}{\partial y}(x, y_u) z \nabla y_u \right] + d'(y_u) z = -d''(y_u)z_{v_1}z_{v_2}$$

$$+ \text{div} \left[ \frac{\partial a}{\partial y}(x, y_u)(z_{v_1} \nabla z_{v_2} + \nabla z_{v_1} z_{v_2}) + \frac{\partial^2 a}{\partial y^2}(x, y_u)z_{v_1}z_{v_2} \nabla y_u \right] \text{ in } \Omega$$

$$z = 0 \text{ on } \Gamma.$$ 

respectively, where $z_{v_i} = G'(u)v_i$, $i = 1, 2$. 

Fredi Tröltzsch (TU Berlin) 
Numerical Analysis 
Workshop Warwick 29 / 42
Other spaces for $G'$

**Additional assumption:** $\forall \ M > 0 \ \exists c_M > 0$ such that

$$\left| \frac{\partial^j a}{\partial y^j}(x_1, y_1) - \frac{\partial^j a}{\partial y^j}(x_2, y_2) \right| \leq d_M \{ |x_1 - x_2| + |y_1 - y_2| \}$$

for all $x_i \in \tilde{\Omega}$, $y_i \in [-M, M]$, $i = 1, 2$ and $j = 1, 2$.

**Theorem**

*Let all previous assumptions be satisfied and $\Gamma$ be of class $C^{1,1}$. Then the control-to-state mapping $G : L^q(\Omega) \rightarrow W^{2,q}(\Omega)$, $G(u) = y_u$, is of class $C^2$ for all $q > n/2$.***
Adjoint equation

With theses prerequisites, first-order necessary and second-order sufficient optimality conditions can be shown. Take $q := 2$ in the sequel.
Adjoint equation

With theses prerequisites, first-order necessary and second-order sufficient optimality conditions can be shown. Take $q := 2$ in the sequel.

**Adjoint equation**: Associated with $u$, the adjoint state $\varphi_u \in H^2(\Omega) \cap H^1_0(\Omega)$ is obtained from

$$-\text{div} [a(x, y_u) \nabla \varphi] + \frac{\partial a}{\partial y}(x, y_u) \nabla y_u \cdot \nabla \varphi + d'(y_u) \varphi = y_u - y_d \quad \text{in } \Omega$$

$$\varphi = 0 \quad \text{on } \Gamma$$
Adjoint equation

With theses prerequisites, first-order necessary and second-order sufficient optimality conditions can be shown. Take \( q := 2 \) in the sequel

**Adjoint equation:** Associated with \( u \), the adjoint state \( \varphi_u \in H^2(\Omega) \cap H^1_0(\Omega) \) is obtained from

\[
-\text{div} \left[ a(x, y_u) \nabla \varphi \right] + \frac{\partial a}{\partial y}(x, y_u) \nabla y_u \cdot \nabla \varphi + d'(y_u) \varphi = y_u - y_d \quad \text{in } \Omega \\
\varphi = 0 \quad \text{on } \Gamma
\]

**Reduced gradient:** Define as before \( f(u) := J(y_u, u) = J(G(u), u) \).

\[
f'(u) \nu = \int_{\Omega} (\varphi_u(x) + \lambda u(x)) \nu(x) \, dx
\]
Adjoints equation

With these prerequisites, first-order necessary and second-order sufficient optimality conditions can be shown. Take \( q := 2 \) in the sequel.

**Adjoints equation:** Associated with \( u \), the adjoint state \( \varphi_u \in H^2(\Omega) \cap H^1_0(\Omega) \) is obtained from

\[
-\text{div} \left[ a(x, y_u) \nabla \varphi \right] + \frac{\partial a}{\partial y}(x, y_u) \nabla y_u \cdot \nabla \varphi + d'(y_u) \varphi = y_u - y_d \quad \text{in} \ \Omega
\]

\[
\varphi = 0 \quad \text{on} \ \Gamma
\]

**Reduced gradient:** Define as before \( f(u) := J(y_u, u) = J(G(u), u) \).

\[
f'(u) v = \int_{\Omega} (\varphi_u(x) + \lambda u(x)) v(x) \, dx
\]

**Riesz identification:** \( f'(u) \approx \varphi_u + \lambda u \).
First-order necessary condition

**Theorem**

If \( \bar{u} \) is locally optimal for \( (P) \) (in the sense of \( L^2 \)) and \( \bar{\varphi} := \varphi \bar{u} \) is the associated adjoint state, then

\[
\int_{\Omega} (\bar{\varphi} + \lambda \bar{u})(u - \bar{u}) \, dx \geq 0 \quad \forall u \in U_{ad}.
\]

This is equivalent to the projection formula

\[
\bar{u}(x) = \mathbb{P}_{[\alpha, \beta]} \left( -\frac{\bar{\varphi}(x)}{\lambda} \right) \quad a.e. \text{ in } \Omega.
\]

This result gives different options for the numerical treatment.
The nonsmooth optimality system

Optimality system

\[- \text{div} [a(x, y) \nabla y] + d(y) = P_{[\alpha, \beta]}(\lambda^{-1} \varphi)\]

\[-\text{div} [a(x, y) \nabla \varphi] + \frac{\partial a}{\partial y}(x, y) \nabla y \cdot \nabla \varphi + d'(y) \varphi = y - y_d\]

(in Ω subject to homogeneous Dirichlet boundary condition.)
The nonsmooth optimality system

Optimality system

\[ - \text{div} [a(x, y) \nabla y] + d(y) = P_{[\alpha, \beta]}(\lambda^{-1} \varphi) \]

\[ - \text{div} [a(x, y) \nabla \varphi] + \frac{\partial a}{\partial y}(x, y) \nabla y \cdot \nabla \varphi + d'(y) \varphi = y - y_d \]

(in \( \Omega \) subject to homogeneous Dirichlet boundary condition.)

Numerical options:

- Semismooth Newton method
- Direct solution of the system by COMSOL Multiphysics

Both methods were tested by V. Dhamo (TU Berlin) – very good experience.
Second-order derivative of $f$

For error estimates and the local convergence of numerical methods we need again second-order sufficient optimality conditions.

**Theorem**

*Under our previous assumptions, the functional $f : L^2(\Omega) \to \mathbb{R}$ is of class $C^2$. We have*

$$J''(u)v_1 v_2 = \int_\Omega \left\{ z_{v_1} z_{v_2} + \lambda v_1 v_2 - \varphi_u d''(u)z_{v_1}z_{v_2}$$

$$- \nabla \varphi_u \left[ \frac{\partial a}{\partial y}(x, y_u)(z_{v_1} \nabla z_{v_2} + \nabla z_{v_1} z_{v_2}) + \frac{\partial^2 a}{\partial y^2}(x, y)z_{v_1}z_{v_2} \nabla y_u \right] \right\} dx$$

*where $\varphi_u \in W^{1,p}_0(\Omega) \cap W^{2,q}(\Omega)$ is the adjoint state associated with $u$ and $z_{v_i} = G'(u)v_i$.***
Second-order sufficient optimality condition

**Theorem**

Assume that \( \bar{u} \in U_{ad} \) satisfies the first-order necessary optimality conditions with the associated adjoint state \( \bar{\varphi} \in W^{1,p}_0(\Omega) \).

Let there exist \( \delta, \tau > 0 \) such that

\[
f''(\bar{u})v^2 \geq \delta \|v\|_{L^2(\Omega)}^2 \quad \forall v \in C_{\bar{u}}^\tau
\]

where

\[
C_{\bar{u}}^\tau = \left\{ v \in L^2(\Omega) : v(x) = \begin{cases} 
\geq 0 & \text{if } \bar{u}(x) = \alpha \\
\leq 0 & \text{if } \bar{u}(x) = \beta \\
= 0 & \text{if } |\bar{\varphi}(x) + \lambda \bar{u}(x)| > \tau 
\end{cases} \text{ for a.e. } x \in \Omega \right\}.
\]

Then \( \bar{u} \) is locally optimal in the sense of \( L^2(\Omega) \).
Remarks

- No two-norm discrepancy (quadratic structure of \( f \)).
- We discussed more general functionals of the form

\[
f(u) = \int_{\Omega} L(x, y_u, u) \, dx.
\]

Here the two-norm discrepancy will occur in general.

- The condition \( f''(\bar{u}) v^2 > 0 \) for all nonzero \( v \) of the critical cone is equivalent to the condition above under some additional requirements on the Hamiltonian.
Approximation by finite elements

Family of regular triangulations: \( \{ T_h \}_{h>0} \) of \( \bar{\Omega} \):

Associate to all \( T \in \mathcal{T}_h \) the numbers \( \rho(T) \) (diameter of \( T \)) and \( \sigma(T) \) (diameter of the largest ball in \( T \)).

\[ h := \max_{T \in \mathcal{T}_h} \rho(T) \text{ (mesh size)} \]

Regularity assumptions:

\[ \exists \rho > 0, \sigma > 0 \text{ such that } \rho(T) \sigma(T) \leq \sigma, \quad h \rho(T) \leq \rho \quad \forall T \in \mathcal{T}_h, \quad h > 0. \]

Define \( \Omega_h = \bigcup_{T \in \mathcal{T}_h} T \) with interior \( \Omega_h \) and boundary \( \Gamma_h \).

Assume that \( \Omega_h \) is convex and that the vertices of \( T_h \) placed on the boundary \( \Gamma_h \) are points of \( \Gamma_h \).
Approximation by finite elements

Family of regular triangulations: \( \{T_h\}_{h>0} \) of \( \bar{\Omega} \):

Associate to all \( T \in T_h \) the numbers \( \rho(T) \) (diameter of \( T \)) and \( \sigma(T) \) (diameter of the largest ball in \( T \)).
Approximation by finite elements

Family of regular triangulations: \( \{ \mathcal{T}_h \}_{h>0} \) of \( \bar{\Omega} \):

Associate to all \( T \in \mathcal{T}_h \) the numbers \( \rho(T) \) (diameter of \( T \)) and \( \sigma(T) \) (diameter of the largest ball in \( T \)).

\[
h := \max_{T \in \mathcal{T}_h} \rho(T) \quad \text{(mesh size)}
\]
Approximation by finite elements

Family of regular triangulations: \( \{I_h\}_{h>0} \) of \( \bar{\Omega} \):

Associate to all \( T \in I_h \) the numbers \( \rho(T) \) (diameter of \( T \)) and \( \sigma(T) \) (diameter of the largest ball in \( T \)).

\[
h := \max_{T \in I_h} \rho(T) \quad \text{(mesh size)}
\]

Regularity assumptions:
Approximation by finite elements

Family of regular triangulations: \( \{ \mathcal{T}_h \}_{h>0} \) of \( \bar{\Omega} \):

Associate to all \( T \in \mathcal{T}_h \) the numbers \( \rho(T) \) (diameter of \( T \)) and \( \sigma(T) \) (diameter of the largest ball in \( T \)).

\[
h := \max_{T \in \mathcal{T}_h} \rho(T) \quad \text{(mesh size)}
\]

Regularity assumptions:

- \( \exists \rho > 0, \sigma > 0 \) such that

\[
\frac{\rho(T)}{\sigma(T)} \leq \sigma, \quad \frac{h}{\rho(T)} \leq \rho \quad \forall \ T \in \mathcal{T}_h, \ h > 0.
\]
Approximation by finite elements

Family of regular triangulations: \( \{ \mathcal{T}_h \}_{h>0} \) of \( \bar{\Omega} \):

Associate to all \( T \in \mathcal{T}_h \) the numbers \( \rho(T) \) (diameter of \( T \)) and \( \sigma(T) \) (diameter of the largest ball in \( T \)).

\[
h := \max_{T \in \mathcal{T}_h} \rho(T) \quad \text{(mesh size)}
\]

Regularity assumptions:

- \( \exists \rho > 0, \sigma > 0 \) such that

\[
\frac{\rho(T)}{\sigma(T)} \leq \sigma, \quad \frac{h}{\rho(T)} \leq \rho \quad \forall \ T \in \mathcal{T}_h, \ h > 0.
\]

- Define \( \bar{\Omega}_h = \bigcup_{T \in \mathcal{T}_h} T \) with interior \( \Omega_h \) and boundary \( \Gamma_h \).
Approximation by finite elements

Family of regular triangulations: \( \{ T_h \}_{h>0} \) of \( \bar{\Omega} \):

Associate to all \( T \in T_h \) the numbers \( \rho(T) \) (diameter of \( T \)) and \( \sigma(T) \) (diameter of the largest ball in \( T \)).

\[
h := \max_{T \in T_h} \rho(T) \quad \text{(mesh size)}
\]

Regularity assumptions:

- \( \exists \rho > 0, \sigma > 0 \) such that

\[
\frac{\rho(T)}{\sigma(T)} \leq \sigma, \quad \frac{h}{\rho(T)} \leq \rho \quad \forall \ T \in T_h, \ h > 0.
\]

- Define \( \bar{\Omega}_h = \bigcup_{T \in T_h} T \) with interior \( \Omega_h \) and boundary \( \Gamma_h \).
  Assume that \( \bar{\Omega}_h \) is convex and that the vertices of \( T_h \) placed on the boundary \( \Gamma_h \) are points of \( \Gamma \).
Finite element approximation

**Assumption:** $\Omega \subset \mathbb{R}^n$ is open, convex and bounded $n \in \{2, 3\}$, with boundary $\Gamma$ of class $C^{1,1}$. For $n = 2$, $\Omega$ is allowed to be polygonal instead of of class $C^{1,1}$. 
**Assumption:** \( \Omega \subset \mathbb{R}^n \) is open, convex and bounded \( n \in \{2, 3\} \), with boundary \( \Gamma \) of class \( C^{1,1} \). For \( n = 2 \), \( \Omega \) is allowed to be polygonal instead of class \( C^{1,1} \).

Then, with some \( C > 0 \).

\[
|\Omega \setminus \Omega_h| \leq C h^2.
\]
Finite element approximation

**Assumption:** $\Omega \subset \mathbb{R}^n$ is open, convex and bounded $n \in \{2, 3\}$, with boundary $\Gamma$ of class $C^{1,1}$. For $n = 2$, $\Omega$ is allowed to be polygonal instead of of class $C^{1,1}$. Then, with some $C > 0$.

$$|\Omega \setminus \Omega_h| \leq Ch^2.$$ 

Piecewise linear approximation of the states:

$$Y_h = \{ y_h \in C(\bar{\Omega}) \mid y_h|_T \in P_1, \text{ for all } T \in \mathcal{T}_h, \text{ and } y_h = 0 \text{ on } \bar{\Omega} \setminus \Omega_h \}.$$
Finite element approximation

Assumption: $\Omega \subset \mathbb{R}^n$ is open, convex and bounded $n \in \{2, 3\}$, with boundary $\Gamma$ of class $C^{1,1}$. For $n = 2$, $\Omega$ is allowed to be polygonal instead of class $C^{1,1}$.

Then, with some $C > 0$.

$$|\Omega \setminus \Omega_h| \leq C h^2.$$ 

Piecewise linear approximation of the states:

$$Y_h = \{y_h \in C(\bar{\Omega}) \mid y_h|_T \in P_1, \text{ for all } T \in \mathcal{T}_h, \text{ and } y_h = 0 \text{ on } \bar{\Omega} \setminus \Omega_h \}.$$ 

Discretized state equation

$$\begin{cases}
\text{Find } y_h \in Y_h \text{ such that, for all } z_h \in Y_h, \\
\int_{\Omega_h} [a(x, y_h(x)) \nabla y_h \cdot \nabla z_h + d(y_h(x)) z_h] \, dx = \int_{\Omega_h} uz_h \, dx.
\end{cases}$$
Local uniqueness of discretized states

By the Brouwer fixed point theorem, the existence of solutions $y_h$ to the discretized equation can be shown.

We did not assume (global) boundedness of $a(x, y)$. To our surprise, we were not able to show uniqueness in this case. If $a$ is bounded, then the uniqueness can be shown for all sufficiently small $h > 0$.

Therefore, in the unbounded case, we had to work with local uniqueness of $y_h$ as in the setting of the implicit function theorem.
By the Brouwer fixed point theorem, the existence of solutions \( y_h \) to the discretized equation can be shown.

We did not assume (global) boundedness of \( a(x, y) \). To our surprise, we were not able to show uniqueness in this case. If \( a \) is bounded, then the uniqueness can be shown for all sufficiently small \( h > 0 \).

Therefore, in the unbounded case, we had to work with local uniqueness of \( y_h \) as in the setting of the implicit function theorem.

Assume for simplicity boundedness of \( a \) and that \( h \) is sufficiently small so that the mapping \( u \mapsto y_h(u) \) is well defined:

**Definition:** For given \( u \in U_{ad} \), \( y_h(u) \) is the solution to the discretized equation.
Discretized optimal control problem

Under the same simplification as above, we define

\[ f_h(u) = \frac{1}{2} \int_{\Omega_h} (y_h(u) - y_d)^2 \, dx + \frac{\lambda}{2} \int_{\Omega_h} u^2 \, dx. \]
Discretized optimal control problem

Under the same simplification as above, we define

\[ f_h(u) = \frac{1}{2} \int_{\Omega_h} (y_h(u) - y_d)^2 \, dx + \frac{\lambda}{2} \int_{\Omega_h} u^2 \, dx. \]

Set of discretized control functions: \( U_{ad}^h \subset U_{ad} \)
Discretized optimal control problem

Under the same simplification as above, we define

\[ f_h(u) = \frac{1}{2} \int_{\Omega_h} (y_h(u) - y_d)^2 \, dx + \frac{\lambda}{2} \int_{\Omega_h} u^2 \, dx. \]

Set of discretized control functions: \( U_h^{ad} \subset U_{ad} \)

\((P_h)\quad \min f_h(u_h), \quad u_h \in U_h^{ad}.\)
Discretized optimal control problem

Under the same simplification as above, we define

\[ f_h(u) = \frac{1}{2} \int_{\Omega_h} (y_h(u) - y_d)^2 \, dx + \frac{\lambda}{2} \int_{\Omega_h} u^2 \, dx. \]

Set of discretized control functions: \( U_{ad}^h \subset U_{ad} \)

\[
(P_h) \quad \min f_h(u_h), \quad u_h \in U_{ad}^h.
\]

We considered the following sets \( U_{ad}^h \):

- \( U_{ad}^h = U_{ad} \quad \forall h > 0 \) (variational discretization)
- All piecewise constant functions on \( \Omega_h \) (constant on each triangle) with values in \([\alpha, \beta]\)
- All piecewise linear functions on \( \Omega_h \) with values in \([\alpha, \beta]\).
Let a locally optimal control $\bar{u}$ of $(P)$ satisfy the second-order sufficient conditions introduced above and let $U^h_{ad}$ be defined by piecewise constant functions. Assume that $\bar{u}_h$ is a sequence of locally optimal (piecewise constant) solutions to $(P_h)$ that converges strongly in $L^2(\Omega)$ to $\bar{u}$. Then there is some constant $C > 0$ not depending on $h$ such that

$$\| \bar{u}_h - \bar{u}\|_{L^2(\Omega_h)} \leq C h \quad \forall h > 0.$$
Theorem (Piecewise constant controls, $L^2$-estimate)

Let a locally optimal control $\bar{u}$ of (P) satisfy the second-order sufficient conditions introduced above and let $U^h_{ad}$ be defined by piecewise constant functions. Assume that $\bar{u}_h$ is a sequence of locally optimal (piecewise constant) solutions to $(P_h)$ that converges strongly in $L^2(\Omega)$ to $\bar{u}$. Then there is some constant $C > 0$ not depending on $h$ such that

$$
\|\bar{u}_h - \bar{u}\|_{L^2(\Omega_h)} \leq C h \quad \forall h > 0.
$$

Survey of other results:

- Same estimate in the $L^\infty$-norm for piecewise constant controls
- Order $h^2$ for variational discretization ($L^2$ and $L^\infty$)
- $\lim_{h \to 0} h^{-1} \|\bar{u}_h - \bar{u}\|_{L^2(\Omega_h)} = 0$ for piecewise linear controls
- $L^2$-estimate of order $h^{3/2}$ for piecewise linear controls under some standard structural assumption on the triangles, where the reduced gradient vanishes on a positive measure.
General tool for error estimates

To simplify the derivation of error estimates, we proved a general theorem on error estimates that is formulated below for our concrete setting.

In our problem, we have a sequence $\varepsilon_h \to 0$ such that 

$$|f'(u) - f'(\bar{u})|v \leq \varepsilon_h \|v\|_{L^2(\Omega)}$$

for all $(u, v) \in U_{ad} \times L^2(\Omega)$ with $v = u_h - \bar{u}$ with $u_h \in U_h^{ad}$.

Theorem

Let $\{\bar{u}_h\}_{h>0}$ be a sequence of local solutions to $(P_h)$ converging strongly to $\bar{u}$ in $L^2(\Omega)$. Under the second-order sufficiency condition, there exist $C>0$ and $h_0>0$ such that 

$$\|\bar{u} - \bar{u}_h\|_{L^2(\Omega)} \leq C[\varepsilon^2_h + \|\bar{u} - u_h\|_{L^2(\Omega)}^2 + f'(\bar{u})(u_h - \bar{u})]^{1/2} \forall u_h \in U_h^{ad}, \forall h < h_0.$$

General tool for error estimates

To simplify the derivation of error estimates, we proved a general theorem on error estimates that is formulated below for our concrete setting.

\[ |f'(u) - f'(u_h)| \leq \varepsilon h \|v\|_{L^2(\Omega)} \]

for all \((u, v) \in U_{ad} \times L^2(\Omega)\) with \(v = u_h - \bar{u}\) with \(u_h \in U_{h_{ad}}\).

**Theorem**

Let \(\{\bar{u}_h\}_{h>0}\) be a sequence of local solutions to \((P_h)\) converging strongly to \(\bar{u}\) in \(L^2(\Omega)\). Under the second-order sufficiency condition, there exist \(C>0\) and \(h_0>0\) such that

\[ \|\bar{u} - u_h\|_{L^2(\Omega)} \leq C \left[ \varepsilon^2 h + \|\bar{u} - u_h\|_{L^2(\Omega)} + f'_{\bar{u}}(u_h - \bar{u}) \right]^{1/2} \]

\(\forall u_h \in U_{h_{ad}}, \forall h < h_0\).

To simplify the derivation of error estimates, we proved a general theorem on error estimates that is formulated below for our concrete setting. In our problem, we have a sequence $\varepsilon_h \to 0$ such that

$$|[f_h'(u) - f'(u)]v| \leq \varepsilon_h \|v\|_{L^2(\Omega)}$$

for all $(u, v) \in U_{ad} \times L^2(\Omega)$ with $v = u_h - \bar{u}$ with $u_h \in U_{ad}^h$. 

General tool for error estimates

To simplify the derivation of error estimates, we proved a general theorem on error estimates that is formulated below for our concrete setting. In our problem, we have a sequence $\varepsilon_h \to 0$ such that

$$||[f'_h(u) - f'(u)]v|| \leq \varepsilon_h \|v\|_{L^2(\Omega)}$$

for all $(u, v) \in U_{ad} \times L^2(\Omega)$ with $v = u_h - \bar{u}$ with $u_h \in U_{ad}^h$.

**Theorem**

Let $\{\bar{u}_h\}_{h>0}$ be a sequence of local solutions to $(P_h)$ converging strongly to $\bar{u}$ in $L^2(\Omega)$. Under the second-order sufficiency condition, there exist $C > 0$ and $h_0 > 0$ such that

$$\|\bar{u} - \bar{u}_h\|_{L^2(\Omega)} \leq C \left[\varepsilon_h^2 + \|\bar{u} - u_h\|^2_{L^2(\Omega)} + f'(\bar{u})(u_h - \bar{u})\right]^{1/2} \quad \forall u_h \in U_{ad}^h, \forall h < h_0.$$
To simplify the derivation of error estimates, we proved a general theorem on error estimates that is formulated below for our concrete setting. In our problem, we have a sequence $\varepsilon_h \to 0$ such that

$$\left| [f'_h(u) - f'(u)]v \right| \leq \varepsilon_h \|v\|_{L^2(\Omega)}$$

for all $(u, v) \in U_{ad} \times L^2(\Omega)$ with $v = u_h - \bar{u}$ with $u_h \in U^h_{ad}$.

**Theorem**

Let $\{\bar{u}_h\}_{h>0}$ be a sequence of local solutions to $(P_h)$ converging strongly to $\bar{u}$ in $L^2(\Omega)$. Under the second-order sufficiency condition, there exist $C > 0$ and $h_0 > 0$ such that

$$\|\bar{u} - \bar{u}_h\|_{L^2(\Omega)} \leq C \left[ \varepsilon_h^2 + \|\bar{u} - u_h\|_{L^2(\Omega)}^2 + f'(\bar{u})(u_h - \bar{u}) \right]^{1/2} \forall u_h \in U^h_{ad}, \forall h < h_0.$$