Online Adjoint Methods for Optimization of PDEs

Justin Sirignano[1] · Konstantinos Spiliopoulos[2]

Abstract

We present and mathematically analyze an online adjoint algorithm for the optimization of partial differential equations (PDEs). Traditional adjoint algorithms would typically solve a new adjoint PDE at each optimization iteration, which can be computationally costly. In contrast, an online adjoint algorithm updates the design variables in continuous-time and thus constantly makes progress towards minimizing the objective function. The online adjoint algorithm we consider is similar in spirit to the the pseudo-time-stepping, one-shot method which has been previously proposed. Motivated by the application of such methods to engineering problems, we mathematically study the convergence of the online adjoint algorithm. The online adjoint algorithm relies upon a time-relaxed adjoint PDE which provides an estimate of the direction of steepest descent. The algorithm updates this estimate continuously in time, and it asymptotically converges to the exact direction of steepest descent as $t \rightarrow \infty$. We rigorously prove that the online adjoint algorithm converges to a critical point of the objective function for optimizing the PDE. Under appropriate technical conditions, we also prove a convergence rate for the algorithm. A crucial step in the convergence proof is a multi-scale analysis of the coupled system for the forward PDE, adjoint PDE, and the gradient descent ODE for the design variables.

Keywords Online algorithm · Adjoint algorithm · Optimization · Partial differential equations (PDEs)

✉ Konstantinos Spiliopoulos
  kspiliop@math.bu.edu

  Justin Sirignano
  Justin.Sirignano@maths.ox.ac.uk

[1] Mathematical Institute, University of Oxford, Oxford, UK

[2] Department of Mathematics and Statistics, Boston University, Boston, USA

⌂ Springer

1 Introduction

Adjoint methods have been widely-used for the optimization of partial differential equations (PDEs), and especially for optimizing PDEs modeling engineering systems. Examples include [2, 4–6, 9–11, 16–18, 20, 24, 26–28], and [15].

Traditional adjoint algorithms consist of an iteration where at each iteration a new adjoint PDE must be solved to calculate the gradient descent step. During the course of optimization, many adjoint PDEs must be solved, which in certain cases can be computationally costly. As an alternative, time-stepping and pseudo-time-stepping methods (often in combination with one-shot methods) have been proposed, where one views time-independent PDEs as stationary states of appropriate dynamical systems and studies the behavior of the latter in the long-time regime, i.e., after their transient phase. We refer the interested reader to [3, 7, 12–14, 19, 31, 32] and the references therein for certainly a non-exhaustive list of representative references.

In this paper, we couple a time-relaxed adjoint PDE with a continuous-time update equation for the variables that are being optimized. The time-relaxed adjoint PDE yields an estimate of the direction of steepest descent, and updates this estimate *continuously* in time. The optimization variables are also updated continuously in time using this online estimate of the direction of steepest descent. The focus of our paper is the mathematical analysis of this "online adjoint algorithm". As $t \to \infty$, the solution of the time-relaxed adjoint PDE asymptotically matches the exact direction of steepest descent. A crucial step in the convergence proof is a multi-scale analysis of the coupled system for the forward PDE, adjoint PDE, and the gradient descent ODE for the design variables.

We prove convergence and convergence rates for the online adjoint algorithm for a certain class of PDEs. Specifically, in our theoretical analysis, we consider the optimization problem where we seek to minimize the objective function:

$$J(\theta) = \frac{1}{2} \int_U \left(u^*(x) - h(x) \right)^2 dx + \frac{\gamma}{2} \|\theta\|_2^2, \tag{1.1}$$

where h is a target profile. $\gamma \|\theta\|_2^2$ is a regularization term where $\gamma > 0$ and $\|\cdot\|_2$ is the ℓ_2 norm. u^* satisfies the elliptic PDE

$$\begin{aligned} Au^*(x) &= f(x, \theta), \quad x \in U \\ u^*(x) &= 0, \quad x \in \partial U, \end{aligned} \tag{1.2}$$

where A is a standard second-order elliptic operator. Thus, we wish to select a parameter θ such that the solution u^* of the PDE (1.2) is as close as possible to the target profile h.

If $A^\dagger$ denotes the formal adjoint operator to A, then the adjoint PDE is

$$\begin{aligned} A^\dagger \hat{u}^*(x) &= u^*(x) - h(x), \quad x \in U \\ \hat{u}^*(x) &= 0, \quad x \in \partial U \end{aligned} \tag{1.3}$$

The gradient of the objective function (1.1) can be evaluated using the solution $\hat{u}^*$ to the adjoint PDE (1.3). By Lemma 2.4 we have that

$$\nabla_\theta J(\theta) = \int_U \hat{u}^*(x)\nabla_\theta f(x,\theta)dx + \gamma\theta. \tag{1.4}$$

Thus, the adjoint PDE (1.3) can be used to evaluate the gradient of the objective function, which in turn can be used to optimize over the PDE (1.2). A key advantage of adjoint methods is that, no matter how large the dimension of θ is, the adjoint PDE (1.3) is the same dimension as the original PDE (1.2).

1.1 The Online Adjoint Algorithm

The online adjoint algorithm optimizes the objective function $J(\theta)$ via a continuous-time equation for the update of the parameter $\theta(t)$; see also [14, 32] for related formulations. The direction of steepest descent is estimated using a time-relaxation of the adjoint PDE. The estimate and the optimization variables are both simultaneously updated continuously in time. An appropriately chosen learning rate parameter is introduced, which allows to guarantee both well posedness of the algorithm for all times (Theorem 2.8) and convergence as $t \to \infty$ (Theorems 3.1 and 4.2).

The online adjoint algorithm satisfies the equations:

$$\begin{aligned}
\frac{\partial u}{\partial t}(t,x) &= -Au(t,x) + f(x,\theta(t)), \quad x \in U, t > 0 \\
\frac{\partial \hat{u}}{dt}(t,x) &= -A^\dagger\hat{u}(t,x) + (u(t,x) - h(x)), \quad x \in U, t > 0 \\
\frac{d\theta}{dt}(t) &= -\alpha(t)\left(\int_U \hat{u}(t,x)\nabla_\theta f(x,\theta(t))dx + \gamma\theta(t) \right) \\
u(t,x) &= \hat{u}(t,x) = 0, \quad x \in \partial U, t > 0 \\
u(0,x) &= u_0(x), \hat{u}(0,x) = \hat{u}_0(x),
\end{aligned} \tag{1.5}$$

where $\alpha(t)$ is an appropriately chosen learning rate. The PDEs for u and $\hat{u}$ can be viewed as time relaxations of the PDE (1.2) and its adjoint PDE (1.3). It is easy to see that $\int_U \hat{u}(t,x)\nabla_\theta f(x,\theta(t))dx$ is an estimate for the direction of steepest descent $\int_U \hat{u}^*(x)\nabla_\theta f(x,\theta(t))dx$.

Apart from the generic formulation of the online adjoint algorithm as presented in (1.5), our main contribution is two-fold. First, we prove that as $t \to \infty$, $\|\nabla J(\theta(t))\| \to 0$. Namely, we prove that $\theta(t)$ converges to a stationary point of $J(\theta)$. We emphasize here that in order to do so, no assumptions on convexity of J are needed. Secondly, if we further assume that $J(\theta)$ is strongly convex, then we also prove a convergence rate of $\theta(t)$ to the global minimum of $J(\theta)$.

In practice, the online adjoint algorithm (1.5) is implemented by simultaneously solving the coupled ODE-PDE system using numerical methods such as finite-difference methods. Either explicit or implicit finite difference methods can be used. For example, an explicit finite difference method for implementing (1.5) would be:

- Update u and $\hat{u}$:

$$u(t + \Delta, x) = u(t, x) + \left(- Au(t, x) + f(x, \theta(t)) \right)\Delta,$$

$$\hat{u}(t + \Delta, x) = \hat{u}(t, x) + \left(- A^{\dagger}\hat{u}(t, x) + (u(t, x) - h(x)) \right)\Delta, \quad (1.6)$$

where Δ is the time-step size.
- Then, update the parameter θ:

$$\theta(t + \Delta) = \theta(t) - \alpha(t)\left(\int_U \hat{u}(t, x)\nabla_\theta f(x, \theta(t))dx + \gamma\theta(t) \right)\Delta. \quad (1.7)$$

The spatial domain x is discretized and a finite-difference method is used to approximate the operator A. The integrals are discretized as appropriate sums.

The focus of our paper is to rigorously prove the convergence of the online adjoint algorithm for linear elliptic PDEs. In practice, real-world applications will typically require optimizing over nonlinear PDEs. The online adjoint algorithm can also be used to optimize over nonlinear PDEs. References [30] and [23] optimize over the Navier–Stokes equation using our online adjoint algorithm. Numerical optimization with pseudo-time-stepping adjoint methods has also been studied in [3, 7, 12–14, 19, 31, 32].

We demonstrate the online adjoint method below for a simple example of a nonlinear PDE. Consider the equation

$$0 = -\theta_1 u \frac{\partial u}{\partial x} - \theta_2 u \frac{\partial u}{\partial y} + \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}, \quad (1.8)$$

where $(x, y) \in [0, 1] \times [0, 1]$ and with boundary conditions $u(0, y) = 1$, $u(1, y) = -1$, $u(x, 0) = 1$, and $u(x, 1) = -1$. The parameters to be optimized over are $\theta = (\theta_1, \theta_2)$ and the objective function is (1.1) with $\gamma = 0$. The target function h is the solution to (1.8) with $\theta = (10, 10)$. That is, our goal is to solve the inverse problem of recovering the parameters in the PDE (1.8) given an observed solution.

The adjoint PDE for (1.8) is

$$0 = \theta_1 u \frac{\partial \hat{u}}{\partial x} + \theta_2 u \frac{\partial \hat{u}}{\partial y} + \frac{\partial^2 \hat{u}}{\partial x^2} + \frac{\partial^2 \hat{u}}{\partial y^2} + (u - h), \quad (1.9)$$

where $(x, y) \in [0, 1] \times [0, 1]$ and with boundary conditions $\hat{u}(0, y) = \hat{u}(1, y) = \hat{u}(x, 0) = \hat{u}(x, 1) = 0$. The gradient of the objective function is given by the formula

$$\frac{\partial J(\theta)}{\partial \theta_1} = -\int_0^1 \int_0^1 \hat{u}u \frac{\partial u}{\partial x}(x, y)dxdy,$$

$$\frac{\partial J(\theta)}{\partial \theta_2} = -\int_0^1 \int_0^1 \hat{u}u \frac{\partial u}{\partial y}(x, y)dxdy.$$

The online adjoint algorithm can be used to minimize the objective function $J(\theta)$. In our numerical experiment, we use an explicit finite difference method for the numerical solution of the time-relaxed PDE and the time-relaxed adjoint PDE. The PDE variables are updated using an Euler scheme. (Although not implemented here, it is worthwhile noting that higher-order accuracy in time could be achieved with a Runge–Kutta scheme.) Uniform mesh sizes for both time and space are chosen. The PDE operator is approximated using a second-order accurate finite-difference method. The parameter ODEs are also solved using an explicit Euler scheme on the same uniform time grid and the spatial integrals are also discretized as sums using the uniform spatial grid. Figure 1 demonstrates that the online adjoint algorithm converges to the correct value for the parameters θ as $t \rightarrow \infty$. The right display in Fig. 1 presents the numerical convergence rate, which satisfies the theoretical convergence rate of $t^{-\frac{1}{2}}$ which we prove for strongly convex objective functions for linear elliptic PDEs in this paper. (In fact, for this specific example, the numerical convergence rate turns out to be faster than $t^{-\frac{1}{2}}$.)

1.2 Organization of the Proof

The rest of the paper is organized as follows. In Sect. 2 we state our assumptions, present in more details the online adjoint algorithm and prove its well posedness in Theorem 2.8. Convergence of $\theta(t)$ to a stationary point of $J(\theta)$ is proven in Sect. 3, Theorem 3.1. We emphasize that no convexity requirements on $J(\theta)$ are needed in order to prove convergence. If in addition, we assume that $J(\theta)$ is strongly convex with a single stationary point, then one can prove a convergence rate, Theorem 4.2. The latter is the content of Sect. 4.
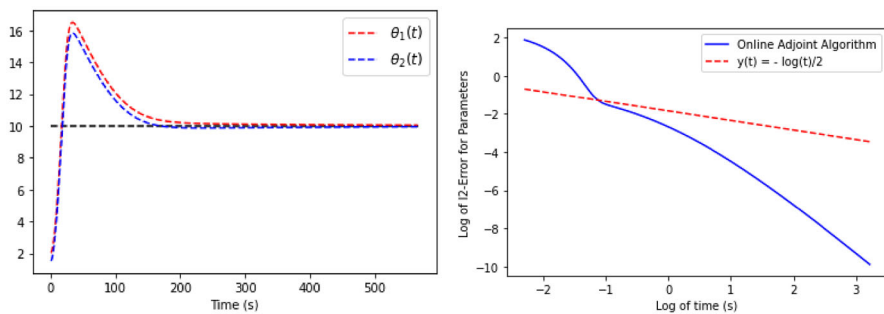
Fig. 1 Left: solution for θ using the online adjoint algorithm versus computational time. Right: numerical convergence rate for $\left\| \theta - \theta^* \right\|_2$ where $\theta^* = (10, 10)$

2 Assumptions, Notation and Well Posedness of the Online Adjoint Algorithm

Let U be an open, bounded subset of $\mathbb{R}^n$. We will denote by $(\cdot, \cdot)$ the usual inner product in $H = L^2(U)$. We shall assume that the operator A is uniformly elliptic, diagonalizable and dissipative, per Assumption 2.1.

Assumption 2.1 The operator A is uniformly elliptic. We also assume that A is diagonalizable and dissipative. Namely there exists a countable complete orthonormal basis $\{e_n\}_{n \in \mathbb{N}} \subset H$ that consists of eigenvectors of A corresponding to a non-negative sequence $\{\lambda_n\}_{n \in \mathbb{N}}$ of eigenvalues such that

$$(-A)e_n = -\lambda_n e_n, \quad n \in \mathbb{N}$$

such that the dissipativity condition $\lambda = \inf_{n \in \mathbb{N}} \lambda_n > 0$ holds.

An example of A is the second order elliptic operator, for definiteness taken to be in divergence form,

$$Au(x) = -\sum_{i,j=1}^{n} \left(a^{i,j}(x) u_{x_i}(x) \right)_{x_j} + \sum_{i=1}^{n} b^i(x) u_{x_i}(x) + c(x)u(x).$$

The diagonalizable condition of Assumption 2.1 is automatically satisfied for example by self-adjoint operators, see [8,Theorem 8.8.37]

Before proceeding with the well-posedness of the online adjoint algorithm, let us recall a few basic results that will be useful for the analysis that follows.

Taking the domain of A to be $D(A) = H_0^1(U) \cap H^2(U)$, we have that it is dense in $H = L^2(U)$. Then, due to Assumption 2.1, elliptic regularity theory gives that the operator A is closed and thus by Hille-Yoshida theorem $(-A)$ is the generator of an analytic strongly contraction semigroup $\{S(t)\}_{t \geq 0}$ on H. The spectral assumption made in Assumption 2.1 guarantees that

$$\|S(t)u\|_H \leq e^{-\lambda t} \|u\|_H. \tag{2.1}$$

The latter also means that A is a coercive operator. In particular, we will heavily use the fact that for $u \in D(A)$, we have

$$(Au, u) \geq \lambda \|u\|_H^2. \tag{2.2}$$

Notice now that because we are dealing with a real Hilbert space and because $H_0^1(U) \cap H^2(U)$ is dense in $L^2(U)$, we obtain that $A^\dagger$ is also a coercive operator. Indeed, by definition of the adjoint operator $A^\dagger$ we shall have that for $u \in H_0^1(U) \cap H^2(U)$

$$(A^\dagger u, u) = (u, Au) \geq \lambda(u, u). \tag{2.3}$$

Notice now that under our assumptions the adjoint operator $(-A^\dagger)$ will also generated an analytic strongly continuous semigroup $\{S^\dagger(t)\}_{t\geq 0}$ on $L^2(U)$. In particular (2.3) implies that the adjoint semigroup $S^\dagger(t)$ will also be exponential stable. Indeed, by definition we have for $u \in H_0^1(U) \cap H^2(U)$

$$\frac{d}{dt} \left\| S^\dagger(t)u \right\|_H^2 = -2(A^\dagger S^\dagger(t)u, S^\dagger(t)u) \leq -2\lambda \left\| S^\dagger(t)u \right\|_H^2, \qquad (2.4)$$

which then due to Gronwall lemma gives

$$\left\| S^\dagger(t)u \right\|_H \leq e^{-\lambda t} \|u\|_H, \qquad (2.5)$$

proving the exponential stability of $S^\dagger(t)$.

Then, if we assume that $f \in L^2(U)$, classical Lax–Milgram theorem (see for example Chap. 5.8 of [8]) says that the elliptic boundary-value problem

$$\begin{aligned} Au^*(x) &= f(x), \quad x \in U \\ u^*(x) &= 0, \quad x \in \partial U \end{aligned} \qquad (2.6)$$

has a unique weak solution $u^* \in H_0^1(U)$. The same conclusion will also be true for the adjoint problem governed by the adjoint operator $A^\dagger$.

Remark 2.2 We also recall here that by classical elliptic regularity results if for given $m \in \mathbb{N}$, $a^{i,j}, b^i, c \in C^{m+1}(\bar{U})$ with $i, j = 1, \ldots, n$ and $\partial U \in C^{m+2}$, then the unique solution u^* to (2.6) is such that $u^* \in H^{m+2}(U)$. Clearly if $a^{i,j}, b^i, c \in C^\infty(\bar{U})$ with $i, j = 1, \ldots, n$ and $\partial U \in C^\infty$, then $u^* \in C^\infty(\bar{U})$. We refer the interested reader to classical manuscripts, e.g., [8,Chap. 8], for more details.

Remark 2.3 For notational convenience and without loss of generality, we have assumed zero boundary conditions for the PDE (2.6). We can consider the PDE (2.6) with non-zero boundary data, say $u^*(x) = g(x), x \in \partial U$, under the assumption $g \in H^1(U)$ for unique solvability of the corresponding PDE (2.6). We refer the interested reader to classical manuscripts, e.g., [8,Chap. 8], for more details.

As briefly presented in the introduction now, let $f(\cdot, \cdot) : \mathbb{R}^n \times \mathbb{R}^d \mapsto \mathbb{R}$ be such that for every $\theta \in \mathbb{R}^d$ $f(\cdot, \theta) \in L^2(U)$. As with (2.6) the linear PDE

$$\begin{aligned} Au^*(x) &= f(x, \theta), \quad x \in U \\ u^*(x) &= 0, \quad x \in \partial U \end{aligned} \qquad (2.7)$$

will have, for each given $\theta \in \mathbb{R}^d$, a unique weak solution $u^* \in H_0^1(U)$. We shall write $u^*(x; \theta)$ when we want to emphasize the dependence on θ.

For a given target profile $h \in L^2(U)$, the goal is to select θ to minimize the objective function

$$J(\theta) = \frac{1}{2}(u^* - h, u^* - h) + \frac{\gamma}{2} \|\theta\|_2^2, \qquad (2.8)$$

where $\gamma > 0$. $\gamma \|\theta\|_2^2$ is a regularization term and $\|\cdot\|_2$ is the ℓ_2 norm.

The adjoint PDE satisfies

$$A^\dagger \hat{u}^*(x) = u^*(x) - h(x), \quad x \in U$$
$$\hat{u}^*(x) = 0, \quad x \in \partial U \tag{2.9}$$

and as with (2.6), Assumption 2.1 and the fact that $u^* - h \in L^2(U)$ guarantee that (2.9) has a unique weak solution $\hat{u}^* \in H_0^1(U)$. Notice that since u^* depends on θ, the same will also be true for $\hat{u}^*$ and we shall write $\hat{u}^*(x; \theta)$ when we want to emphasize that. The following lemma provides a useful representation for $\nabla_\theta J(\theta)$, which will also motivate the form of the online adjoint algorithm.

Lemma 2.4 *Let Assumption 2.1 and assume that $h \in L^2(U)$. Then, we can write*

$$\nabla_\theta J(\theta) = (\hat{u}^*, \nabla_\theta f(\theta)) + \gamma \theta. \tag{2.10}$$

The proof of Lemma 2.4 is presented at the end of this section. In terms of the learning rate $\alpha(t)$ we make the following assumption.

Assumption 2.5 We assume that the learning rate $\alpha(t)$ is such that $\lim\limits_{t \to \infty} \alpha(t) = 0$ and

- $\int_0^\infty \alpha(s)ds = \infty$ and $\int_0^\infty \alpha^2(s)ds < \infty$.
- $\sup\limits_{t \geq 0} \int_0^t \alpha(s)e^{-\gamma \int_s^t \alpha(r)dr}ds < \infty$ and $\lim\limits_{t \to \infty} \dfrac{\alpha'(t)}{\alpha(t)} = 0$.

The first part of Assumption 2.5 on the learning rate is classical and is also the same used in discrete time algorithms, see for example classical references such as [1, 21]. The second part of Assumption 2.5 comes up while proving that $\|\theta(t)\|$ stays bounded for all times and later on in the convergence proof of $\theta(t)$ to a stationary point of J. An example of a learning rate that satisfies both parts of Assumption 2.5 is $\alpha(t) = \frac{1}{1+t}$.

In terms of the parametric model $f(\cdot, \cdot)$ we make the following assumption

Assumption 2.6 We assume the following conditions:

- For each fixed $\theta \in \mathbb{R}^d$, $f(\cdot, \theta)$, $\nabla_\theta f(\cdot, \theta)$ and $\nabla_\theta^2 f(\cdot, \theta)$ are in $L^2(U)$. For each fixed $x \in \mathbb{R}^n$, $f(x, \cdot)$, $\nabla_\theta f(x, \cdot)$ and $\nabla_\theta^2 f(x, \cdot)$ are bounded. In other words we assume that there exists $C < \infty$ such that

$$\sup_{\theta \in \mathbb{R}^d} \left(\|f(\theta)\|_{L^2(U)} + \|\nabla_\theta f(\theta)\|_{L^2(U)} + \left\|\nabla_\theta^2 f(\theta)\right\|_{L^2(U)} \right) \leq C.$$

- For each $x \in U$, $f(x, \cdot)$ is globally Lipschitz L^2 Lipschitz constant in x.

In terms of the associated cost function $J(\cdot)$ we make the following Assumption 2.7.

Assumption 2.7 We assume that $J(\cdot) \in C^2$ and globally Lipschitz.

We emphasize that Assumption 2.7 does not impose any convexity type of assumptions on $J(\theta)$. The online adjoint algorithm satisfies the time-dependent PDEs

$$
\begin{aligned}
\frac{\partial u}{\partial t}(t, x) &= -Au(t, x) + f(x, \theta(t)), \quad x \in U, t > 0 \\
\frac{\partial \hat{u}}{dt}(t, x) &= -A^{\dagger}\hat{u}(t, x) + (u(t, x) - h(x)), \quad x \in U, t > 0 \\
\frac{d\theta}{dt}(t) &= -\alpha(t)\left((\nabla_\theta f(\theta(t)), \hat{u}(t)) + \gamma\theta(t)\right) \\
u(t, x) &= \hat{u}(t, x) = 0, \quad x \in \partial U, t > 0 \\
u(0, x) &= u_0(x), \hat{u}(0, x) = \hat{u}_0(x)
\end{aligned}
\tag{2.11}
$$

Notice that (2.11) is a non-local coupled system of PDEs. Theorem 2.8 is about the well-posedness of system (2.11). Also, with slight abuse of notation, with $\theta = \theta(t)$ from (2.11) we shall denote the solutions to (2.7) and (2.9), by $u^*(t, x)$ and $\hat{u}^*(t, x)$ respectively.

Theorem 2.8 *Assume Assumptions 2.1, 2.5 and 2.6 and that $u_0, \hat{u}_0, h \in L^2(U)$. There exists a unique mild solution $u, \hat{u} \in \mathcal{C}((0, \infty); W_0^{2,2}(U)) \cap \mathcal{C}^1((0, \infty); L^2(U))$ and $\theta \in \mathcal{C}^1((0, \infty))$ to equation (2.11). If in addition $h \in L^\infty$, then $u, \hat{u} \in \mathcal{C}((0, \infty); W_0^{2,p}(U)) \cap \mathcal{C}^1((0, \infty); L^p(U))$ for any $p \geq 2$ and if further we assume that $u_0, \hat{u}_0 \in W^{2,p}(U)$, then $u, \hat{u} \in \mathcal{C}([0, \infty); W_0^{2,p}(U)) \cap \mathcal{C}^1([0, \infty); L^p(U))$ for any $p \geq 2$. In addition, we have that there exists some constant $K < \infty$ such that*

$$
\sup_{t \geq 0}\left[\|u(t)\|_{L^2(U)} + \|\hat{u}(t)\|_{L^2(U)} + \|\theta(t)\|_2\right] < K.
\tag{2.12}
$$

Remark 2.9 At this point we mention that even though in Assumption 2.6 we have assumed that $\|f(\theta)\|_{L^2(U)}$ is uniformly bounded, an investigation of the proof of Theorem 2.8 shows that this assumption can be relaxed. In particular, at the expense of slightly more elaborate estimates, the results of this paper (which heavily rely on (2.12) being true) hold if we assume instead that $\|f(\theta)\|_{L^2(U)}$ grows linearly in $\|\theta\|_{\ell_2}$ with bounded derivatives, i.e., $\|f(\theta)\|_{L^2} \leq C(1 + \|\theta\|_{\ell_2})$, still with $\left(\|\nabla_\theta f(\theta)\|_{L^2} + \|\nabla_\theta^2 f(\theta)\|_{L^2(U)}\right) < C$ and additionally that the regularization coefficient $\gamma > 0$ is large enough depending on the L^2 norms of $u_0(x)$ and $\hat{u}_0(x)$. We have chosen to present the results for uniformly bounded $\|f(\theta)\|_{L^2(U)}$ for presentation purposes and because in this case we do not need any additional restriction on the magnitude of γ other than being strictly positive.

Let us conclude this section with the proofs of Lemma 2.4 and Theorem 2.8.

Proof of Lemma 2.4 (2.10) can be derived using the definition of the adjoint PDE (2.9). Define $\tilde{u} = \nabla_\theta u^*$. Differentiating (2.7) yields

$$A\tilde{u}(x) = \nabla_\theta f(x, \theta), \quad x \in U$$
$$\tilde{u}(x) = 0, \quad x \in \partial U$$

Integration by parts yields

$$(\hat{u}^*, A\tilde{u}) = (A^\dagger \hat{u}^*, \tilde{u}).$$

Due to (2.7), this yields the equation

$$(A^\dagger \hat{u}^*, \tilde{u}) = (\hat{u}^*, \nabla_\theta f(\theta)).$$

Using the definition of the adjoint PDE (2.9),

$$(u^* - h, \tilde{u}) = (\hat{u}^*, \nabla_\theta f(\theta)).$$

Recalling the objective function (2.8), we then write

$$\nabla_\theta J(\theta) = (u^* - h, \tilde{u}) + \gamma\theta$$
$$= (\hat{u}^*, \nabla_\theta f(\theta)) + \gamma\theta,$$

which yields (2.10). $\square$

Proof of Theorem 2.8 Let us define the index set $\mathcal{G} = \{1, 2, 3\}$ and the space $\Theta = U \times \mathcal{G}$. Define the variable $y = (x, \zeta) \in \Theta$ and the measure $dn = dx \otimes d\iota$ on Θ where $d\iota$ denotes the counting measure on $\mathcal{G}$. Define now the Banach space $X^2 = L^2(\Theta, dn)$. Similarly we denote by $H^2(\Theta) = W^{2,2}(\Theta)$ the Banach space of functions f on Θ such that for each $\zeta \in \mathcal{G}$ we have $f(\cdot, \zeta) \in H^2(U)$ with norm $\|f\|_{H^2(\Theta)} = \sum_{\zeta=1}^{3} \|f(\cdot, \zeta)\|_{H^2(U)}$.

Setting $v(t, x) = (u(t, x), \hat{u}(t, x), \theta(t))$ and $\rho(t, y) = v_\zeta(t, x)$ (the ζ'th component of the vector v) for $y = (x, \zeta) \in U \times \mathcal{G}$ we consider the evolution equation on Θ given by

$$\partial_t \rho(t, y) = \mathcal{L}[\rho](t, y) + \mathcal{R}[\rho](t, y), y \in \Theta \tag{2.13}$$

where

$$\mathcal{L}[\rho](t, x, 1) = -Au(t, x), \quad \mathcal{L}[\rho](t, x, 2) = -A^\dagger \hat{u}(t, x), \quad \mathcal{L}[\rho](t, x, 3) = 0$$

and

$$\mathcal{R}[\rho](t, x, 1) = f(x, \theta(t)), \quad \mathcal{R}[\rho](t, x, 2) = u(t, x) - h(x),$$
$$\mathcal{R}[\rho](t, x, 3) = -\alpha(t)\left((\nabla_\theta f(\theta(t)), \hat{u}(t)) + \gamma\theta(t)\right).$$

We note that we have slightly abused notation here because $\mathcal{L}[\rho](t, x, 3) = 0$. However, this notation is convenient because it allows us to describe the PDE in question in the form (2.13) as a single vector valued evolution equation.

Let us define the norm $\|w\|_{2,T} = \sup_{t \in [0,T]} \|w(t)\|_{L^2(U)}$ if $w = w(t, x)$ and $\|w\|_{2,T} = \sup_{t \in [0,T]} \|w(t)\|_{\ell_2}$ if $w = w(t)$. Here ℓ_2 denotes the standard Euclidean norm. In particular, for $v(t, x) = (u(t, x), \hat{u}(t, x), \theta(t))$ we shall have

$$\|v\|_{2,T} = \|u\|_{2,T} + \|\hat{u}\|_{2,T} + \|\theta\|_{2,T}.$$

where, the first two components depend on (t, x) and the last component depends only on t.

We will be working with mild solutions. Due to Assumption 2.1, the operators A and $A^\dagger$ are generators of analytic contraction semigroups $\{S(t)\}_{t \geq 0}$ and $\{S^\dagger(t)\}_{t \geq 0}$ respectively on $L^2(U)$.

Then, we can write for the mild solution of (2.13) that

$$\rho(t, y) = H[\rho](t, y),$$

where

$$H[\rho](t, y) = \begin{cases} H[\rho](t, x, 1) \\ H[\rho](t, x, 2) \\ H[\rho](t, x, 3) \end{cases} = \begin{cases} S(t)u_0(x) + \int_0^t S(t - s)f(x, \theta(s))ds \\ S^\dagger(t)\hat{u}_0(x) + \int_0^t S^\dagger(t - s)(u(s, x) - h(x))ds \\ -\alpha(t)\Big((\nabla_\theta f(\theta(t)), \hat{u}(t)) + \gamma\theta(t)\Big). \end{cases}$$

(2.14)

Now the properties of the analytic contraction semigroups $\{S(t)\}_{t \geq 0}$ and $\{S^\dagger(t)\}_{t \geq 0}$ guarantee that there exist an increasing continuous function $D(t)$ with $\lim_{t \to 0} D(t) = 0$ (possible different from line to line below) such that

$$\|H[\rho](\cdot, 1)\|_{2,T} \leq \|u_0\|_{L^2} + D(T)$$
$$\|H[\rho](\cdot, 2)\|_{2,T} \leq \|\hat{u}_0\|_{L^2} + D(T)(1 + \|u\|_{2,T})$$
$$\|H[\rho](\cdot, 3)\|_{2,T} \leq \|\theta_0\|_{\ell_2} + D(T)(\|\hat{u}\|_{2,T} + \|\theta\|_{2,T})$$

(2.15)

and for $\rho = (u_1, \hat{u}_1, \theta_1)$ and $q = (u_2, \hat{u}_2, \theta_2)$

$$\|H[\rho](\cdot, 1) - H[q](\cdot, 1)\|_{2,T} \leq D(T)\|\theta_1 - \theta_2\|_{2,T}$$
$$\|H[\rho](\cdot, 2) - H[q](\cdot, 2)\|_{2,T} \leq D(T)\|u_1 - u_2\|_{2,T}$$
$$\|H[\rho](\cdot, 3) - H[q](\cdot, 3)\|_{2,T} \leq D(T)\Big(\|\hat{u}_1 - \hat{u}_2\|_{2,T} + \|\theta_1 - \theta_2\|_{2,T}(1 + \|\hat{u}_1\|_{2,T})\Big)$$ (2.16)

The linear growth bounds of the operator H as given by (2.15) together with the local Lipschitz continuity property demonstrated in (2.16) allow us to conclude via the classical Picard–Lindelöf theorem for Banach valued ODE's that for every $\rho_0 \in$

X^2 there exists a unique local mild solution $\rho \in \mathcal{C}([0, T_0], X^2)$ of (2.13) for some sufficiently small $0 < T_0 < \infty$.

Next we want to show that this solution can be extended globally. To do so it is enough to establish a global bound for the $\|\cdot\|_{X^2}$ of solutions. Indeed, the analytic contraction semigroups $\{S(t)\}_{t\geq 0}$ and $\{S^\dagger(t)\}_{t\geq 0}$ guarantee that there is a constant $K < \infty$ (independent of t) such that

$$
\begin{aligned}
\|u(t)\|_{L^2} &\leq \|S(t)u(0)\|_{L^2} + \int_0^t \|S(t-s)f(\theta(s))\|_{L^2}\, ds \\
&\leq \|u(0)\|_{L^2} + C \int_0^t e^{-\lambda(t-s)} ds \\
&\leq K.
\end{aligned}
\tag{2.17}
$$

Analogously, for a potentially different constant $K < \infty$ and using estimate (2.17)

$$
\begin{aligned}
\|\hat{u}(t)\|_{L^2} &\leq \left\| {}^\dagger S(t)\hat{u}(0) \right\|_{L^2} + \int_0^t \left\| S^\dagger(t-s)(u(s)-h) \right\|_{L^2} ds \\
&\leq \|\hat{u}(0)\|_{L^2} + \int_0^t e^{-\lambda(t-s)}(\|u(s)\|_{L^2} + \|h\|_{L^2}) ds \\
&\leq K.
\end{aligned}
\tag{2.18}
$$

Let us next show that $\theta(t)$ is uniformly bounded in time. Define the quantity $Q(t) = \left\| (\nabla_\theta f(\theta(t)), \hat{u}(t)) \right\|_2^2$, and notice that due to estimate (2.18) and the bound on $\nabla_\theta f(\theta)$ by Assumption 2.6, we obtain that $\sup_{t\geq 0} Q(t) < C$ for some appropriate constant $C < \infty$.

By direct calculation, keeping in mind the equation that $\theta(t)$ satisfies and Hölder inequality, we obtain

$$
\begin{aligned}
\frac{d}{dt}\|\theta(t)\|_2^2 &= -2\gamma\alpha(t)\|\theta(t)\|_2^2 - 2\alpha(t)\theta_t(\nabla_\theta f(\theta(t)), \hat{u}(t)) \\
&\leq -\gamma\alpha(t)\|\theta(t)\|_2^2 + \frac{4}{\gamma}\alpha(t)Q(t).
\end{aligned}
$$

By comparison principle we then have that

$$
\begin{aligned}
\|\theta(t)\|_2^2 &\leq e^{-\gamma \int_0^t \alpha(s)ds}\|\theta(0)\|_2^2 + \frac{4}{\gamma}\int_0^t \alpha(s)e^{-\gamma \int_s^t \alpha(r)dr}Q(s)ds \\
&\leq e^{-\gamma \int_0^t \alpha(s)ds}\|\theta(0)\|_2^2 + C\int_0^t \alpha(s)e^{-\gamma \int_s^t \alpha(r)dr}ds
\end{aligned}
\tag{2.19}
$$

and the result follows by requiring that for some $C < \infty$ (see Assumption 2.5)

$$
\sup_{t\geq 0} e^{-\gamma \int_0^t \alpha(s)ds} + \sup_{t\geq 0}\int_0^t \alpha(s)e^{-\gamma \int_s^t \alpha(r)dr}ds \leq C
$$

All in all we obtain the a-priori global estimate

$$\|\rho\|_{2,\infty} \leq K$$

for some finite constant $K < \infty$. With this a-priori bound the solution can be extended indefinitely in time. Thus a unique global solution exists. This means that there exists a unique global mild solution $\rho \in \mathcal{C}([0,\infty), X^2)$.

Essentially the same argument as above shows that if the initial data are in L^q and $h \in L^q$, for $q > 2$, then we will have that there is a unique global mild solution $\rho \in \mathcal{C}([0,\infty), L^q(\Theta, dn))$.

Let us now discuss regularity. We will prove that for initial data and h in L^q, we actually have that $u, \hat{u} \in \mathcal{C}([0,\infty), L^p)$ for any $p \in [q,\infty)$. We will use a bootstrap argument. Due to Sobolev embedding theorem and Riesz-Thorin theorem the following $L^q \to L^p$ estimate for the semigroup $S(t)$ holds

$$\|S(t)g\|_{L^p(U)} \leq C(t \wedge 1)^{-\frac{n}{2}\left(\frac{1}{q} - \frac{1}{p}\right)} \|g\|_{L^q(U)} \tag{2.20}$$

where n is the spatial dimension, $p \geq q$ and g a test function. Then, let us consider $p > q$ such that $\frac{1}{n} = \frac{1}{q} - \frac{1}{p}$ and assume that we know $\|u\|_{L^q} < \infty$. Consider an initial time $t = \epsilon$ for some fixed $\epsilon > 0$ and using (2.20) we have for u (we use the mild formulation of the solution)

$$\|u(t+\epsilon)\|_{L^p} \leq C\left[t^{-\frac{1}{2}} \|u(\epsilon)\|_{L^q} + \int_0^t (t-s)^{-\frac{1}{2}} \|u(s+\epsilon)\|_{L^q} \, ds \right]$$

$$\leq C\left[t^{-\frac{1}{2}} \|u(\epsilon)\|_{L^q} + t^{\frac{1}{2}} \sup_{s \in [\epsilon, \epsilon+t]} \|u(s)\|_{L^q} \right] ds.$$

Next consider the solution u starting at time $t = 2\epsilon$ with initial data $u(2\epsilon) \in L^p$. Then, we will have that $u \in \mathcal{C}([2\epsilon,\infty), L^p)$. Notice now that $\epsilon > 0$ is arbitrary. Thus we obtain that $u \in \mathcal{C}([0,\infty), L^p)$ for $p \geq q$ such that $\frac{1}{n} = \frac{1}{q} - \frac{1}{p}$. Using this argument inductively, first with $q = 2$ and $p > 2$ such that $\frac{1}{n} = \frac{1}{2} - \frac{1}{p}$ we then get that $u \in \mathcal{C}([0,\infty), L^p)$ for any $p \in [2,\infty)$.

Following exactly the same process and using that $u \in \mathcal{C}([0,\infty), L^p)$ for any $p \in [2,\infty)$ we then obtain that if $h \in L^p$ for all $p \geq 2$, then $\hat{u} \in \mathcal{C}([0,\infty), L^p)$ for any $p \in [2,\infty)$ as well.

Next, we notice that the forcing term $\mathcal{R}$ in (2.13) is in $L^p(U)$, so by the parabolic estimates in Sect. IV.3 of [22], we get that $u, \hat{u} \in \mathcal{C}((0,\infty); W_0^{2,p}(U)) \cap \mathcal{C}^1((0,\infty); L^p(U))$ and if the initial data $u_0, \hat{u}_0 \in W^{2,p}(U)$, then $u, \hat{u} \in \mathcal{C}([0,\infty); W_0^{2,p}(U)) \cap \mathcal{C}^1([0,\infty); L^p(U))$ for any $p \geq 2$. This concludes the proof of the theorem. $\qquad\square$

3 Convergence to a Stationary Point

The main result of this section is the convergence result for $\theta(t)$. It says that as $t \to \infty$, $\theta(t)$ converges to a stationary point of the cost function $J(\theta)$.

Theorem 3.1 *Assume Assumptions 2.1, 2.5, 2.6 and 2.7. Then, we have that*

$$\lim_{t \to \infty} \|\nabla J(\theta(t))\|_2 = 0.$$

The proof of Theorem 3.1 is be a consequence of series of lemmas. In Sect. 3.1 we establish necessary decay rates for the solution to (1.5). These results are then used in Sect. 3.2 to characterize the behavior of $\theta(t)$ for large times and eventually prove Theorem 3.1.

3.1 Decay Rates for the Online Adjoint Algorithm (1.5)

In this subsection, we establish some necessary decay rates for the online adjoint algorithm (1.5).

First, Lemma 3.2 has a critical bound on $\left\| \frac{\partial u^*}{\partial t} \right\|_H$ and on $\left\| \frac{\partial \hat{u}^*}{\partial t} \right\|_H$.

Lemma 3.2 *Under Assumptions 2.1 and 2.6, there exists a constant $C < \infty$ such that*

$$\left\| \frac{\partial u^*}{\partial t} \right\|_H + \left\| \frac{\partial \hat{u}^*}{\partial t} \right\|_H < C\alpha(t)$$

and consequently $\lim_{t \to \infty} \left\| \frac{\partial u^*}{\partial t} \right\|_H = \lim_{t \to \infty} \left\| \frac{\partial \hat{u}^*}{\partial t} \right\|_H = 0.$

Proof First, we study $\left\| \frac{\partial u^*}{\partial t} \right\|_H$. We see that $\frac{\partial u^*}{\partial t}(t, x)$ satisfies the PDE

$$A\frac{\partial u^*}{\partial t}(t, x) = \nabla_\theta f(x, \theta(t))^\top \frac{d\theta}{dt}$$

$$= -\alpha(t)\nabla_\theta f(x, \theta(t))^\top \left(\left(\nabla_\theta f(\theta(t)), \hat{u}(t) \right) + \gamma\theta(t) \right), \quad x \in U, t > 0$$

$$\frac{\partial u^*}{\partial t} = 0, \quad x \in \partial U, t > 0 \tag{3.1}$$

By the coercivity assumption on A by Assumption 2.1 we subsequently obtain

$$\left\| \frac{\partial u^*}{\partial t} \right\|_H^2 \leq \frac{1}{\lambda} \left(A\frac{\partial u^*}{\partial t}, \frac{\partial u^*}{\partial t} \right)$$

$$\leq \frac{1}{\lambda}\alpha(t) \left| \left(\nabla_\theta f(\theta(t))^\top \left(\left(\nabla_\theta f(\theta(t)), \hat{u}(t) \right) + \gamma\theta(t) \right), \frac{\partial u^*}{\partial t} \right) \right|$$

$$\leq \frac{1}{\lambda}\alpha(t) \left\| \nabla_\theta f(\theta(t))^\top \left(\left(\nabla_\theta f(\theta(t)), \hat{u}(t) \right) + \gamma\theta(t) \right) \right\|_H \left\| \frac{\partial u^*}{\partial t} \right\|_H$$

and the result follows directly by Assumption 2.6 on $\nabla_\theta f$ and estimate (2.12).

Let us now turn our attention to $\left\| \frac{\partial \hat{u}^*}{\partial t} \right\|_H$. By differentiation, we obtain that $\frac{\partial u^*}{\partial t}(t, x)$ satisfies the PDE

$$A^\dagger \frac{\partial \hat{u}^*}{\partial t}(t, x) = \frac{\partial u^*}{\partial t}(t, x), \quad x \in U, t > 0$$
$$\frac{\partial \hat{u}^*}{\partial t} = 0, \quad x \in \partial U, t > 0 \tag{3.2}$$

The result then follows by the coercivity condition on $A^\dagger$ by Assumption 2.1 and due to the fact that $\left\| \frac{\partial u^*}{\partial t} \right\|_H \leq C\alpha(t)$. This completes the proof of the lemma. $\qquad\square$

Let us consider the difference

$$\phi(t, x) = u(t, x) - u^*(t, x). \tag{3.3}$$

$\phi(t)$ satisfies the PDE

$$\frac{\partial \phi}{\partial t}(t, x) = -A\phi(t, x) - \frac{\partial u^*}{\partial t}(t, x), \quad x \in U, t > 0$$
$$\phi(t, x) = 0, \quad x \in \partial U, t > 0$$
$$\phi(0, x) = u_0(x) - u^*(0, x), \quad x \in U \tag{3.4}$$

Lemma 3.3 *Under Assumptions 2.1, 2.5 and 2.6 we have that*

$$\lim_{t \to \infty} \|\phi(t)\|_H = 0, \tag{3.5}$$

and there is some finite $T^* < \infty$ such that for all $t \geq T^*$

$$\|\phi(t)\|_H \leq C \left(e^{-\lambda t} + \alpha(t) \right), \tag{3.6}$$

where $C < \infty$ is an unimportant constant.

Proof We begin by proving that $\frac{\partial u^*}{\partial t}(t)$ is globally Lipschitz in time. Differentiating the elliptic PDE that u^* satisfies twice with respect to t yields for $x \in U$ and $t \geq 0$

$$A \frac{\partial^2 u^*}{\partial t^2} = -\alpha'(t) \nabla_\theta f(\theta(t))^\top \left(\left(\nabla_\theta f(\theta(t)), \hat{u}(t) \right) + \gamma\theta(t) \right)$$
$$- \alpha(t) \frac{\partial}{\partial t} \left[\nabla_\theta f(\theta(t))^\top \left(\left(\nabla_\theta f(\theta(t)), \hat{u}(t) \right) + \gamma\theta(t) \right) \right].$$

Therefore, using the bounds on $f(\theta), \nabla f(\theta), \theta(t), u(t),$ and $\hat{u}(t)$, we can show, using the coercivity Assumption 2.1 and the Cauchy-Schwarz inequality as in Lemma 3.2,

that $\sup\limits_{t<\infty} \left\| \dfrac{\partial^2 u^*}{\partial t^2} \right\|_H \leq C$. We provide the necessary calculations below for completeness.

$$\left\| \frac{\partial^2 u^*}{\partial t^2} \right\|_H^2 \leq \frac{1}{\lambda}(A\frac{\partial^2 u^*}{\partial t^2}, \frac{\partial^2 u^*}{\partial t^2}) \leq \frac{1}{\lambda}\left\| A\frac{\partial^2 u^*}{\partial t^2} \right\|_H \left\| \frac{\partial^2 u^*}{\partial t^2} \right\|_H$$
$$\leq \frac{K}{\lambda}\left\| \frac{\partial^2 u^*}{\partial t^2} \right\|_H,$$

where K is a constant. Re-arranging yields, for $t \geq 0$,

$$\left\| \frac{\partial^2 u^*}{\partial t^2} \right\|_H \leq C,$$

where C is a constant. This, then gives

$$\left\| \frac{\partial u^*}{\partial t}(t) - \frac{\partial u^*}{\partial t}(s) \right\|_H = \left\| \int_s^t \frac{\partial^2 u^*(\rho)}{\partial \rho^2}d\rho \right\|_H \leq \int_s^t \left\| \frac{\partial^2 u^*(\rho)}{\partial \rho^2} \right\|_H d\rho \leq C|t-s|. \tag{3.7}$$

Therefore, we can write

$$\phi(t) = S(t)\phi(0) - \int_0^t S(t-\tau)\frac{\partial u^*}{\partial \tau}d\tau,$$

where $\|S(t)\|_H \leq e^{-\lambda t}$ with $\lambda > 0$ by Assumptions 2.1.

Due to Lemma 3.2, for any $\epsilon > 0$, there exists a s such that $\left\| \dfrac{\partial u^*}{\partial \tau} \right\|_H < \epsilon$ for $t > s$. By the triangle inequality,

$$\|\phi(t)\|_H \leq \|S(t)\phi(0)\|_H + \int_0^t \left\| S(t-\tau)\frac{\partial u^*}{\partial \tau} \right\|_H d\tau$$
$$\leq e^{-\lambda t}\|\phi(0)\|_H + C_2 \int_0^t e^{-\lambda(t-\tau)}\alpha(\tau)d\tau.$$

Let us now define $I(t) = \int_0^t e^{-\lambda(t-\tau)}\alpha(\tau)d\tau$. Next, it is easy to show that $\lim_{t\to\infty} I(t) = 0$ and in particular that the integral term goes to zero at the rate of $\alpha(t)$ in the sense that $\lim_{t\to\infty} \frac{I(t)}{\alpha(t)} = \frac{1}{\lambda}$. These observations imply that there is a finite $T^* < \infty$ such that for all $t \geq T^*$, we have that $I(t) \leq C\alpha(t)$ for some constant $C < \infty$.

Hence we indeed get that both (3.5) and (3.6) hold, concluding the proof of the lemma. □

Define $\Psi(t, x) = \hat{u}(t, x) - \hat{u}^*(t, x)$. A similar lemma can also be proven for the limit of $\Psi(t)$.

Lemma 3.4 *Under Assumptions* 2.1, 2.5 *and* 2.6 *we have that*

$$\lim_{t \to \infty} \|\Psi(t)\|_H = 0, \tag{3.8}$$

and there is some finite $T^* < \infty$ such that for all $t \geq T^*$

$$\|\Psi(t)\|_H \leq C \left(e^{-\lambda t} t + \alpha(t) \right), \tag{3.9}$$

where $C < \infty$ is an unimportant constant.

Proof $\Psi(t)$ satisfies the PDE

$$\frac{\partial \Psi}{\partial t}(t, x) = -A^\dagger \Psi(t, x) + \phi(t, x) - \frac{\partial \hat{u}^*}{\partial t}(t, x). \tag{3.10}$$

Exactly, as it was done in Lemma 3.3 we can show that $t \mapsto \frac{\partial \hat{u}^*}{\partial t}(t)$ is globally Lipschitz. Together with Assumption 2.1 we write

$$\Psi(t) = S^\dagger(t)\Psi(0) + \int_0^t S^\dagger(t - \tau) \left(\phi(\tau) - \frac{\partial \hat{u}^*}{\partial t}(\tau) \right) d\tau,$$

where $S^\dagger(t)$ is the analytic contraction semigroup generated by $A^\dagger$ satisfying $\left\| S^\dagger(t) \right\|_H \leq e^{-\lambda t}$ with $\lambda > 0$. Using the same reasoning as in Lemma 3.3, and the fact that $\lim_{t \to \infty} \left(\|\phi(t)\|_H + \left\| \frac{\partial \hat{u}^*}{\partial t}(t) \right\|_H \right) = 0$, we can prove (3.8). The decay rates for $\|\phi(t)\|_H$ and $\left\| \frac{\partial \hat{u}^*}{\partial t}(t) \right\|_H$ from Lemma 3.3 and 3.2 respectively prove (3.9), the same way (3.6) was proven. This concludes the proof of the lemma. $\qquad\square$

3.2 Proof of Theorem 3.1

Let us now return to the equation for $\theta(t)$.

$$\begin{aligned}
\frac{d\theta}{dt} &= -\alpha(t) \Big(\big(\nabla_\theta f(\theta(t)), \hat{u}(t) \big) + \gamma \theta(t) \Big) \\
&= -\alpha(t) \nabla_\theta J(\theta(t)) - \alpha(t) \Big(\big(\nabla_\theta f(\theta(t)), \hat{u}^*(t) \big) - \big(\nabla_\theta f(\theta(t)), \hat{u}(t) \big) \Big) \\
&= -\alpha(t) \nabla_\theta J(\theta(t)) - \alpha(t) \big(\nabla_\theta f(\theta(t)), \Psi(t) \big). \tag{3.11}
\end{aligned}$$

The second term on the RHS will converge to zero as $t \to \infty$ due to (3.8). This means that asymptotically θ will be updated in the direction of steepest descent. Theorem 3.1 rigorously proves this along with proving convergence of $\theta(t)$ to a critical point of $J(\theta)$.

The structure of the proof proceeds in a spirit similar to [29] with certain differences that will be highlighted below as needed. For completeness, we present the whole argument with the proper adjustments.

Let $\epsilon > 0$ be given and let $\mu = \mu(\epsilon) > 0$ to be chosen later on. We define the following cycle of times

$$0 = \sigma_0 \leq \tau_1 \leq \sigma_1 \leq \tau_2 \leq \sigma_2 \leq \cdots$$

where for $k = 1, 2, \ldots$

$$\tau_k = \inf \{t > \sigma_{k-1} : \|\nabla J(\theta(t))\| \geq \epsilon\},$$

$$\sigma_k = \sup \left\{t > \tau_k : \frac{\|\nabla J(\theta(\tau_k))\|}{2} \leq \|\nabla J(\theta(s))\| \leq 2\|\nabla J(\theta(\tau_k))\| \text{ for all } s \in [\tau_k, t] \right.$$

$$\left. \text{and} \int_{\tau_k}^t \alpha(s)ds \leq \mu \right\}.$$

Essentially, the sequence of times $\{\sigma_k\}_{k \in \mathbb{N}}$ and $\{\tau_k\}_{k \in \mathbb{N}}$ keep track of the times in which $\|\nabla J(\theta(t))\|$ is within a ball of radius ϵ and away from it.

Next, let us define the corresponding intervals of time $J_k = [\sigma_{k-1}, \tau_k)$ and $I_k = [\tau_k, \sigma_k)$. Clearly, when $t \in J_k$, then we have that $\|\nabla J(\theta(t))\| < \epsilon$.

Let us now go back to (3.11) and define the integral term

$$\Delta_{s,t} = \int_s^t \alpha(\rho) \left(\nabla_\theta f(\theta(\rho)), \Psi(\rho)\right) d\rho. \tag{3.12}$$

We first show that $\Delta_{\tau_k, \sigma_k}$ decays to zero. In particular, we have the following lemma.

Lemma 3.5 *Assume that Assumptions 2.1, 2.5 and 2.6 hold. Let us fix some $\eta > 0$. Then, we have that $\lim_{k \to \infty} \left\|\Delta_{\tau_k, \sigma_k + \eta}\right\|_2 = 0$.*

Proof of Lemma 3.5 We notice that for some constant $C < \infty$ that may change from line to line

$$\sup_{t > 0} \left\|\Delta_{0,t}\right\|_2 \leq \int_0^\infty \alpha(\rho) \|(\nabla_\theta f(\theta(\rho)), \Psi(\rho))\|_2 \, d\rho$$

$$\leq \int_0^\infty \alpha(\rho) \|\nabla_\theta f(\theta(\rho))\|_H \|\Psi(\rho)\|_H \, d\rho$$

$$\leq C \int_0^\infty \alpha(\rho) \|\Psi(\rho)\|_H \, d\rho$$

$$\leq C \int_0^\infty \left[\alpha^2(\rho) + \alpha(\rho)\rho e^{-\lambda\rho}\right] d\rho$$

$$\leq C < \infty,$$

where the boundedness of $\|\nabla_\theta f(\theta)\|_H$ together with the decay rates from Lemma 3.4 were used. This immediately proves that $\lim_{k \to \infty} \left\|\Delta_{\tau_k, \sigma_k + \eta}\right\|_2 = 0$ concluding the proof of the lemma. □

Lemma 3.6 *Assume that Assumptions* 2.1, 2.5 *and* 2.6 *hold. Denote by $L_{\nabla J}$ to be the Lipschitz constant of ∇J. For given $\epsilon > 0$, let μ be such that $3\mu + \frac{\mu}{8\epsilon} = \frac{1}{2L_{\nabla J}}$. Then, for k large enough and for $\eta > 0$ small enough (potentially depending on k), one has $\int_{\tau_k}^{\sigma_k + \eta} \alpha(s)ds > \mu$. In addition, we also have $\frac{\mu}{2} \leq \int_{\tau_k}^{\sigma_k} \alpha(s)ds \leq \mu$.*

Proof of Lemma 3.6 The proof proceeds by contradiction. Let us assume that $\int_{\tau_k}^{\sigma_k + \eta} \alpha(s) ds \leq \mu$ and let $\delta > 0$ be such that $\delta < \mu/8$. In addition, without loss of generality, we can assume that for the given k, η is so small such that for any $s \in [\tau_k, \sigma_k + \eta]$ one has $\|\nabla J(\theta(s))\|_2 \leq 3 \|\nabla J(\theta(\tau_k))\|_2$.

Then, invoking (3.11) we have

$$\|\theta(\sigma_k + \eta) - \theta(\tau_k)\|_2 \leq \int_{\tau_k}^{\sigma_k + \eta} \alpha(t) \|\nabla_\theta J(\theta(t))\|_2 \, dt + \|\Delta_{\tau_k, \sigma_k + \eta}\|_2$$
$$\leq 3 \|\nabla J(\theta(\tau_k))\|_2 \, \mu + \|\Delta_{\tau_k, \sigma_k + \eta}\|_2.$$

By Lemma 3.5 we have that for k large enough, $\|\Delta_{\tau_k, \sigma_k + \eta}\|_2 \leq \delta < \mu/8$. In addition, we also have by definition that $\frac{\epsilon}{\|\nabla J(\theta(\tau_k))\|_2} \leq 1$. The combination of these two results gives

$$\|\theta(\sigma_k + \eta) - \theta(\tau_k)\| \leq \|\nabla J(\theta(\tau_k))\|_2 \left(3\mu + \frac{\mu}{8\epsilon}\right) \leq \|\nabla J(\theta(\tau_k))\|_2 \frac{1}{2L_{\nabla J}}.$$

This means that

$$\|\nabla J(\theta(\sigma_k + \eta)) - \nabla J(\theta(\tau_k))\|_2 \leq L_{\nabla J} \|\theta(\sigma_k + \eta) - \theta(\tau_k)\|_2 \leq \frac{1}{2} \|\nabla J(\theta(\tau_k))\|_2.$$

Then, this would yield

$$\frac{1}{2} \|\nabla J(\theta(\tau_k))\|_2 \leq \|\nabla J(\theta(\sigma_k + \eta))\|_2 \leq 2 \|\nabla J(\theta(\tau_k))\|_2.$$

However, this is a contradiction, because that would mean that $\int_{\tau_k}^{\sigma_k + \eta} \alpha(s)ds > \mu$, since otherwise $\sigma_k + \eta \in [\tau_k, \sigma_k]$ which cannot happen because $\eta > 0$. This concludes the proof of the first part of the lemma. The proof of the second part of the lemma goes as follows. By its own definition, we have that $\int_{\tau_k}^{\sigma_k} \alpha(s)ds \leq \mu$. Next, we show that $\int_{\tau_k}^{\sigma_k} \alpha(s)ds \geq \mu/2$. We have shown that $\int_{\tau_k}^{\sigma_k + \eta} \alpha(s)ds > \mu$. For k large enough and η small enough we can choose that $\int_{\sigma_k}^{\sigma_k + \eta} \alpha(s)ds \leq \mu/2$. The conclusion then follows. This concludes the proof of the lemma. □

Lemma 3.7 *Assume that Assumptions* 2.1, 2.5 *and* 2.6 *hold. Assume that there exists an infinite number of intervals $I_k = [\tau_k, \sigma_k)$. Then, there is a fixed $\zeta_1 > 0$ that depends on ϵ such that for k large enough*

$$J(\theta_{\sigma_k}) - J(\theta_{\tau_k}) \leq -\zeta_1.$$

Proof of Lemma 3.7 By chain rule we have that

$$J(\theta(\sigma_k)) - J(\theta(\tau_k)) = -\int_{\tau_k}^{\sigma_k} \alpha(\rho) \|\nabla J(\theta(\rho))\|_2 \, d\rho$$

$$+ \int_{\tau_k}^{\sigma_k} \alpha(\rho) \nabla J(\theta(\rho)) \cdot (\nabla_\theta f(\theta(\rho)), \Psi(\rho)) \, d\rho$$

$$= M_{1,k} + M_{2,k}. \tag{3.13}$$

Let us first consider $M_{1,k} = -\int_{\tau_k}^{\sigma_k} \alpha(\rho) \|\nabla J(\theta(\rho))\|_2 \, d\rho$. For $\rho \in [\tau_k, \sigma_k]$ we have that $\frac{\|\nabla J(\theta(\tau_k))\|_2}{2} \leq \|\nabla J(\theta(\rho))\|_2 \leq 2 \|\nabla J(\theta(\tau_k))\|_2$. Thus, for sufficiently large k, we have by Lemma 3.6

$$M_{1,k} \leq -\frac{\|\nabla J(\theta(\tau_k))\|_2^2}{4} \int_{\tau_k}^{\sigma_k} \alpha(\rho) d\rho \leq -\frac{\|\nabla J(\theta(\tau_k))\|_2^2}{8} \mu.$$

Next, we address $M_{2,k} = \int_{\tau_k}^{\sigma_k} \alpha(\rho) \nabla J(\theta(\rho)) \cdot (\nabla_\theta f(\theta(\rho)), \Psi(\rho)) \, d\rho$. Let us define

$$\hat{M}_{s,t} = \int_s^t \alpha(\rho) \nabla J(\theta(\rho)) \cdot (\nabla_\theta f(\theta(\rho)), \Psi(\rho)) \, d\rho.$$

Clearly, we have that $M_{2,k} = \hat{M}_{\tau_k, \sigma_k}$.

We claim that $\sup_{\rho \geq 0} \|\nabla J(\theta(\rho))\|_2 < \infty$. For this purpose, we shall use the representation of $\nabla J(\theta)$ by (2.10) together with the a-priori H norm bounds for $\hat{u}^*$, $\nabla_\theta f(x, \theta)$ as well as the uniform bound on $\sup_{\rho \geq 0} \|\theta(\rho)\|$ by (2.12). These imply that indeed $\sup_{\rho \geq 0} \|\nabla J(\theta(\rho))\|_2 < \infty$.

Then, we have for some constant $C < \infty$ that may change from line to line

$$\sup_{t \geq 0} \hat{M}_{0,t} \leq C \int_0^\infty \alpha(\rho) \|\nabla J(\theta(\rho))\|_2 \|\nabla_\theta f(\theta(\rho))\|_H \|\Psi(\rho)\|_H \, d\rho$$

$$\leq C \int_0^\infty \alpha(\rho) \|\Psi(\rho)\|_H \, d\rho$$

$$\leq C \int_0^\infty \left(\alpha^2(\rho) + \alpha(\rho)e^{-\lambda\rho}\rho\right) d\rho$$

$$\leq C < \infty.$$

where we used the decay rate bound from Lemma 3.8. The latter, then means that $M_{2,k} \to 0$ as $k \to \infty$.

Putting the above together, we get for k large enough such that $|M_{2,k}| \leq \delta < \frac{\mu}{16}\epsilon^2$

$$J(\theta(\sigma_k)) - J(\theta(\tau_k)) \leq -\frac{\|\nabla J(\theta(\tau_k))\|_2^2}{8}\mu + \delta$$

$$\leq -\frac{\mu}{8}\epsilon^2 + \frac{\mu}{16}\epsilon^2 = -\frac{\mu}{8}\epsilon^2.$$

Setting $\zeta_1 = \frac{\mu}{8}\epsilon^2$ we conclude the proof of the lemma. □

Lemma 3.8 *Assume that Assumptions 2.1, 2.5 and 2.6 hold. Assume that there exists an infinite number of intervals $I_k = [\tau_k, \sigma_k)$. Then, there is a fixed $0 < \zeta_2 < \zeta_1$ such that for k large enough*

$$J(\theta_{\tau_k}) - J(\theta_{\sigma_{k-1}}) \leq \zeta_2.$$

Proof of Lemma 3.8 Recall that $\|\nabla J(\theta(t))\|_2 \leq \epsilon$ for $t \in J_k = [\sigma_{k-1}, \tau_k]$. By chain rule we have

$$
\begin{aligned}
J(\theta(\tau_k)) - J(\theta(\sigma_{k-1})) &= -\int_{\sigma_{k-1}}^{\tau_k} \alpha(\rho)\|\nabla J(\theta(\rho))\|_2 \, d\rho \\
&\quad + \int_{\sigma_{k-1}}^{\tau_k} \alpha(\rho)\nabla J(\theta(\rho)) \cdot (\nabla_\theta f(\theta(\rho)), \Psi(\rho)) \, d\rho \\
&\leq \int_{\sigma_{k-1}}^{\tau_k} \alpha(\rho)\nabla J(\theta(\rho)) \cdot (\nabla_\theta f(\theta(\rho)), \Psi(\rho)) \, d\rho.
\end{aligned}
$$

As in the proof of Lemma 3.7 we get that for k large enough, the right hand side of the last display can be arbitrarily small. This concludes the proof of the lemma. □

Now we can conclude the proof of Theorem 3.1.

Proof of Theorem 3.1 Fix an $\epsilon > 0$. If there are finitely number of τ_k, then there is a finite T^* such that $\|\nabla J(\theta(t))\|_2 < \epsilon$ for $t \geq T^*$, which proves the theorem. So, we basically need to prove that there can only be finitely many τ_k. So, let us assume that there are infinitely many instances of τ_k. By Lemmas 3.7 and 3.8 we have for sufficiently large k that

$$
\begin{aligned}
J(\theta_{\sigma_k}) - J(\theta_{\tau_k}) &\leq -\zeta_1 \\
J(\theta_{\tau_k}) - J(\theta_{\sigma_{k-1}}) &\leq \zeta_2
\end{aligned}
$$

with $0 < \zeta_2 < \zeta_1$. Let N large enough so that the above relations hold simultaneously. Then we have

$$
\begin{aligned}
J(\theta(\tau_{n+1})) - J(\theta(\tau_N)) &= \sum_{k=N}^{n} \left[J(\theta(\sigma_k)) - J(\theta(\tau_k)) + J(\theta(\tau_{k+1})) - J(\theta(\sigma_k)) \right] \\
&\leq \sum_{k=N}^{n} (-\zeta_1 + \zeta_2) < 0.
\end{aligned}
$$

Letting $n \to \infty$, we get that $J(\theta(\tau_{n+1})) \to -\infty$, which is a contradiction, since by definition $J(\theta) \geq 0$. Thus, there must be at most finitely many τ_k. This concludes the proof of the theorem. □

4 Convergence Rate in the Strongly Convex Case

We will now prove a convergence rate for $\theta(t)$. First, we need to strengthen the assumptions on the cost function $J(\cdot)$. Namely, we now assume that $J(\theta)$ is strongly convex.

Assumption 4.1 We assume the following conditions:

- $J(\cdot) \in C^2$.
- There exists a unique global minimum θ^* where $\nabla_\theta J(\theta^*) = 0$.
- $H(\theta) = \nabla_{\theta\theta} J(\theta)$ is globally Lipschitz and $H(\theta^*)$ is positive definite. That is, there exists a constant $q > 0$ such that

$$\xi^\top H(\theta^*)\xi \geq q \, \|\xi\|_2^2 . \tag{4.1}$$

- The learning rate satisfies $\alpha(t) = C_\alpha a(t)$ where the learning rate magnitude C_α is selected such that $C_\alpha q > 1$. The learning rate function $a(t)$ satisfies $\lim_{t\to\infty} a(t) = 0$ and

$$\int_0^\infty a(s)ds = \infty \text{ and } \int_0^\infty a^2(s)ds < \infty.$$
$$\sup_{t\geq 0} \int_0^t a(s)e^{-\gamma \int_s^t a(r)dr} ds < \infty \text{ and } \lim_{t\to\infty} \frac{a'(t)}{a(t)} = 0.$$

An example is $a(t) = \frac{1}{1+t}$.

In this section we will assume without loss of generality that the learning rate is $\alpha(t) = \frac{1}{1+t}$. Even though this specific choice of the learning rate is not necessary for the results to hold, it will simplify the derivation of the convergence rate. The main result of this section is Theorem 4.2.

Theorem 4.2 *Let us assume that $\alpha(t) = 1/(1 + t)$. In addition, assume that Assumptions 2.1, 2.6 and 4.1 hold. Then we have that there exists a time $0 < t_0 < \infty$, such that for all $t \geq t_0$*

$$\left\| \theta(t) - \theta^* \right\|_2 \leq Ct^{-1/2}. \tag{4.2}$$

The proof of this theorem will be a consequence of a series of lemmas. Let us recall the function $\phi(t, x) = u(t, x) - u^*(t, x)$ satisfying (3.4)

$$\frac{\partial \phi}{\partial t}(t, x) = -A\phi(t, x) - \frac{\partial u^*}{\partial t}(t, x)$$

such that by Lemma 3.2, $t \mapsto \frac{\partial u^*}{\partial t}$ is globally Lipschitz and $\left\| \frac{\partial u^*}{\partial t} \right\|_H \leq C\frac{1}{1+t} < C\frac{1}{t}$.

Then, we have the following lemma.

Lemma 4.3 *Consider the setting of Theorem 4.2. We have that there is a finite constant $C < \infty$ such that*

$$\limsup_{t\to\infty} t^2 \left(\phi(t), \phi(t)\right) \le C. \tag{4.3}$$

In addition, there exists a $t_0 > 0$ such that for all $t \ge t_0$ and any $0 < p < 1$,

$$\left(\phi(t), \phi(t)\right) \le K t^{-2p}.$$

Proof of Lemma 4.3 For notational convenience we set below $Y(t) = (\phi(t), \phi(t))$. First we calculate

$$\frac{dY}{dt}(t) = -2\left(\phi(t), A\phi(t)\right) - 2\left(\phi(t), \frac{\partial u^*}{\partial t}(t)\right).$$

For $\epsilon > 0$ to be chosen later on and using the coercivity Assumption 2.1, we then have the inequality (omitting the argument t for notational convenience)

$$
\begin{aligned}
\frac{dY}{dt} &\le -2\left(\phi, A\phi\right) + 2\left|\left(\phi\epsilon, \frac{1}{\epsilon}\frac{\partial u^*}{\partial t}(t)\right)\right| \\
&\le -2\lambda(\phi, \phi) + \frac{1}{2}\epsilon^2(\phi, \phi) + \frac{1}{2\epsilon^2}\left(\frac{\partial u^*}{\partial t}, \frac{\partial u^*}{\partial t}\right) \\
&\le -2(\lambda - \frac{\epsilon^2}{2})Y + \frac{C}{2\epsilon^2}t^{-2} \\
&= -2b_\epsilon Y + C_\epsilon t^{-2},
\end{aligned}
$$

where we have used Young's inequality. The constant $b_\epsilon = \lambda - \frac{\epsilon^2}{2}$ and $C_\epsilon = \frac{C}{2\epsilon^2}$. We can select ϵ such that $b_\epsilon > 0$.

Denoting now for notational convenience $b = b^\epsilon$ and with some abuse of notation setting $C = C_\epsilon$, let's construct the ODE

$$
\begin{aligned}
\frac{dv}{dt} &= -2bv + Ct^{-2}, \\
v(1) &= Y(1).
\end{aligned}
$$

Define $\xi = Y - v$. Then, we have that $\xi(1) = 0$ and for $t \ge 1$,

$$
\begin{aligned}
\frac{d\xi}{dt} &= \frac{dY}{dt} - \frac{dv}{dt} \\
&\le -2bY + Ct^{-2} - \left(-2bv + Ct^{-2}\right) \\
&= -2b(Y - v) \\
&= -2b\xi.
\end{aligned}
$$

By Gronwall's inequality $\xi \leq 0$ and therefore $Y \leq v$. If we can establish a convergence rate for v, we then have a convergence rate for Y.

The solution v is

$$v(t) = e^{-2bt} \int_1^t e^{2bs} s^{-2} C ds,$$

We know that

$$\lim_{t \to \infty} t^2 v(t) = \frac{1}{2b}.$$

Therefore, for a finite constant $C < \infty$ we have that

$$\limsup_{t \to \infty} t^2 Y \leq \limsup_{t \to \infty} t^2 v \leq C.$$

We also consequently know that there exists a $t_0 > 0$ such that for all $t \geq t_0$ and any $0 < p < 1$,

$$|v(t)| \leq K t^{-2p}.$$

Therefore, for all $t \geq t_0$,

$$Y(t) \leq v(t) \leq K t^{-2p}.$$

concluding the proof of the lemma. $\square$

Let us now recall that $\Psi(t, x) = \hat{u}(t, x) - \hat{u}^*(t, x)$. Next, let's prove a convergence rate for Ψ.

Lemma 4.4 *Consider the setting of Theorem* 4.2. *Then, we have that*

$$\limsup_{t \to \infty} t^2 (\Psi(t), \Psi(t)) \leq C, \tag{4.4}$$

for some finite constant $C < \infty$.

Proof of Lemma 4.4 We first calculate that

$$\frac{\partial \Psi(t)}{\partial t}(t, x) = -A^\dagger \Psi(t, x) + \phi(t, x) - \frac{\partial \hat{u}^*}{\partial t}(t, x).$$

Define $W(t) = t^2 (\Psi(t), \Psi(t))$. Then, omitting for notational convenience the time argument, we have

$$\frac{dW}{dt} = 2t (\Psi, \Psi) + 2t^2 \left(\Psi, \frac{\partial \Psi}{\partial t} \right)$$

$$= \frac{2}{t}W - 2t^2\left(\Psi, A^\dagger\Psi\right) + 2t^2\left(\Psi, \phi\right) - 2t^2\left(\Psi, \frac{\partial \hat{u}^*}{\partial t}\right)$$

$$\leq \frac{2}{t}W - 2\lambda W + 2t^2\left(\Psi, \phi\right) - 2t^2\left(\Psi, \frac{\partial \hat{u}^*}{\partial t}\right),$$

where we used the assumed coercivity of $A^\dagger$ (consequence of Assumption 2.1). Let's select a $t_0 > 1$ such that $t_0^{-1} < \delta \ll c$. Then, for $t \geq t_0$ and for a constant $C < \infty$ that may change from line to line,

$$\frac{dW}{dt} \leq -2(\lambda - \delta)W + 2t^2\left(\Psi, \phi\right) + 2t^2\left|\left(\Psi, \frac{\partial \hat{u}^*}{\partial t}\right)\right|$$

$$\leq -2(\lambda - \delta)W + C\epsilon^2 t^2\left(\Psi, \Psi\right) + C\epsilon^{-2}t^2\left(\phi, \phi\right) + C\epsilon^{-2}t^2\left(\frac{\partial \hat{u}^*}{\partial t}, \frac{\partial \hat{u}^*}{\partial t}\right)$$

$$= -bV + Ct^2\left(\phi, \phi\right) + Ct^2\left(\frac{\partial \hat{u}^*}{\partial t}, \frac{\partial \hat{u}^*}{\partial t}\right).$$

where we have chosen an $\epsilon > 0$ such that $b = \lambda - \delta - C\epsilon^2 > 0$.

Let's construct the ODE

$$\frac{d\hat{q}}{dt} = -b\hat{q} + Ct^2\|\phi\|_H^2 + Ct^2\left\|\frac{\partial \hat{u}^*}{\partial t}\right\|_H^2, \quad t \geq t_0,$$

$$\hat{q}(t_0) = V(t_0).$$

which then, by (4.3) and Lemma 3.2, satisfies

$$\hat{q}(t) = e^{-bt}\hat{q}(t_0) + Ce^{-bt}\int_{t_0}^t e^{bs}s^2\left(\|\phi(s)\|_H^2 + \left\|\frac{\partial \hat{u}^*}{\partial t}(s)\right\|_H^2\right)ds$$

$$\leq e^{-bt}\hat{q}(t_0) + Ce^{-bt}\int_{t_0}^t e^{bs}ds$$

$$\leq C.$$

Therefore, using the same ODE comparison principle as before, we have the bound

$$\limsup_{t\to\infty} t^2\left(\Psi(t), \Psi(t)\right) \leq C,$$

which concludes the Proof of the Lemma. □

We now present the Proof of Theorem 4.2 on the convergence rate for θ.

Proof of Theorem 4.2 Recall that $H(\theta) = \nabla_{\theta\theta}J(\theta)$ is the Hessian matrix. At the stationary point θ^*, $H(\theta^*)$ is positive definite, i.e. there exists some constant $q > 0$ such that $\xi^\top H(\theta^*)\xi \geq q\|\xi\|_2^2$.

Since, according to Theorem 3.1, we have already proven convergence, we know that for θ^* such that $\nabla J(\theta^*) = 0$, we have that $\lim_{t\to\infty} \theta(t) = \theta^*$. The parameter updates satisfy

$$\frac{d\theta}{dt} = -\alpha(t)\nabla_\theta J(\theta(t)) - \alpha(t)\left(\nabla_\theta f(\theta(t)), \Psi(t)\right)$$

$$= -\alpha(t)H(\theta^*)(\theta(t) - \theta^*)$$

$$- \alpha(t)\nabla_\theta H(\bar{\theta}(t))_{j,k}(\theta(t) - \theta^*)_k(\theta(t) - \theta^*)_j - \alpha(t)\left(\nabla_\theta f(\theta(t)), \Psi(t)\right),$$

where $\bar{\theta}(t) \in [\theta(t), \theta^*]$, $\nabla_\theta H(\bar{\theta}(t))_{j,k} \in \mathbb{R}^d$, and $\nabla_\theta H(\bar{\theta}(t))_{j,k}(\theta(t) - \theta^*)_k(\theta(t) -$

$\theta^*)_j = \sum\limits_{k,j=1}^{d} \nabla_\theta H(\bar{\theta}(t))_{j,k}(\theta(t) - \theta^*)_k(\theta(t) - \theta^*)_j$. $H(\theta)$ is the Hessian matrix and

$H(\theta)_{j,k}$ is the (j, k)-th element of the matrix.

Define

$$V(t) = \left\|\theta(t) - \theta^*\right\|_2^2.$$

V satisfies the ODE

$$\frac{dV}{dt} = 2(\theta(t) - \theta^*)^\top \frac{d\theta}{dt}$$

$$= 2(\theta(t) - \theta^*)^\top$$

$$\left[-\alpha(t)H(\theta^*)(\theta(t) - \theta^*) - \alpha(t)\nabla_\theta H(\bar{\theta}(t))_{j,k}(\theta(t) - \theta^*)_k(\theta(t) - \theta^*)_j \right.$$

$$\left. - \alpha(t)\left(\nabla_\theta f(\theta(t)), \Psi(t)\right) \right]$$

$$\leq -2\alpha(t)qV(t) - 2\alpha(t)C\left\|\theta(t) - \theta^*\right\|V(t) - 2\alpha(t)(\theta(t) - \theta^*)^\top\left(\nabla_\theta f(\theta(t)), \Psi(t)\right).$$

Since $\lim_{t\to\infty} \theta(t) = \theta^*$, there exists a t_0 such that for all $t \geq t_0$

$$q - C\left\|\theta(t) - \theta^*\right\| > b_0 > 0.$$

Therefore, for $t \geq t_0$ large enough, we have

$$\frac{dV}{dt} \leq -\alpha(t)b_0V + \alpha(t)\left|(\theta(t) - \theta^*)^\top\left(\nabla_\theta f(\theta(t)), \Psi(t)\right)\right|$$

$$= -\alpha(t)b_0V + \epsilon\alpha(t)\left|(\theta(t) - \theta^*)^\top\left(\nabla_\theta f(\theta(t)), \frac{\Psi(t)}{\epsilon}\right)\right|$$

$$\leq -\alpha(t)b_0V + \epsilon^2\alpha(t)^2V + C\epsilon^{-2}(\Psi, \Psi),$$

where we have used Young's inequality and the Cauchy–Schwartz inequality. Here we have set $C = \sup_{\theta\in\mathbb{R}^d} \|\nabla_\theta f(\theta)\|_{L^2(U)}^2 < \infty$ by Assumption 2.6.

Let's select an ϵ small enough such that $b = b_0 - \epsilon^2 > 0$. Then,

$$\frac{dV}{dt} \leq -\alpha(t)bV + C(\Psi, \Psi).$$

Let $\hat{q}$ satisfy the ODE

$$\frac{d\hat{q}}{dt} = -\alpha(t)b\hat{q} + C(\Psi, \Psi), \quad t \geq t_0,$$
$$\hat{q}(t_0) = V(t_0).$$

The following comparison principle holds

$$V(t) \leq \hat{q}(t).$$

Using an integrating factor and $\limsup\limits_{t\to\infty} t^2 (\Psi, \Psi) \leq C$ by Lemma 4.4, we have that

$$
\begin{aligned}
\hat{q}(t) &= C_1 t^{-b} + C_2 t^{-b} \int_{t_0}^{t} s^b (\Psi(s), \Psi(s)) \, ds \\
&\leq C_1 t^{-b} + C_2 t^{-b} \int_{t_0}^{t} s^{b-2} s^2 (\Psi(s), \Psi(s)) \, ds \\
&\leq C_1 t^{-b} + C_2 t^{-b} \int_{t_0}^{t} s^{b-2} ds \\
&= C_1 t^{-b} + C_2 t^{-b} \left(t^{b-1} - t_0^{b-1} \right) \\
&\leq C t^{-1},
\end{aligned}
$$

where the constant C_2 may change from line to line and we have used the assumption $C_\alpha b > 1$. This concludes the convergence rate proof of $\theta(t)$. $\qquad\square$

Declarations

References

1. Benveniste, A., Metivier, M., Priouret, P.: Adaptive Algorithms and Stochastic Approximations. Springer, New York (2012)
2. Brandenburg, C., Lindemann, F., Ulbrich, M., Ulbrich, S.: A continuous adjoint approach to shape optimization for Navier Stokes flow. In: Optimal Control of Coupled Systems of Partial Differential Equations. In: Internat. Ser. Numer. Math., vol. 158. Birkhäuser Verlag, Basel, pp. 35–56 (2009)

3. Bosse, T., Gauger, N.R., Griewank, A., Günther, S., Schulz, V.: One-shot approaches to design optimization. Int. Ser. Numer. Math. **165**, 43–66 (2014)
4. Bueno-Orovio, A., Castro, C.C., Palacios, F., Zuazua, E.: Continuous adjoint approach for the Spalart-Allmaras model in aerodynamic optimization. AIAA J. **50**(3), 631–646 (2012)
5. Cagnetti, F., Gomes, D., Tran, H.: Adjoint methods for obstacle problems and weakly coupled systems of PDE. ESAIM **19**(3), 754–779 (2013)
6. Duta, M., Giles, M., Campobasso, M.: The harmonic adjoint approach to unsteady turbomachinery design. Int. J. Numer. Methods Fluids **40**(3–4), 323–332 (2002)
7. Gauger, N., Griewank, A., Hamdi, A., Kratzenstein, C., Özkaya, E., Slawig, T.: Automated extension of fixed point PDE solvers for optimal design with bounded retardation. In: Leugering, G., Engell, S., Griewank, A., Hinze, M., Rannacher, R., Schulz, V., Ulbrich, M., Ulbrich, S. (eds.) Constrained Optimization and Optimal Control for Partial Differential Equations, pp. 99–122. Springer, Basel (2012)
8. Gilbard, D., Trudinger, N.S.: Elliptic Partial Differential Equations of Second Order, 2nd edn. Springer, Berlin (2001)
9. Giles, M., Pierce, N.: An introduction to the adjoint approach to design. Flow Turbul. Combust. **65**, 393–415 (2000)
10. Giles, M., Ulbrich, S.: Convergence of linearized and adjoint approximations for discontinuous solutions of conservation laws. Part 1: Linearized approximations and linearized output functionals. SIAM J. Numer. Anal. **48**(3), 882–904 (2010)
11. Giles, M., Ulbrich, S.: Convergence of linearized and adjoint approximations for discontinuous solutions of conservation laws. Part 2: Adjoint approximations and extensions. SIAM J. Numer. Anal. **48**(3), 905–921 (2010)
12. Günther, S., Gauger, N.R., Wang, Q.: Simultaneous single-step one-shot optimization with unsteady PDEs. J. Comput. Appl. Math. **294**, 12–22 (2016)
13. Hazra, S.B.: Direct treatment of state constraints in aerodynamic shape optimization using simultaneous pseudo-time-stepping. AIAA J. **45**(8), 1988–1997 (2007)
14. Hazra, S.B., Schulz, V.: Simultaneous pseudo-timestepping for PDE-model based optimization problems. Bit Numer. Math. **44**(3), 457–472 (2004)
15. Hinze, M., Pinnau, R., Ulbrich, M., Ulbrich, S.: Optimization with PDE constraints. In: Mathematical Modelling: Theory and Applications, vol. 23. Springer, New York (2009)
16. Jameson, A.: Aerodynamic Shape Optimization Using the Adjoint Method. Lectures at the Von Karman Institute, Brussels (2003)
17. Jameson, A., Martinelli, L., Pierce, N.: Optimum aerodynamic design using the Navier-Stokes equations. Theoret. Comput. Fluid Dyn. **10**(1), 213–237 (1998)
18. Jameson, A., Kim, S.: Reduction of the adjoint gradient formula in the continuous limit. In: 41st Aerospace Sciences Meeting and Exhibit, p. 40 (2003)
19. Kaland, L., De Los Reyes, J.C., Gauger, N.R.: One-shot methods in function space for PDE-constrained optimal control problems. Optim. Methods Softw. **29**(2), 376–405 (2014)
20. Knopoff, D.A., Fernández, D.R., Torres, C.V.: Adjoint method for a tumor growth PDE-contrained optimization problem. Comput. Math. Appl. **66**, 1104–1119 (2013)
21. Kushner, H., Yin, G.: Stochastic Approximation and Recursive Algorithms and Applications, 2nd edn. Springer, New York (2003)
22. Ladyženskaja, O.A., Solonnikov, V.A., Ural'ceva, N.N.: Linear and quasilinear equations of parabolic type. American Mathematical Society, Providence (1968). (Translations of Mathematical Monographs, vol. 23, by S. Smith)
23. MacArt, J.F., Sirignano, J., Panesi, M.: Deep learning closure of the Navier-Stokes equations for transitional flows. In: AIAA SciTech (2021)
24. Nadarajah, S., Jameson, A.: A comparison of the continuous and discrete adjoint approach to automatic aerodynamic optimization. In: 38th Aerospace Sciences Meeting and Exhibit, p. 667 (2000)
25. Pazy, A.: Semigroups of Linear Operators and Applications to Partial Differential Equations. Springer, New York (1983)
26. Pierce, N., Giles, M.: Adjoint recovery of superconvergent functionals from PDE approximations. SIAM Rev. **42**(2), 247–264 (2000)
27. Protas, B.: Adjoint-based optimization of PDE systems with alternative gradients. J. Comput. Phys. **227**(13), 6490–6510 (2008)

28. Reuther, J., Jameson, A., Farmer, J., Martinelli, L., Saunders, D.: Aerodynamic shape optimization of complex aircraft configurations via an adjoint formulation. In: 34th Aerospace Sciences Meeting and Exhibit, p. 94 (1996)
29. Sirignano, J., Spiliopoulos, K.: Stochastic gradient descent in continuous time. SIAM J. Financ. Math. **8**(1), 933–961 (2017)
30. Sirignano, J., MacArt, J.F., Spiliopoulos, K.: PDE-constrained models with neural network terms: optimization and global convergence, arXiv:2105.08633 (2021)
31. Ta'asan, S.: One shot methods for optimal control of distributed parameter systems 1: finite dimensional control. Technical report, DTIC Document (1991)
32. Ta'asan, S.: Pseudo-time methods for constrained optimization problems governed by PDE, ICASE Report No. 95-32 (1995)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.