Weierstraß-Institut
für Angewandte Analysis und Stochastik

Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

Stochastic augmented Lagrangian method in shape spaces

Caroline Geiersbach[1], Tim Suchan[2], Kathrin Welker [3]

submitted: April 25, 2023

[1] Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: caroline.geiersbach@wias-berlin.de

[2] Helmut-Schmidt-Universität /
Universität der Bundeswehr Hamburg
Holstenhofweg 85
022043 Hamburg
Germany
E-Mail: suchan@hsu-hh.de

[3] Technische Universität Bergakademie Freiberg
Akademiestr. 6
09599 Freiberg
Germany
E-Mail: Kathrin.Welker@math.tu-freiberg.de

No. 3010
Berlin 2023

2020 *Mathematics Subject Classification.* 49Q10, 60H35, 35R15, 49K20, 41A25, 60H15, 60H30, 35R60.

Key words and phrases. Augmented Lagrangian, stochastic optimization, uncertainties, inequality constraints, Riemannian manifold, shape optimization, geometric constraints.

This work has been partly supported by the German Research Foundation (DFG) within the priority program SPP 1962 under contract number WE 6629/1-1 and by the state of Hamburg (Germany) within the Landesforschungsförderung under project “Simulation-Based Design Optimization of Dynamic Systems Under Uncertainties” (SENSUS) with project number LFF-GK11. Computational resources (HPC cluster HSUper) have been provided by the project hpc.bw, funded by dtec.bw – Digitalization and Technology Research Center of the Bundeswehr. dtec.bw is funded by the European Union – NextGenerationEU.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Stochastic augmented Lagrangian method in shape spaces

Caroline Geiersbach, Tim Suchan, Kathrin Welker

Abstract

In this paper, we present a stochastic Augmented Lagrangian approach on (possibly infinite-dimensional) Riemannian manifolds to solve stochastic optimization problems with a finite number of deterministic constraints. We investigate the convergence of the method, which is based on a stochastic approximation approach with random stopping combined with an iterative procedure for updating Lagrange multipliers. The algorithm is applied to a multi-shape optimization problem with geometric constraints and demonstrated numerically.

1 Introduction

In this paper, we concentrate on stochastic optimization problems of the form

$$\begin{aligned} &\min_{\boldsymbol{u}\in\mathcal{U}^N} \{j(\boldsymbol{u}) := \mathbb{E}[J(\boldsymbol{u},\boldsymbol{\xi})] = \int_\Omega J(\boldsymbol{u},\boldsymbol{\xi}(\omega))\,\mathrm{d}\mathbb{P}(\omega)\} \\ &\text{subject to (s.t.)} \quad h_i(\boldsymbol{u}) = 0 \quad i\in\mathcal{E}, \quad h_i(\boldsymbol{u}) \le 0 \quad i\in\mathcal{I}. \end{aligned} \tag{P}$$

Here, $\mathcal{U}^N$ is a Riemannian manifold and $\boldsymbol{\xi}\colon \Omega \to \Xi \subset \mathbb{R}^m$ is a random vector defined on a given probability space. We assume that we have deterministic constraints of the form $\boldsymbol{h}\colon \mathcal{U}^N \to \mathbb{R}^n$, $\boldsymbol{u} \mapsto \boldsymbol{h}(\boldsymbol{u}) = (h_1(\boldsymbol{u}),\dots,h_n(\boldsymbol{u}))^\top$, where we distinguish between the index set $\mathcal{E}$ of equality constraints and the index set $\mathcal{I}$ of inequality constraints.

Our investigations are motivated by applications in shape optimization, where an objective function is supposed to be minimized with respect to a shape, or a subset of $\mathbb{R}^d$. Finding a correct model to describe the set of shapes is one of the main challenges in shape optimization. From a theoretical and computational point of view, it is attractive to optimize in Riemannian manifolds because algorithmic ideas from [1] can be combined with approaches from differential geometry as outlined in [11]. Often more than one shape needs to be considered, which leads to so-called multi-shape optimization problems. As applications, we can mention electrical impedance tomography, where the material distribution of electrical properties such as electric conductivity and permittivity inside the body is examined [8, 21, 22], and the optimization of biological cell composites in the human skin [27, 28]. In [11], a shape is seen as a point on an abstract manifold so that a collection of shapes can be viewed as a vector of points $\boldsymbol{u} = (u_1,\dots,u_N)$ in a product manifold $\mathcal{U}^N = \mathcal{U}_1 \times \cdots \times \mathcal{U}_N$.

A central difficulty in (P) is that the constraints lead to a nonsmooth stochastic optimization problem that cannot be handled using standard techniques such as gradient descent or Newton's method; additionally, the numerical solution of the problem may be intractable on account of the expectation. In this work, we propose a stochastic augmented Lagrangian method to solve problems of the form (P). The proposed method combines the smoothing properties of the augmented Lagrangian method with a reduction in complexity granted by stochastic approximation.

The augmented Lagrangian method has been extensively studied; see [5, 6] for an introduction to the method in finite dimensions. Substantial theory can be found in the literature for PDE-constrained

DOI 10.20347/WIAS.PREPRINT.3010 Berlin 2023

optimization, where convergence has been studied in function spaces; see [19, 29, 16, 18, 17]. This theory does not apply even for deterministic counterparts of (P) since the control variable $\boldsymbol{u}$ belongs to a Riemannian manifold, not a Banach space. Stochastic approximation is a class of algorithms that originated from the paper [24] and has developed in recent decades due to its applicability to high-dimensional stochastic optimization problems. The most basic algorithm is the stochastic gradient method, which can be used to solve an unconstrained version of (P), i.e., the problem of minimizing the expectation. Recently, the stochastic gradient method was proposed to handle PDE-constrained shape optimization problems [11, 10]. In [10], asymptotic convergence was proven for optimization variables belonging to a Riemannian manifold and the connection was made to shape optimization following the ideas in [30]. However, the stochastic gradient method cannot solve nonsmooth problems of the form (P).

While both augmented Lagrangian and stochastic approximation methods are well-developed, the combined method—what we call the stochastic augmented Lagrangian method—is not. In the context of training neural networks, a combined stochastic gradient/augmented Lagrangian approach in the same spirit as ours can be found in the paper [9]. Our method, however, involves a novel use of the randomized multi-batch stochastic gradient method from [14, 15], where a random number of stochastic gradient steps are chosen. We use this strategy to solve the inner loop optimization problem for fixed Lagrange multipliers and penalty parameters. A central consequence of the random stopping rule from [14, 15] is that convergence rates of the expected value of the norm of the gradient can be obtained, even in the nonconvex case. The random stopping rule in combination with an outer loop procedure can be used to adaptively adjust step sizes and batch sizes for a tractable algorithm where asymptotic convergence to stationary points of the original nonsmooth problem is guaranteed.

The paper is structured as follows. In Section 2, we present the stochastic augmented Lagrangian method for optimization on Riemannian manifolds and analyze its convergence. Then, an application for our method is introduced and results of numerical tests are presented in Section 3.

2 Optimization approach

In this section, we introduce the stochastic augmented Lagrangian method for Riemannian manifolds. In view of our later application to shape optimization, where convexity of the objective function j cannot be expected, we focus on providing results for the nonconvex case. First, in Section 2.1, we will provide background material that will be of use in our analysis. The algorithm is presented in Section 2.2. Convergence of the method is proven in two parts: in Section 2.3, we provide an efficiency estimate for the inner loop procedure, corresponding to a randomized multi-batch stochastic gradient method. Then, in Section 2.4, convergence rates with respect to the outer loop procedure, which corresponds to a stochastic augmented Lagrangian method, are given.

2.1 Background and notation

For a differentiable Riemannian manifold (M, g), $g = (g_u)_{u\in M}$ denotes the Riemannian metric. The induced norm is denoted by $\|\cdot\|_g := \sqrt{g(\cdot,\cdot)}$. The derivative of a mapping $f\colon M \to S$ between two differentiable manifolds M and S is defined using the pushforward. In a point $u \in M$, it is defined by $(f_*)_u\colon T_uM \to T_{f(u)}S$ with $(f_*)_u(c) := \frac{\mathrm{d}}{\mathrm{d}t} f(c(t))|_{t=0} = (f\circ c)'(0)$, where $c'(t) \in T_uM$ and $c\colon I \subset \mathbb{R} \to M$ is a differentiable curve. In particular, $f\colon M \to S$ is called $\mathcal{C}^k$ if $\psi_\beta \circ f \circ \phi_\alpha^{-1}$ is k-times continuously differentiable for all charts (U_α, ϕ_α) of M and (V_β, ψ_β) of S with $f(U_\alpha) \subset V_\beta$.

The pullback by f in u is the mapping $f_u^*\colon T^*_{f(u)}S \to T^*_u M$. In the case $S = \mathbb{R}$, a Riemannian gradient $\nabla f(u) \in T_u M$ is defined by the relation

$$(f_*)_u w = g_u(\nabla f(u), w) \quad \forall w \in T_u M. \tag{1}$$

We denote the exponential mapping at u by $\exp_u\colon T_u M \to M$, $v \mapsto \exp_u(v)$, which assigns to every tangent vector v the value $\gamma(1)$ of the geodesic $\gamma\colon [0,1] \to M$ satisfying $\gamma(0) = u$ and $\gamma'(0) = v$. Let the length of a $\mathcal{C}^1$-curve $c\colon [0,1] \to M$ be denoted by $\mathsf{L}(c) = \int_0^1 \|c'(t)\|_g \,\mathrm{d}t$. Then the distance $\mathrm{d}\colon M \times M \to \mathbb{R}$ between points $u, q \in M$ is given by

$$\mathrm{d}(u,q) = \inf\{\mathsf{L}(c)\colon c\colon [0,1] \to M \text{ is a piecewise smooth curve with } c(0) = u \text{ and } c(1) = q\}.$$

The injectivity radius i_u at a point $u \in M$ is defined as

$$i_u := \sup\{r > 0\colon \exp_u|_{B_r(0_u)} \text{ is a diffeomorphism}\},$$

where 0_u denotes the zero element of $T_u M$ and $B_r(0_u) \subset T_u M$ is a ball centered at $0_u \in T_u M$ with radius r. The injectivity radius of the manifold M is the number $i(M) := \inf_{u \in M} i_u$.

As mentioned in the introduction, in this paper we will work with a (possibly infinite-dimensional) connected Riemannian product manifold $(M, g) = (\mathcal{U}^N, \mathcal{G}^N)$ equipped with the family of inner products $\mathcal{G}^N = (\mathcal{G}^N_{\boldsymbol{u}})_{\boldsymbol{u} \in \mathcal{U}^N}$. As described in [11], the tangent space $T\mathcal{U}^N$ can be identified with the product of tangent spaces $T\mathcal{U}_1 \times \cdots \times T\mathcal{U}_N$ via $T_{\boldsymbol{u}}\mathcal{U}^N \cong T_{u_1}\mathcal{U}_1 \times \cdots \times T_{u_N}\mathcal{U}_N$. Additionally, the product metric $\mathcal{G}^N$ to the corresponding product shape space $\mathcal{U}^N$ can be defined via $\mathcal{G}^N = \sum_{i=1}^N \pi_i^* \mathcal{G}^i$, where

$$\mathcal{G}^N_{\boldsymbol{u}}(\boldsymbol{v}, \boldsymbol{w}) = \sum_{i=1}^N \mathcal{G}^i_{\pi_i(u)}(\pi_{i*}\boldsymbol{v}, \pi_{i*}\boldsymbol{w}) \qquad \forall\, \boldsymbol{v}, \boldsymbol{w} \in T_u\mathcal{U}^N, \tag{2}$$

and $\pi_i\colon \mathcal{U}^N \to \mathcal{U}_i$, $i = 1, \ldots, N$, correspond to canonical projections. The multi-exponential map is denoted by

$$\exp^N_{\boldsymbol{u}}\colon T_{\boldsymbol{u}}\mathcal{U}^N \to \mathcal{U}^N,\ \boldsymbol{v} = (v_1, \ldots, v_N) \mapsto (\exp_{u_1} v_1, \ldots, \exp_{u_N} v_N)$$

for the vector $\boldsymbol{u} = (u_1, \ldots, u_N)$, where $\exp_{u_i}\colon T_{u_i}\mathcal{U}_i \to \mathcal{U}_i,\ v_i \mapsto \exp_{u_i}(v_i)$ for all $i = 1, \ldots, N$.

The triple $(\Omega, \mathcal{F}, \mathbb{P})$ denotes a (complete) probability space, where $\mathcal{F} \subset 2^\Omega$ is the σ-algebra of events and $\mathbb{P}\colon \Omega \to [0,1]$ is a probability measure. The expectation of a random variable $X\colon \Omega \to \mathbb{R}$ is defined by $\mathbb{E}[X] = \int_\Omega X(\omega)\,\mathrm{d}\mathbb{P}(\omega)$. A filtration is a sequence $\{\mathcal{F}_n\}$ of sub-σ-algebras of $\mathcal{F}$ such that $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}$. If for an event $F \in \mathcal{F}$ it holds that $\mathbb{P}(F) = 1$, then we say F occurs almost surely (a.s.). Given an integrable random variable $X\colon \Omega \to \mathbb{R}$ and a sub-σ-algebra $\mathcal{F}_n$, the conditional expectation is denoted by $\mathbb{E}[X|\mathcal{F}_n]$, which is a random variable that is $\mathcal{F}_n$-measurable and satisfies $\int_A \mathbb{E}[X|\mathcal{F}_n](\omega)\,\mathrm{d}\mathbb{P}(\omega) = \int_A X(\omega)\,\mathrm{d}\mathbb{P}(\omega)$ for all $A \in \mathcal{F}_n$.

We will frequently use the convention $\boldsymbol{\xi} \in \Xi$ to denote a realization (i.e., the deterministic value $\boldsymbol{\xi}(\omega) \in \Xi$ for some ω) of the vector $\boldsymbol{\xi}\colon \Omega \to \Xi \subset \mathbb{R}^m$; based on the context, there should be no confusion as to whether a realization or a random vector is meant. Let $J\colon \mathcal{U}^N \times \mathbb{R}^m \to \mathbb{R}$ be a parametrized objective as in problem (P) and define $J_{\boldsymbol{\xi}} := J(\cdot, \boldsymbol{\xi})$. The gradient $\nabla_{\boldsymbol{u}} J(\boldsymbol{u}, \boldsymbol{\xi}) := \nabla J_{\boldsymbol{\xi}}(\boldsymbol{u})$ of J with respect to $\boldsymbol{u}$ is defined by the relation

$$((J_{\boldsymbol{\xi}})_*)_{\boldsymbol{u}} \boldsymbol{w} = \mathcal{G}^N_{\boldsymbol{u}}(\nabla_{\boldsymbol{u}} J(\boldsymbol{u}, \boldsymbol{\xi}), \boldsymbol{w}) \quad \forall \boldsymbol{w} \in T_{\boldsymbol{u}}\mathcal{U}^N. \tag{3}$$

Following [10], if $\nabla_{\boldsymbol{u}} J \colon \mathcal{U}^N \times \mathbb{R}^m \to T_{\boldsymbol{u}}\mathcal{U}^N$ is $\mathbb{P}$-integrable, equation (3) is fulfilled for all $\boldsymbol{u}$ almost surely, and $\mathbb{E}[\nabla_{\boldsymbol{u}} J(\boldsymbol{u}, \boldsymbol{\xi})] = \nabla j(\boldsymbol{u})$, we call $\nabla_{\boldsymbol{u}} J$ a stochastic gradient.

Let the Lagrangian for problem (P) be the mapping $\mathcal{L} \colon \mathcal{U}^N \times \mathbb{R}^n \to \mathbb{R}$ defined by

$$\mathcal{L}(\boldsymbol{u}, \boldsymbol{\lambda}) := j(\boldsymbol{u}) + \boldsymbol{\lambda}^\top \boldsymbol{h}(\boldsymbol{u}).$$

The gradient $\nabla h_i(\boldsymbol{u}) \in T_{\boldsymbol{u}}\mathcal{U}^N$ of $h_i \colon \mathcal{U}^N \to \mathbb{R}$ is defined by the relation

$$((h_i)_*)_{\boldsymbol{u}} \boldsymbol{w} = \mathcal{G}^N(\nabla h_i(\boldsymbol{u}), \boldsymbol{w})$$

for all $\boldsymbol{w} \in T_{\boldsymbol{u}}\mathcal{U}^N$. The gradient of the corresponding vector $\boldsymbol{h} \colon \mathcal{U}^N \to \mathbb{R}^n$ is the vector $\nabla \boldsymbol{h}(\boldsymbol{u}) = (\nabla h_1(\boldsymbol{u}), \dots, \nabla h_n(\boldsymbol{u}))^\top$.

In the following, we define a *Karush–Kuhn–Tucker* (KKT) point. In order for the following KKT conditions to be necessary optimality conditions for problem (P), we need additional regularity conditions; we refer to [31, 4] for their treatment in manifolds.

Definition 2.1. *The pair $(\hat{\boldsymbol{u}}, \hat{\boldsymbol{\lambda}}) \in \mathcal{U}^N \times \mathbb{R}^n$ is called a KKT point for problem (P) if it satisfies the following conditions:*

$$\nabla j(\hat{\boldsymbol{u}}) + \sum_{i=1}^{n} \hat{\lambda}_i \nabla h_i(\hat{\boldsymbol{u}}) = \boldsymbol{0}_{\hat{\boldsymbol{u}}}, \tag{4a}$$

$$h_i(\hat{\boldsymbol{u}}) = 0, \quad \forall i \in \mathcal{E}, \tag{4b}$$

$$h_i(\hat{\boldsymbol{u}}) \leq \boldsymbol{0}, \quad \hat{\lambda}_i \geq \boldsymbol{0}, \quad \hat{\lambda}_i h_i(\hat{\boldsymbol{u}}) = 0, \quad \forall i \in \mathcal{I}. \tag{4c}$$

The closed cone corresponding to the geometric constraints, the distance to the cone, and the projection are defined, respectively, by

$$\boldsymbol{K} := \{\boldsymbol{y} \in \mathbb{R}^n : y_i = 0 \ \forall i \in \mathcal{E}, y_i \leq 0 \ \forall i \in \mathcal{I}\},$$
$$\operatorname{dist}_{\boldsymbol{K}}(\boldsymbol{y}) := \inf_{\boldsymbol{k} \in \boldsymbol{K}} \|\boldsymbol{y} - \boldsymbol{k}\|_2, \qquad \pi_{\boldsymbol{K}}(\boldsymbol{y}) := \operatorname{argmin}_{\boldsymbol{k} \in \boldsymbol{K}} \|\boldsymbol{y} - \boldsymbol{k}\|_2.$$

The ith component of $\boldsymbol{K}$ is denoted by K_i. For $y \in \mathbb{R}$, the projection has the formula $\pi_{K_i}(y) = 0$ if $i \in \mathcal{E}$, and $\pi_{K_i}(y) = \min(0, y)$ if $i \in \mathcal{I}$. We have $\pi_{\boldsymbol{K}}(\boldsymbol{y}) = (\pi_{K_1}(y_1), \dots, \pi_{K_n}(y_n))^\top$. The normal cone of $\boldsymbol{K}$ in a point $\boldsymbol{s} \in \boldsymbol{K}$ is defined by $N_{\boldsymbol{K}}(\boldsymbol{s}) = \{\boldsymbol{v} \in \mathbb{R}^n \colon \boldsymbol{v}^\top(\boldsymbol{s} - \boldsymbol{y}) \geq 0 \ \forall \boldsymbol{y} \in \boldsymbol{K}\}$; the normal cone is the empty set if $\boldsymbol{s}$ is not contained in $\boldsymbol{K}$. To define the augmented Lagrangian, we first introduce a slack variable $\boldsymbol{s} \in \boldsymbol{K}$ to obtain the equivalent, equality-constrained problem

$$\min_{(\boldsymbol{u}, \boldsymbol{s}) \in \mathcal{U}^N \times \boldsymbol{K}} \{j(\boldsymbol{u}) = \mathbb{E}[J(\boldsymbol{u}, \boldsymbol{\xi})]\} \qquad \text{s.t.} \quad \boldsymbol{h}(\boldsymbol{u}) - \boldsymbol{s} = \boldsymbol{0}.$$

The corresponding augmented Lagrangian for a fixed parameter μ is the mapping $\mathcal{L}_A^{\boldsymbol{s}} \colon \mathcal{U}^N \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ defined by

$$\begin{aligned} \mathcal{L}_A^{\boldsymbol{s}}(\boldsymbol{u}, \boldsymbol{s}, \boldsymbol{\lambda}; \mu) &= j(\boldsymbol{u}) + \boldsymbol{\lambda}^\top(\boldsymbol{h}(\boldsymbol{u}) - \boldsymbol{s}) + \frac{\mu}{2}\|\boldsymbol{h}(\boldsymbol{u}) - \boldsymbol{s}\|_2^2 \\ &= j(\boldsymbol{u}) + \frac{\mu}{2}\left\|\boldsymbol{h}(\boldsymbol{u}) + \frac{\boldsymbol{\lambda}}{\mu} - \boldsymbol{s}\right\|_2^2 - \frac{\|\boldsymbol{\lambda}\|_2^2}{2\mu}. \end{aligned}$$

Notice that $\min_{\boldsymbol{s} \in \boldsymbol{K}} \|\boldsymbol{h}(\boldsymbol{u}) + \frac{\boldsymbol{\lambda}}{\mu} - \boldsymbol{s}\|_2^2 = \operatorname{dist}_{\boldsymbol{K}}(\boldsymbol{h}(\boldsymbol{u}) + \frac{\boldsymbol{\lambda}}{\mu})^2$. Hence, it is possible to eliminate the slack variable to obtain, again for fixed μ, the augmented Lagrangian $\mathcal{L}_A \colon \mathcal{U}^N \times \mathbb{R}^n \to \mathbb{R}$ defined by

$$\mathcal{L}_A(\boldsymbol{u}, \boldsymbol{\lambda}; \mu) = j(\boldsymbol{u}) + \frac{\mu}{2} \operatorname{dist}_{\boldsymbol{K}}\left(\boldsymbol{h}(\boldsymbol{u}) + \frac{\boldsymbol{\lambda}}{\mu}\right)^2 - \frac{\|\boldsymbol{\lambda}\|_2^2}{2\mu}. \tag{5}$$

2.2 Augmented Lagrangian method on Riemannian manifolds

In this section, we present Algorithm 1, which relies on stochastic approximation. For this, we need the function $L_A\colon \mathcal{U}^N \times \mathbb{R}^n \times \Xi \to \mathbb{R}$ defined by

$$L_A(\boldsymbol{u}, \boldsymbol{\lambda}, \boldsymbol{\xi}; \mu) := J(\boldsymbol{u}, \boldsymbol{\xi}) + \frac{\mu}{2}\mathsf{dist}_{\boldsymbol{K}}\left(\boldsymbol{h}(\boldsymbol{u}) + \frac{\boldsymbol{\lambda}}{\mu}\right)^2 - \frac{\|\boldsymbol{\lambda}\|_2^2}{2\mu}.$$

Additionally, it will be convenient to define a feasibility measure and its induced sequence by

$$H(\boldsymbol{u}, \boldsymbol{\lambda}; \mu) := \left\| \boldsymbol{h}(\boldsymbol{u}) - \pi_{\boldsymbol{K}}\left(\boldsymbol{h}(\boldsymbol{u}) + \frac{\boldsymbol{\lambda}}{\mu}\right)\right\|_2, \quad H_k := H(\boldsymbol{u}^k, \boldsymbol{w}^{k-1}; \mu_{k-1}). \tag{6}$$

The stochastic augmented Lagrangian (AL) method is shown in Algorithm 1. The inner loop is an adaptation of the randomized mini-batch stochastic gradient (RSG) method from [15]. In deterministic AL methods, the inner loop is in practice only solved up to a given error tolerance, leading to an *inexact* augmented Lagrangian method. Deterministic termination conditions for the inner loop typically rely on conditions of the following type: $\boldsymbol{u}^{k+1}$ is chosen as the first point of the corresponding iterative procedure satisfying

$$\nabla_{\boldsymbol{u}} \mathcal{L}_A(\boldsymbol{u}^{k+1}, \boldsymbol{w}^k; \mu^k) = \boldsymbol{\varepsilon}_k$$

with the error disappearing asymptotically, i.e., $\boldsymbol{\varepsilon}_k \to \boldsymbol{0}$ as $k \to \infty$. Stochastic methods like the kind used here can only provide probabilistic error bounds; termination conditions are based on a priori estimates and result in stochastic errors. The outer loop corresponds to the augmented Lagrangian (AL) method with a safeguarding procedure as described in [16]; see also [29]. A feature of this procedure is that instead of using the Lagrange multiplier in the subproblem in line 4, one chooses a proxy function from a bounded set B, which is essential for achieving global convergence. In practice, this should be chosen in such a way so that the projection is easy to compute, i.e., box constraints are appropriate. A natural choice is $\boldsymbol{w}^k := \pi_B(\boldsymbol{\lambda}^k)$.

Algorithm 1 Stochastic Augmented Lagrangian Method

1: **Input:** Initial point $\boldsymbol{u}^1 = (u_1^1, \dots, u_N^1) \in \mathcal{U}^N$, AL parameters $\gamma > 1$, $\tau \in (0,1)$, $B \subset \mathbb{R}^n$
2: **Initialization:** $\mu_1 > 0$, $\boldsymbol{\lambda}^1 \in \mathbb{R}^n$, $k := 1$
3: **while** $\boldsymbol{u}^k$, $\boldsymbol{\lambda}^k$ not converged **do**
4: Choose $\boldsymbol{w}^k \in B$, step size t_k, iteration limit N_k, and batch size m_k
5: $\boldsymbol{z}^{k,1} := \boldsymbol{u}^k$
6: Take a sample R_k from the uniform distribution on $\{1, \dots, N_k\}$
7: **for** $j = 1, \dots, R_k$ **do**
8: Take i.i.d. samples $\{\boldsymbol{\xi}^{k,j,1}, \dots, \boldsymbol{\xi}^{k,j,m_k}\}$ according to probability distribution $\mathbb{P}$
9: $\boldsymbol{z}^{k,j+1} := \exp^N_{\boldsymbol{z}^{k,j}}(-\frac{t_k}{m_k}\sum_{s=1}^{m_k} \nabla_{\boldsymbol{u}} L_A(\boldsymbol{z}^{k,j}, \boldsymbol{w}^k, \boldsymbol{\xi}^{k,j,s}; \mu_k))$
10: $\boldsymbol{u}^{k+1} := \boldsymbol{z}^{k,j+1}$
11: $\boldsymbol{\lambda}^{k+1} := \mu_k\left(\boldsymbol{h}(\boldsymbol{u}^{k+1}) + \frac{\boldsymbol{w}^k}{\mu_k} - \pi_{\boldsymbol{K}}\left(\boldsymbol{h}(\boldsymbol{u}^{k+1}) + \frac{\boldsymbol{w}^k}{\mu_k}\right)\right)$
12: If $H_{k+1} \le \tau H_k$ or $k = 1$ satisfied, set $\mu_{k+1} = \mu_k$. Otherwise, set $\mu_{k+1} := \gamma\mu_k$.
13: $k := k + 1$

2.3 Convergence of inner loop

To prove convergence of the RSG procedure in Algorithm 1, we make the following assumptions about the manifold, which are adapted from [10].

Assumption 1. *We assume that (i) the distance $\mathrm{d}(\cdot,\cdot)$ is non-degenerate,*
(ii) the manifold $(\mathcal{U}^N, \mathcal{G}^N)$ has a positive injectivity radius $i(\mathcal{U}^N)$, and
(iii) for all $\boldsymbol{u} \in \mathcal{U}^N$ and all $\tilde{\boldsymbol{u}} \in B_{i_{\boldsymbol{u}}}(0_{\boldsymbol{u}})$, the minimizing geodesic between $\boldsymbol{u}$ and $\tilde{\boldsymbol{u}}$ is completely contained in $B_{i_{\boldsymbol{u}}}(0_{\boldsymbol{u}})$.

As pointed out in [10], the conditions in Assumption 1 are strong for infinite-dimensional manifolds. In infinite dimensions, Riemannian metrics are generally weak, so that gradients may not exist; in the stochastic setting, we need additional assumptions to ensure integrability. In the following, a function $g\colon \mathcal{U}^N \to \mathbb{R}$ is called L_g-Lipschitz continuously differentiable if the function is $\mathcal{C}^1$ and there exists a constant $L_g > 0$ such that for all $\boldsymbol{u}, \tilde{\boldsymbol{u}} \in \mathcal{U}^N$ with $\mathrm{d}(\boldsymbol{u}, \tilde{\boldsymbol{u}}) < i(\mathcal{U}^N)$, we have

$$\|P_{1,0}\nabla j(\tilde{\boldsymbol{u}}) - \nabla j(\boldsymbol{u})\|_{\mathcal{G}^N} \leq L_j \mathrm{d}(\boldsymbol{u}, \tilde{\boldsymbol{u}})$$

where $P_{1,0}\colon T_{\gamma(1)}\mathcal{U}^N \to T_{\gamma(0)}\mathcal{U}^N$ is the parallel transport along the unique geodesic such that $\gamma(0) = \boldsymbol{u}$ and $\gamma(1) = \tilde{\boldsymbol{u}}$.

Assumption 2. *(i) The functions j and h_i ($i = 1, \dots, n$) are L_j-Lipschitz and L_{h_i}-Lipschitz continuously differentiable and the gradients ∇j and ∇h_i ($i = 1, \dots, n$) exist for all $\boldsymbol{u} \in \mathcal{U}^N$.*

(ii) The stochastic gradient $\nabla_{\boldsymbol{u}} J$ defined by (3) *exists and there exists $M > 0$ such that:*

$$\mathbb{E}[\|\nabla_{\boldsymbol{u}} J(\boldsymbol{u}, \boldsymbol{\xi}) - \nabla j(\boldsymbol{u})\|^2_{\mathcal{G}^N}] \leq M^2 \quad \forall \boldsymbol{u} \in \mathcal{U}^N. \tag{7}$$

We begin our investigations with the following useful property.

Lemma 2.1. *Under Assumption 1 and assuming the gradients ∇j and ∇h_i ($i = 1, \dots, n$) exist, the iterates of Algorithm 1 satisfy*

$$\nabla_{\boldsymbol{u}} \mathcal{L}_A(\boldsymbol{u}^{k+1}, \boldsymbol{w}^k; \mu_k) = \nabla_{\boldsymbol{u}} \mathcal{L}(\boldsymbol{u}^{k+1}, \boldsymbol{\lambda}^{k+1}) \quad \textit{for all } k.$$

Proof. We have $\nabla \mathrm{dist}^2_{\boldsymbol{K}} = 2(\mathrm{Id}_{\mathbb{R}^n} - \pi_{\boldsymbol{K}})$; see [3, Corollary 12.31]. Let $f(\boldsymbol{u}) := \mathcal{L}_A(\boldsymbol{u}, \boldsymbol{w}; \mu)$. Then, the chain rule yields

$$(f_*)_{\boldsymbol{u}}\boldsymbol{v} = (j_*)_{\boldsymbol{u}}\boldsymbol{v} + \mu \sum_{i=1}^n \left(h_i(\boldsymbol{u}) + \frac{w_i}{\mu} - \pi_{K^i}\left(h_i(\boldsymbol{u}) + \frac{w_i}{\mu} \right) \right) ((h_i)_*)_{\boldsymbol{u}}\boldsymbol{v}.$$

From this, thanks to the identity (1), we can follow that

$$\nabla f(\boldsymbol{u}) = \nabla j(\boldsymbol{u}) + \mu \nabla \boldsymbol{h}(\boldsymbol{u})^\top \left(\boldsymbol{h}(\boldsymbol{u}) + \frac{\boldsymbol{w}}{\mu} - \pi_{\boldsymbol{K}}\left(\boldsymbol{h}(\boldsymbol{u}) + \frac{\boldsymbol{w}}{\mu} \right) \right),$$

and using the definition of $\boldsymbol{\lambda}^{k+1}$ from Algorithm 1, we obtain

$$\nabla_{\boldsymbol{u}} \mathcal{L}_A(\boldsymbol{u}^{k+1}, \boldsymbol{w}^k; \mu_k) = \nabla j(\boldsymbol{u}^{k+1}) + \nabla \boldsymbol{h}(\boldsymbol{u}^{k+1})^\top \boldsymbol{\lambda}^{k+1}.$$

Using the fact that $\nabla_{\boldsymbol{u}} \mathcal{L}(\boldsymbol{u}, \boldsymbol{\lambda}) = \nabla j(\boldsymbol{u}) + \nabla \boldsymbol{h}(\boldsymbol{u})^\top \boldsymbol{\lambda}$, we have proven the claim. □

Now, we turn to an efficiency estimate for the inner loop. First, we define the functions

$$F_k(\boldsymbol{u}, \boldsymbol{\xi}) := L_A(\boldsymbol{u}, \boldsymbol{w}^k, \boldsymbol{\xi}; \mu_k), \quad f_k(\boldsymbol{u}) := \mathbb{E}[L_A(\boldsymbol{u}, \boldsymbol{w}^k, \boldsymbol{\xi}; \mu_k)] = \mathcal{L}_A(\boldsymbol{u}, \boldsymbol{w}^k; \mu_k).$$

Recall the convention $\boldsymbol{\xi} \in \Xi$ being used in the definition of F_k and $\boldsymbol{\xi}\colon \Omega \to \Xi$ being used in the definition of f_k.

Lemma 2.2. *Suppose that Assumption 1 and Assumption 2 are satisfied and let $\hat{B}_k \subset \mathcal{U}^N$ be a bounded set such that $\mathrm{d}(\tilde{\boldsymbol{u}}, \boldsymbol{u}) \leq i(\mathcal{U}^N)$ for all $\tilde{\boldsymbol{u}}, \boldsymbol{u} \in \hat{B}_k$. Then, f_k is L_k-Lipschitz continuously differentiable with L_k depending on $L_j, L_{h_1}, \ldots, L_{h_n}$, and $\hat{B}_k$. Moreover, for all $\tilde{\boldsymbol{u}}, \boldsymbol{u} \in \hat{B}_k$ with $\boldsymbol{v} := \exp_{\boldsymbol{u}}^{-1}(\tilde{\boldsymbol{u}})$, we have*

$$f_k(\tilde{\boldsymbol{u}}) - f_k(\boldsymbol{u}) \leq \mathcal{G}^N(\nabla f_k(\boldsymbol{u}), \boldsymbol{v}) + \frac{L_k}{2}\|\boldsymbol{v}\|_{\mathcal{G}^N}^2. \tag{8}$$

Proof. Let $P_{1,0}$ denote the parallel transport as defined directly before Assumption 2 and set $g_i(\boldsymbol{u}) := h_i(\boldsymbol{u}) + \frac{w_i^k}{\mu_k} - \pi_{K_i}(h_i(\boldsymbol{u}) + \frac{w_i^k}{\mu_k})$. Since h_i is L_{h_i}-Lipschitz continuously differentiable and $\hat{B}_k$ is bounded, there exists $C_{i,k} > 0$ such that $\|\nabla h_i(\boldsymbol{u})\|_{\mathcal{G}^N} \leq C_{i,k}$. Now, we have

$$\begin{aligned}
&\Big\| \sum_{i=1}^n P_{1,0}\nabla h_i(\tilde{\boldsymbol{u}}) g_i(\tilde{\boldsymbol{u}}) - \nabla h_i(\boldsymbol{u}) g_i(\boldsymbol{u}) \Big\|_{\mathcal{G}^N} \\
&\quad \leq \sum_{i=1}^n \|P_{1,0}\nabla h_i(\tilde{\boldsymbol{u}}) - \nabla h_i(\boldsymbol{u})\|_{\mathcal{G}^N} |g_i(\boldsymbol{u})| + \|\nabla h_i(\boldsymbol{u})\|_{\mathcal{G}^N} |g_i(\tilde{\boldsymbol{u}}) - g_i(\boldsymbol{u})| \\
&\quad \leq \sum_{i=1}^n L_{h_i} \mathsf{d}(\boldsymbol{u}, \tilde{\boldsymbol{u}}) |g_i(\boldsymbol{u})| + C_{i,k} |g_i(\tilde{\boldsymbol{u}}) - g_i(\boldsymbol{u})| \\
&\quad \leq \sum_{i=1}^n L_{h_i} \mathsf{d}(\boldsymbol{u}, \tilde{\boldsymbol{u}}) |g_i(\boldsymbol{u})| + 2C_{i,k} |h_i(\tilde{\boldsymbol{u}}) - h_i(\boldsymbol{u})|,
\end{aligned} \tag{9}$$

where in the last step, we used the contraction property of the projection operator. Notice that

$$|h_i(\tilde{\boldsymbol{u}}) - h_i(\boldsymbol{u})| \leq C_i' \mathrm{d}(\tilde{\boldsymbol{u}}, \boldsymbol{u}) \tag{10}$$

for some $C_i' > 0$ since h_i is $\mathcal{C}^1$. Additionally, we have

$$|g_i(\boldsymbol{u})| \leq \Big| h_i(\boldsymbol{u}) + \frac{w_i^k}{\mu_k} \Big| \quad (i \in \mathcal{E}) \tag{11}$$

and

$$|g_i(\boldsymbol{u})| = \begin{cases} h_i(\boldsymbol{u}) + \frac{w_i^k}{\mu_k} & \text{if } h_i(\boldsymbol{u}) + \frac{w_i^k}{\mu_k} \geq 0, \\ 0 & \text{else} \end{cases} \quad (i \in \mathcal{I}). \tag{12}$$

Since $\hat{B}_k$ is bounded, (11) and (12) together imply that there exists $C_{i,k}'' > 0$ such that $|g_i(\boldsymbol{u})| \leq C_{i,k}''$. As a consequence of (9) and (10), we have

$$\Big\| \sum_{i=1}^n P_{1,0}\nabla h_i(\tilde{\boldsymbol{u}}) g_i(\tilde{\boldsymbol{u}}) - \nabla h_i(\boldsymbol{u}) g_i(\boldsymbol{u}) \Big\|_{\mathcal{G}^N} \leq \mathrm{d}(\boldsymbol{u}, \tilde{\boldsymbol{u}}) \sum_{i=1}^n L_{h_i} C_{i,k}'' + 2C_{i,k} C_i'.$$

Setting $\tilde{L}_{\boldsymbol{h},k} := \sum_{i=1}^n L_{h_i} C_{i,k}'' + 2C_{i,k} C_i'$, we have

$$\begin{aligned}
&\|P_{1,0}\nabla f_k(\tilde{\boldsymbol{u}}) - \nabla f_k(\boldsymbol{u})\|_{\mathcal{G}^N} \\
&\leq \|P_{1,0}\nabla j(\tilde{\boldsymbol{u}}) - \nabla j(\boldsymbol{u})\|_{\mathcal{G}^N} + \mu_k \Big\| \sum_{i=1}^n P_{1,0}\nabla h_i(\tilde{\boldsymbol{u}}) g_i(\tilde{\boldsymbol{u}}) - \nabla h_i(\boldsymbol{u}) g_i(\boldsymbol{u}) \Big\|_{\mathcal{G}^N} \\
&\leq (L_j + \mu_k \tilde{L}_{\boldsymbol{h},k}) \mathrm{d}(\tilde{\boldsymbol{u}}, \boldsymbol{u})
\end{aligned}$$

Therefore, f_k is L_k-Lipschitz with $L_k := L_j + \mu_k \tilde{L}_{\boldsymbol{h},k}$. Applying [10, Theorem 2.6], we obtain (8). □

Remark. *In the previous lemma, we introduced a bounded set $\hat{B}_k$. For the following results, we will need the existence of these sets containing the iterates almost surely within each k. Conditions ensuring boundedness can, e.g., be guaranteed by including constraints of the form $\boldsymbol{u} \in C \subset \mathcal{U}^N$ for some bounded set C, or growth conditions on the gradient in combination with a regularizer; see [12].*

Our first result concerning the convergence of Algorithm 1 handles the efficiency of the inner loop process, which corresponds to a stochastic gradient method that is randomly stopped after R_k iterations. We follow the arguments in [15, Corollary 3]. It is possible to choose non-constant step sizes t_{k_j}; see [15, Theorem 2], but for the sake of clarity we observe step sizes that are constant in the inner loop here.

To handle the analysis, we interpret R_k as a realization of a stopping time $\tau_k \colon \Omega \to \{1, \dots, N_k\}$. Let $\boldsymbol{\xi}^{k,j} := (\boldsymbol{\xi}^{k,j,1}, \dots, \boldsymbol{\xi}^{k,j,m_k})$ be the batch associated with iteration j for a given outer loop k and let $\mathcal{F}_{k,n} = \sigma(\boldsymbol{\xi}^{\ell,i} : \ell \in \{1, \dots, k\}, i \in \{1, \dots, n\})$ define the corresponding natural filtration. We define the filtration associated with the randomly stopped stochastic process by $\mathcal{F}^{\tau_k} = \{\mathcal{F}_{\ell, n \wedge \tau_k} : \ell \in \{1, \dots, k\}, n \in \{1, \dots, N_k\}\}$.

Theorem 2.1. *Suppose Assumption 1 and Assumption 2 are satisfied. Observe a fixed iteration k from Algorithm 1. Suppose the iterates $\{\boldsymbol{z}^{k,j}\}$ are a.s. contained in a bounded set $\hat{B}_k \subset \mathcal{U}^N$, where $\mathrm{d}(\boldsymbol{u}, \tilde{\boldsymbol{u}}) \leq i(\mathcal{U}^N)$ for all $\boldsymbol{u}, \tilde{\boldsymbol{u}} \in \hat{B}_k$. Then, if the step sizes $\{t_k\}$ satisfy $t_k = \alpha_k / L_k$ for $\alpha_k \in (0, 2)$, we have*

$$\mathbb{E}[\|\nabla f_k(\boldsymbol{u}^{k+1})\|_{\mathcal{G}^N}^2 | \mathcal{F}^{\tau_k}] \leq \frac{2L_k(f_k(\boldsymbol{u}^k) - f_k^*)}{(2\alpha_k - \alpha_k^2)N_k} + \frac{\alpha_k M^2}{(2 - \alpha_k)m_k}, \tag{13}$$

where $f_k^* := \inf_{u \in \hat{B}_k} f_k(\boldsymbol{u})$. Moreover, if $\hat{B}_\infty := \cup_{k=1}^\infty \hat{B}_k$ is bounded, $\mathrm{d}(\boldsymbol{u}, \tilde{\boldsymbol{u}}) \leq i(\mathcal{U}^N)$ for all $\boldsymbol{u}, \tilde{\boldsymbol{u}} \in \hat{B}_\infty$, the maximum iterations $\{N_k\}$ are chosen such that $N_k = \beta_k L_k$ for $\beta_k > 0$, and

$$\sum_{k=1}^\infty \frac{1}{(2\alpha_k - \alpha_k^2)\beta_k} + \frac{\alpha_k}{(2 - \alpha_k)m_k} < \infty, \tag{14}$$

then we have $\|\nabla f_k(\boldsymbol{u}^{k+1})\|_{\mathcal{G}^N} \to 0$ a.s. as $k \to \infty$.

Proof. Let k be fixed. We define $\delta^j := \frac{1}{m_k} \sum_{i=1}^{m_k} \nabla_{\boldsymbol{u}} F_k(\boldsymbol{z}^{k,j}, \boldsymbol{\xi}^{k,j,i}) - \nabla f_k(\boldsymbol{z}^{k,j})$. With $\boldsymbol{v}^j := \exp_{\boldsymbol{z}^{k,j}}^{-1}(\boldsymbol{z}^{k,j+1}) = -\frac{1}{L_k m_k} \sum_{i=1}^{m_k} \nabla_{\boldsymbol{u}} F_k(\boldsymbol{z}^{k,j}, \boldsymbol{\xi}^{k,j,i})$, Lemma 2.2 yields

$$\begin{aligned}
&f_k(\boldsymbol{z}^{k,j+1}) - f_k(\boldsymbol{z}^{k,j}) \\
&\leq -t_k \mathcal{G}^N \left(\nabla f_k(\boldsymbol{z}^{k,j}), \frac{1}{m_k} \sum_{i=1}^{m_k} \nabla_{\boldsymbol{u}} F_k(\boldsymbol{z}^{k,j}, \boldsymbol{\xi}^{k,j,i}) \right) \\
&\quad + \frac{L_k t_k^2}{2} \left\| \frac{1}{m_k} \sum_{i=1}^{m_k} \nabla_{\boldsymbol{u}} F_k(\boldsymbol{z}^{k,j}, \boldsymbol{\xi}^{k,j,i}) \right\|_{\mathcal{G}^N}^2 \\
&= -\frac{\alpha_k}{L_k} \|\nabla f_k(\boldsymbol{z}^{k,j})\|_{\mathcal{G}^N}^2 - \frac{\alpha_k}{L_k} \mathcal{G}^N(\nabla f_k(\boldsymbol{z}^{k,j}), \delta^j) \\
&\quad + \frac{\alpha_k^2}{2L_k} \left(\|\nabla f_k(\boldsymbol{z}^{k,j})\|_{\mathcal{G}^N}^2 + 2\mathcal{G}^N(\nabla f_k(\boldsymbol{z}^{k,j}), \delta^j) + \|\delta^j\|_{\mathcal{G}^N}^2 \right) \\
&= \left(-\frac{\alpha_k}{L_k} + \frac{\alpha_k^2}{2L_k} \right) \|\nabla f_k(\boldsymbol{z}^{k,j})\|_{\mathcal{G}^N}^2 + \left(-\frac{\alpha_k}{L_k} + \frac{\alpha_k^2}{L_k} \right) \mathcal{G}^N(\nabla f_k(\boldsymbol{z}^{k,j}), \delta^j) \\
&\quad + \frac{\alpha_k^2}{2L_k} \|\delta^j\|_{\mathcal{G}^N}^2.
\end{aligned}$$

Taking the sum with respect to j on both sides and rearranging, we obtain

$$\begin{aligned}\sum_{\ell=1}^{N_k}\|\nabla f_k(\boldsymbol{z}^{k,\ell})\|_{\mathcal{G}^N}^2 &\leq \frac{2L_k}{2\alpha_k-\alpha_k^2}(f_k(\boldsymbol{z}^{k,1})-f_k^*)\\ &+\frac{2(\alpha_k-1)}{2-\alpha_k}\sum_{\ell=1}^{N_k}\mathcal{G}^N(\nabla f_k(\boldsymbol{z}^{k,\ell}),\delta^\ell)+\frac{\alpha_k}{2-\alpha_k}\sum_{\ell=1}^{N_k}\|\delta^\ell\|_{\mathcal{G}^N}^2\end{aligned} \tag{15}$$

since $f_k^* \leq f_k(\boldsymbol{z}^{k,N_k+1})$ and $0<\alpha_k<2$. Since $\nabla_{\boldsymbol{u}}F_k$ is a stochastic gradient, we have

$$\mathbb{E}[\mathcal{G}^N(\nabla f_k(\boldsymbol{z}^{k,j}),\delta^j)|\mathcal{F}_{k,j}]=\mathcal{G}^N(\nabla f_k(\boldsymbol{z}^{k,j}),\mathbb{E}[\delta^j|\mathcal{F}_{k,j}])=0.$$

Notice that due to (7), we have

$$\begin{aligned}&\mathbb{E}\left[\|\nabla_{\boldsymbol{u}}F_k(\boldsymbol{z}^{k,j},\boldsymbol{\xi}^{k,j,i})-\nabla f_k(\boldsymbol{z}^{k,j})\|_{\mathcal{G}^N}^2|\mathcal{F}_{k,j}\right]\\ &=\mathbb{E}\left[\|\nabla_{\boldsymbol{u}}J(\boldsymbol{z}^{k,j},\boldsymbol{\xi}^{k,j,i})-\nabla j(\boldsymbol{z}^{k,j})\|_{\mathcal{G}^N}^2|\mathcal{F}_{k,j}\right]\\ &=\mathbb{E}\left[\|\nabla_{\boldsymbol{u}}J(\boldsymbol{z}^{k,j},\boldsymbol{\xi})-\nabla j(\boldsymbol{z}^{k,j})\|_{\mathcal{G}^N}^2\right]\leq M^2.\end{aligned} \tag{16}$$

With (16), we obtain

$$\begin{aligned}\mathbb{E}[\|\delta^j\|_{\mathcal{G}^N}^2|\mathcal{F}_{k,j}] &= \frac{1}{m_k^2}\mathbb{E}\left[\Big\|\sum_{i=1}^{m_k}\left(\nabla_{\boldsymbol{u}}F(\boldsymbol{z}^{k,j},\boldsymbol{\xi}^{k,j,i})-\nabla f_k(\boldsymbol{z}^{k,j})\right)\Big\|_{\mathcal{G}^N}^2\Big|\mathcal{F}_{k,j}\right]\\ &\leq \frac{1}{m_k^2}\sum_{i=1}^{m_k}\mathbb{E}\left[\|\nabla_{\boldsymbol{u}}F(\boldsymbol{z}^{k,j},\boldsymbol{\xi}^{k,j,i})-\nabla f_k(\boldsymbol{z}^{k,j})\|_{\mathcal{G}^N}^2|\mathcal{F}_{k,j}\right]\leq\frac{M^2}{m_k},\end{aligned} \tag{17}$$

where we used Jensen's inequality, the linearity of the expectation, and (16). Taking the expectation on both sides of (17), using (15), and using the tower property, we get the inequality

$$\sum_{\ell=1}^{N_k}\mathbb{E}[\|\nabla f_k(\boldsymbol{z}^{k,\ell})\|_{\mathcal{G}^N}^2]\leq\frac{2L_k(f_k(\boldsymbol{z}^{k,1})-f_k^*)}{2\alpha_k-\alpha_k^2}+\frac{\alpha_k}{2-\alpha_k}\frac{M^2N_k}{m_k}. \tag{18}$$

Due to the law of total expectation, we have

$$\begin{aligned}\mathbb{E}[\|\nabla f_k(\boldsymbol{z}^{k,R_k})\|_{\mathcal{G}^N}^2|\mathcal{F}^{\tau_k}] &= \mathbb{E}[\|\nabla f_k(\boldsymbol{z}^{k,\tau_k})\|_{\mathcal{G}^N}^2|\mathcal{F}^{\tau_k}]\\ &=\sum_{\ell=1}^{N_k}\mathbb{E}[\|\nabla f(\boldsymbol{z}^{k,\ell})\|_{\mathcal{G}^N}^2|\mathcal{F}_{k,\ell}]\mathbb{P}\{\tau_k=\ell\}\\ &=\frac{1}{N_k}\sum_{\ell=1}^{N_k}\mathbb{E}[\|\nabla f(\boldsymbol{z}^{k,\ell})\|_{\mathcal{G}^N}^2].\end{aligned}$$

Note that $f_k(\boldsymbol{z}^{k,R_k})=f_k(\boldsymbol{u}^{k+1})$ and $f_k(\boldsymbol{z}^{k,1})=f_k(\boldsymbol{u}^k)$. Returning to (18), we obtain

$$\mathbb{E}[\|\nabla f_k(\boldsymbol{u}^{k+1})\|_{\mathcal{G}^N}^2|\mathcal{F}^{\tau_k}]\leq\frac{2L_k(f_k(\boldsymbol{u}^k)-f_k^*)}{(2\alpha_k-\alpha_k^2)N_k}+\frac{\alpha_kM^2}{(2-\alpha_k)m_k},$$

so we have shown (13).

Now, to prove almost sure convergence, we first observe that if all iterates are contained in $\hat{B}_\infty$, we have

$$f_k(\boldsymbol{u}^k)-f_k^*\leq 2\sup_{\boldsymbol{u}\in\hat{B}_\infty}|f_k(\boldsymbol{u})|\leq C$$

for some $C > 0$ due to the assumed smoothness of f_k on $\mathcal{U}^N$. Taking the total expectation of (13), Markov's inequality in combination with Jensen's inequality gives

$$\begin{aligned}\mathbb{P}\{\|\nabla f_k(\boldsymbol{u}^{k+1})\|_{\mathcal{G}^N} \geq \varepsilon\} &\leq \varepsilon^{-2}\mathbb{E}[\|\nabla f_k(\boldsymbol{u}^{k+1})\|_{\mathcal{G}^N}^2] \\ &\leq \varepsilon^{-2}\left(\frac{2L_kC}{(2\alpha_k-\alpha_k^2)N_k} + \frac{\alpha_k M^2}{(2-\alpha_k)m_k}\right).\end{aligned}$$

Since $N_k = \beta_k L_k$ and (14) holds, the infinite sum of the right-hand side is finite for every $\varepsilon > 0$, implying the almost sure convergence of $\{\|\nabla f_k(\boldsymbol{u}^{k+1})\|_{\mathcal{G}^N}\}$ to zero. □

For the choice $t_k = 1/L_k$, the efficiency estimate (13) evidently simplifies to $\mathbb{E}[\|\nabla f_k(\boldsymbol{u}^{k+1})\|_{\mathcal{G}^N}^2] \leq \frac{2L_k(f_k(\boldsymbol{u}^k)-f_k^*)}{N_k} + \frac{M^2}{m_k}$. In the next section, we will investigate optimality of the solution in the limit as k is taken to infinity. Since the Lipschitz constant L_k has a potential to be unbounded due to the penalty term μ_k, the maximal number of iterations N_k needs to be balanced appropriately in this case. To obtain almost sure convergence, we required $N_k = \beta_k L_k$ for $\beta_k > 0$. Alternatively, if it can be guaranteed that L_k is bounded for all k (for instance by bounding μ_k), then one could (asymptotically) choose $t_k = \alpha_k/L$ with $L = \sup_k L_k$. Regarding complexity, it is possible to establish the inner loop's complexity as argued in [15, Section 4.2]. We define a (ε_k, η_k)-solution to the problem $\min_{\boldsymbol{u}\in\mathcal{U}^N}\{f_k(\boldsymbol{u}) = \mathbb{E}[F_k(\boldsymbol{u},\boldsymbol{\xi})]\}$ as the point $\hat{\boldsymbol{u}}$ that satisfies $\mathbb{P}\{\|\nabla f_k(\hat{\boldsymbol{u}})\|_{\mathcal{G}^N}^2 \leq \varepsilon_k\} \geq 1-\eta_k$. Ignoring some constants, for the choice $t_k = 1/L_k$, the complexity can be bounded by $\mathcal{O}\left((\eta_k\varepsilon_k)^{-1} + M^2\eta_k^{-2}\varepsilon_k^{-2}\right)$.

2.4 Convergence of outer loop

In the final part of this section, we analyze the behavior of the outer loop of Algorithm 1 adapting arguments from [29, 18]. We define an optimality measure and its induced sequence by

$$r(\boldsymbol{u},\boldsymbol{\lambda}) = \|\nabla_{\boldsymbol{u}}\mathcal{L}(\boldsymbol{u},\boldsymbol{\lambda})\|_{\mathcal{G}^N} + \|\boldsymbol{h}(\boldsymbol{u}) - \pi_{\boldsymbol{K}}(\boldsymbol{h}(\boldsymbol{u})+\boldsymbol{\lambda})\|_2, \quad r_k := r(\boldsymbol{u}^k,\boldsymbol{\lambda}^k)$$

and make the following assumptions on iterates induced by Algorithm 1.

Assumption 3. *We assume that (i) the sequence $\{\boldsymbol{u}^k\}$ is a.s. contained in a bounded set $\hat{B}_\infty$ such that $\mathrm{d}(\boldsymbol{u},\tilde{\boldsymbol{u}}) \leq i(\mathcal{U}^N)$ for all $\boldsymbol{u},\tilde{\boldsymbol{u}} \in \hat{B}_\infty$,*
(ii) $\|\nabla_{\boldsymbol{u}}\mathcal{L}_A(\boldsymbol{u}^{k+1},\boldsymbol{w}^k;\mu_k)\|_{\mathcal{G}^N} \to 0$ a.s. as $k\to\infty$,
(iii) $\{(\boldsymbol{u}^k,\boldsymbol{\lambda}^k)\}$ converges a.s. to the set of KKT points,
(iv) for k sufficiently large, we have $\boldsymbol{w}^k = \boldsymbol{\lambda}^k$.

Note that Theorem 2.1 implies Assumption 3(ii). In the absence of constraint qualifications, one can still work with asymptotic KKT (AKKT) conditions; under certain conditions, it can even be shown that they are necessary conditions (see, e.g., [18, Theorem 5.3]). We will say that a feasible point $\hat{\boldsymbol{u}}$ satisfies the AKKT conditions if there exists a sequence $\{\boldsymbol{u}^k\}$ such that $\mathrm{d}(\boldsymbol{u}^k,\hat{\boldsymbol{u}}) \to 0$ and a sequence $\{\boldsymbol{\lambda}^k\}$ contained in the dual cone $\boldsymbol{K}^\oplus := \{\boldsymbol{y}\in\mathbb{R}^n\colon \boldsymbol{y}^\top\boldsymbol{k} \geq 0\,\forall \boldsymbol{k}\in\boldsymbol{K}\}$ such that

$$\|\nabla j(\boldsymbol{u}^k) + \nabla\boldsymbol{h}(\boldsymbol{u}^k)^\top\boldsymbol{\lambda}^k\|_{\mathcal{G}^N} \to 0 \quad \text{and} \quad \pi_{\boldsymbol{K}}(-\boldsymbol{h}(\boldsymbol{u}^k))^\top\boldsymbol{\lambda}^k \to 0 \tag{19}$$

as $k\to\infty$.

A fundamental difference in the stochastic variant of the augmented Lagrangian method is that limit points, as limits of the stochastic process $(\boldsymbol{u}^k,\boldsymbol{\lambda}^k)$, are random. In the following, we will consider a fixed limit point $(\hat{\boldsymbol{u}},\hat{\boldsymbol{\lambda}})$ and the corresponding set of paths converging to it. This motivates the definition of the set

$$E_{\hat{\boldsymbol{u}},\hat{\boldsymbol{\lambda}}} := \{\omega : (\boldsymbol{u}^k(\omega),\boldsymbol{\lambda}^k(\omega)) \to (\hat{\boldsymbol{u}},\hat{\boldsymbol{\lambda}}) \quad \text{a.s. on a subsequence}\}. \tag{20}$$

Theorem 2.2. *Suppose Assumption 1–Assumption 3(i)-(ii) are satisfied. Let* $E := \{\omega \in \Omega : \mu_k(\omega)$ *is a.s. bounded*$\}$*. Then,* $\{\boldsymbol{\lambda}^k(\omega)\}$ *is a.s. bounded on* E *and any limit point* $(\hat{\boldsymbol{u}}, \hat{\boldsymbol{\lambda}})$ *of the random sequence* $\{(\boldsymbol{u}^k(\omega), \boldsymbol{\lambda}^k(\omega)) : \omega \in E, k \in \mathbb{N}\}$ *is a KKT point. On the set* $\Omega\backslash E$*, if a limit point* $\hat{\boldsymbol{u}}$ *is feasible, then it is a AKKT point.*

Proof. We will make arguments in several parts.

Part 1: Bounded μ_k. We first show that the sequence $\{\boldsymbol{\lambda}^k\}$ is a.s. bounded. Let $\boldsymbol{v}^{k+1} := \boldsymbol{h}(\boldsymbol{u}^{k+1}) + \frac{\boldsymbol{w}^k}{\mu_k}$ and $\boldsymbol{y}^{k+1} := \pi_{\boldsymbol{K}}(\boldsymbol{v}^{k+1})$. By definition of $\boldsymbol{\lambda}^k$, we have

$$\boldsymbol{h}(\boldsymbol{u}^{k+1}) = \frac{1}{\mu_k}(\boldsymbol{\lambda}^{k+1} - \boldsymbol{w}^k) + \boldsymbol{y}^{k+1}. \tag{21}$$

Now, observe that the boundedness of $\{\mu_k\}$ on E implies that there exists a maximal iterate $\bar{k}$ in Algorithm 1 such that $H_{k+1} \leq \tau H_k \leq \tau M$ is satisfied for every $k \geq \bar{k}$ and some $M > 0$. This M exists since $\boldsymbol{h}$ is $\mathcal{C}^1$ and $\boldsymbol{u}^k$, $\boldsymbol{w}^k$, and μ_k are all bounded by assumption. In particular, $H_k \to 0$ as $k \to \infty$ on E. In turn, (21) combined with the definition of H_k implies the a.s. convergence of $\|\boldsymbol{\lambda}^{k+1} - \boldsymbol{w}^k\|_2/\mu_k$ to zero, in turn implying $\|\boldsymbol{\lambda}^{k+1} - \boldsymbol{w}^k\|_2 \to 0$ for $k \to 0$. The boundedness of $\boldsymbol{w}^k$ guaranteed by Algorithm 1 means therefore that $\{\boldsymbol{\lambda}^k\}$ is bounded on E.

Now, we prove that for any $\boldsymbol{y} \in \boldsymbol{K}$, there exists a nonnegative sequence γ_k converging to zero and such that

$$(\boldsymbol{y} - \boldsymbol{h}(\boldsymbol{u}^k))^\top \boldsymbol{\lambda}^k \leq \gamma_k, \quad \omega \in E, k \in \mathbb{N}. \tag{22}$$

With [3, Theorem 3.14], the projection formula

$$(\boldsymbol{v}^{k+1} - \boldsymbol{y}^{k+1})^\top (\boldsymbol{y}^{k+1} - \boldsymbol{y}) \geq 0$$

holds for all $\boldsymbol{y} \in \boldsymbol{K}$, implying that $\boldsymbol{\lambda}^{k+1} = \mu_{k+1}(\boldsymbol{v}^{k+1} - \boldsymbol{y}^{k+1}) \in N_{\boldsymbol{K}}(\boldsymbol{y}^{k+1})$. Now, using $\boldsymbol{\lambda}^{k+1} \in N_{\boldsymbol{K}}(\boldsymbol{y}^{k+1})$ and (21), we have

$$\begin{aligned}
(\boldsymbol{y} - \boldsymbol{h}(\boldsymbol{u}^{k+1}))^\top \boldsymbol{\lambda}^{k+1} &= \left(\boldsymbol{y} - \frac{1}{\mu_k}(\boldsymbol{\lambda}^{k+1} - \boldsymbol{w}^k) - \boldsymbol{y}^{k+1}\right)^\top \boldsymbol{\lambda}^{k+1} \\
&\leq \frac{1}{\mu_k}((\boldsymbol{w}^k)^\top \boldsymbol{\lambda}^{k+1} - \|\boldsymbol{\lambda}^{k+1}\|_2^2) \\
&= (\boldsymbol{y}^{k+1} - \boldsymbol{h}(\boldsymbol{u}^{k+1}))^\top \boldsymbol{\lambda}^{k+1} =: \gamma_{k+1}.
\end{aligned}$$

We have shown (22). That $\{\gamma_k\}$ is a.s. a null sequence follows from the fact that $\|\boldsymbol{\lambda}^{k+1} - \boldsymbol{w}^k\|_2/\mu_k$ a.s. converges to zero.

Consider a subsequence of $\{(\boldsymbol{u}^k(\omega), \boldsymbol{\lambda}^k(\omega))\}$ that converge to a limit point $(\hat{\boldsymbol{u}}, \hat{\boldsymbol{\lambda}})$ for a fixed $\omega \in E_{\hat{\boldsymbol{u}}, \hat{\boldsymbol{\lambda}}}$. We will prove that the limit point satisfies the KKT conditions (4). Continuity of $\nabla_{\boldsymbol{u}}\mathcal{L}$ gives $\lim_{k\to\infty} \nabla_{\boldsymbol{u}}\mathcal{L}(\boldsymbol{u}^k(\omega), \boldsymbol{\lambda}^k(\omega)) = \nabla_{\boldsymbol{u}}\mathcal{L}(\hat{\boldsymbol{u}}, \hat{\boldsymbol{\lambda}})$ and $\|\nabla_{\boldsymbol{u}}\mathcal{L}(\hat{\boldsymbol{u}}, \hat{\boldsymbol{\lambda}})\|_{\mathcal{G}^n} = 0$ due to Assumption 3(ii). By definition, $\nabla_{\boldsymbol{u}}\mathcal{L}(\hat{\boldsymbol{u}}, \hat{\boldsymbol{\lambda}}) \in T_{\hat{\boldsymbol{u}}}\mathcal{U}^N$, and the only element in $T_{\hat{\boldsymbol{u}}}\mathcal{U}^N$ having norm zero is $0_{\hat{\boldsymbol{u}}}$, thus (4a) is fulfilled. Since $\alpha_k \to 0$ a.s., we have that $(\boldsymbol{y} - \boldsymbol{h}(\hat{\boldsymbol{u}}))^\top \hat{\boldsymbol{\lambda}} \leq 0$ for all $\boldsymbol{y} \in \boldsymbol{K}$, implying that $\hat{\boldsymbol{\lambda}} \in N_{\boldsymbol{K}}(\boldsymbol{h}(\hat{\boldsymbol{u}}))$. This immediately implies (4b)–(4c).

Part 2: Unbounded μ_k. Consider a fixed $\omega \in \Omega\backslash E$ and a sequence $\{\boldsymbol{u}^k(\omega)\}$ such that (possibly on a subsequence that we do not relabel) $\mathrm{d}(\boldsymbol{u}^k(\omega), \hat{\boldsymbol{u}}) \to 0$ as $k \to \infty$. Assumption 3(ii) gives the first AKKT condition in (19). It remains to prove that $\pi_{\boldsymbol{K}}(-\boldsymbol{h}(\boldsymbol{u}^k(\omega)))^\top \boldsymbol{\lambda}^k(\omega) \to 0$. Now, we define

$$\boldsymbol{p}^k(\omega) := (\mu_k(\omega)\boldsymbol{h}(\boldsymbol{u}^{k+1}(\omega)) + \boldsymbol{w}^k(\omega))^\top \pi_{\boldsymbol{K}}(-\boldsymbol{h}(\boldsymbol{u}^{k+1}(\omega))).$$

For readability, we will suppress the dependence on ω. Since

$$\boldsymbol{\lambda}^{k+1} = \mu_k \left(\boldsymbol{h}(\boldsymbol{u}^{k+1}) + \frac{\boldsymbol{w}^k}{\mu_k} - \pi_{\boldsymbol{K}} \left(\boldsymbol{h}(\boldsymbol{u}^{k+1}) + \frac{\boldsymbol{w}^k}{\mu_k} \right) \right)$$

it is evidently enough to prove $\boldsymbol{p}^k \to 0$, since due to the contraction property of the projection, we have $\pi_{\boldsymbol{K}}(\boldsymbol{a}^k)^\top \boldsymbol{b}^k \to 0$ implies $\pi_{\boldsymbol{K}}(\boldsymbol{a}^k)^\top \pi_{\boldsymbol{K}}(\boldsymbol{b}^k) \to 0$ for any $\boldsymbol{a}^k, \boldsymbol{b}^k \in \mathbb{R}^n$. Note that at least on a subsequence, we have $\boldsymbol{h}(\boldsymbol{u}^{k+1}) \to \boldsymbol{h}(\hat{\boldsymbol{u}})$ and $|\boldsymbol{h}(\boldsymbol{u}^k)|$ is bounded.

Consider first the case that $h_i(\hat{\boldsymbol{u}}) < 0$. Then $\boldsymbol{h}(\boldsymbol{u}^{k+1}) \to \boldsymbol{h}(\hat{\boldsymbol{u}})$ implies that $w_i^k + \mu_k h_i(\boldsymbol{u}^{k+1}) < 0$ for k sufficiently large, implying $\boldsymbol{p}^k \to 0$.

Consider now the case that $h_i(\hat{\boldsymbol{u}}) = 0$. For a fixed k, if $h_i(\boldsymbol{u}^{k+1}) \geq 0$ then $\boldsymbol{p}^k = 0$. Else if $h_i(\boldsymbol{u}^{k+1}) < 0$, then $p_i^k = (\mu_k h_i(\boldsymbol{u}^{k+1}) + w_i^k)\pi_{\boldsymbol{K}}(-h_i(\boldsymbol{u}^{k+1})) \leq w_i^k |h_i(\boldsymbol{u}^{k+1})|$. If $h_i(\boldsymbol{u}^{k+1}) < 0$ infinitely many times, then $w_i^k |h_i(\boldsymbol{u}^{k+1})| \to 0$, meaning $\boldsymbol{p}^k \to 0$.

Since $\boldsymbol{p}^k$ in both cases converges to zero and $\omega \in \Omega \backslash E$ was arbitrary, we have proven the claim. □

We now turn to local convergence statements. In the spirit of a local argument, we restrict our investigations to the study around a limit point for *only those realizations converging to it*. Again, we consider the set $E_{\hat{\boldsymbol{u}},\hat{\boldsymbol{\lambda}}}$ defined in (20).

Lemma 2.3. *Suppose Assumptions 1–3 hold. Let $(\hat{\boldsymbol{u}}, \hat{\boldsymbol{\lambda}})$ be a limit point satisfying for some $c_1, c_2 > 0$*

$$c_1 r(\boldsymbol{u}, \boldsymbol{\lambda}) \leq \mathrm{d}(\boldsymbol{u}, \hat{\boldsymbol{u}}) + \|\boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}\|_2 \leq c_2 r(\boldsymbol{u}, \boldsymbol{\lambda}) \tag{23}$$

for all $(\boldsymbol{u}, \boldsymbol{\lambda})$ with $\boldsymbol{u}$ near $\hat{\boldsymbol{u}}$ and $r(\boldsymbol{u}, \boldsymbol{\lambda})$ sufficiently small. Then we have for sufficiently large k

$$\left(1 - \frac{c_2}{\mu_k}\right) r_{k+1} \leq \|\nabla_{\boldsymbol{u}} \mathcal{L}_A(\boldsymbol{u}^{k+1}, \boldsymbol{w}^k; \mu_k)\|_{\mathcal{G}^N} + \frac{c_2}{\mu_k} r_k \quad \textit{a.s. on } E_{\hat{\boldsymbol{u}},\hat{\boldsymbol{\lambda}}}.$$

Proof. We have using Lemma 2.1 and $\boldsymbol{w}^k = \boldsymbol{\lambda}^k$ that

$$r_{k+1} = \|\nabla_{\boldsymbol{u}} \mathcal{L}_A(\boldsymbol{u}^{k+1}, \boldsymbol{\lambda}^k; \mu_k)\|_{\mathcal{G}^N} + \|\boldsymbol{h}(\boldsymbol{u}^{k+1}) - \pi_{\boldsymbol{K}}(\boldsymbol{h}(\boldsymbol{u}^{k+1}) + \boldsymbol{\lambda}^{k+1})\|_2. \tag{24}$$

Let $\boldsymbol{v}^{k+1} := \boldsymbol{h}(\boldsymbol{u}^{k+1}) + \frac{\boldsymbol{w}^k}{\mu_k}$ and $\boldsymbol{y}^{k+1} := \pi_{\boldsymbol{K}}(\boldsymbol{v}^{k+1})$. Then it follows that

$$\|\boldsymbol{y}^{k+1} - \pi_{\boldsymbol{K}}(\boldsymbol{y}^{k+1} + \boldsymbol{\lambda}^{k+1})\|_2 = 0 \tag{25}$$

since $\boldsymbol{\lambda}^{k+1} \in N_{\boldsymbol{K}}(\boldsymbol{y}^{k+1})$ as argued in Part 1 of the proof of Theorem 2.2. Note that $\mathrm{Id}_{\mathbb{R}^n} - \pi_{\boldsymbol{K}}$ is (firmly) nonexpansive (cf. [3, Prop. 12.27]). It is an easy exercise to deduce that the mapping $\boldsymbol{y} \mapsto \boldsymbol{y} - \pi_{\boldsymbol{K}}(\boldsymbol{y} + \boldsymbol{\lambda}^{k+1})$ is nonexpansive as well, from which we can conclude

$$\begin{aligned} &\Big| \|\boldsymbol{h}(\boldsymbol{u}^{k+1}) - \pi_{\boldsymbol{K}}(\boldsymbol{h}(\boldsymbol{u}^{k+1}) + \boldsymbol{\lambda}^{k+1})\|_2 - \|\boldsymbol{y}^{k+1} - \pi_{\boldsymbol{K}}(\boldsymbol{y}^{k+1} + \boldsymbol{\lambda}^{k+1})\|_2 \Big| \\ &\quad \leq \|\boldsymbol{h}(\boldsymbol{u}^{k+1}) - \pi_{\boldsymbol{K}}(\boldsymbol{h}(\boldsymbol{u}^{k+1}) + \boldsymbol{\lambda}^{k+1}) - \boldsymbol{y}^{k+1} + \pi_{\boldsymbol{K}}(\boldsymbol{y}^{k+1} + \boldsymbol{\lambda}^{k+1})\|_2 \\ &\quad \leq \|\boldsymbol{h}(\boldsymbol{u}^{k+1}) - \boldsymbol{y}^{k+1}\|_2. \end{aligned} \tag{26}$$

Using the definition of $\boldsymbol{y}^{k+1}$ and $\boldsymbol{w}^k = \boldsymbol{\lambda}^k$, notice that

$$\begin{aligned} &\|\boldsymbol{h}(\boldsymbol{u}^{k+1}) - \pi_{\boldsymbol{K}}(\boldsymbol{h}(\boldsymbol{u}^{k+1}) + \boldsymbol{\lambda}^{k+1})\|_2 \\ &\leq \|\boldsymbol{h}(\boldsymbol{u}^{k+1}) - \pi_{\boldsymbol{K}}(\boldsymbol{h}(\boldsymbol{u}^{k+1}) + \boldsymbol{\lambda}^k/\mu_k)\|_2 \\ &= \frac{1}{\mu_k} \|\mu_k \boldsymbol{h}(\boldsymbol{u}^{k+1}) + \boldsymbol{\lambda}^k - \mu_k \pi_{\boldsymbol{K}}(\boldsymbol{h}(\boldsymbol{u}^{k+1}) + \boldsymbol{\lambda}^k/\mu_k) - \boldsymbol{\lambda}^k\|_2 \\ &= \frac{1}{\mu_k} \|\boldsymbol{\lambda}^{k+1} - \boldsymbol{\lambda}^k\|_2. \end{aligned}$$

Returning to (24), we obtain

$$r_{k+1} \leq \|\nabla_{\boldsymbol{u}} \mathcal{L}_A(\boldsymbol{u}^{k+1}, \boldsymbol{\lambda}^k; \mu_k)\|_{\mathcal{G}^N} + \frac{1}{\mu_k}\Big(\|\boldsymbol{\lambda}^{k+1} - \hat{\boldsymbol{\lambda}}\|_2 + \|\boldsymbol{\lambda}^k - \hat{\boldsymbol{\lambda}}\|_2\Big). \tag{27}$$

Since $\lim_{k\to\infty} \mathrm{d}(\boldsymbol{u}^k, \hat{\boldsymbol{u}}) = 0$ a.s. on $E_{\hat{\boldsymbol{u}},\hat{\boldsymbol{\lambda}}}$, then for any $\varepsilon > 0$ there exists $\bar{k}$ such that $\mathrm{d}(\boldsymbol{u}^k, \hat{\boldsymbol{u}}) < \varepsilon$ for all $k \geq \bar{k}$ a.s. Possibly choosing $\bar{k}$ even larger, Assumption 3 combined with the positive injectivity radius further implies $\|\boldsymbol{\lambda}^k(\omega) - \hat{\boldsymbol{\lambda}}\|_2 \leq c_2 r_k$ for almost all $\omega \in E_{\hat{\boldsymbol{u}},\hat{\boldsymbol{\lambda}}}$. Using (27), we conclude that for almost all $\omega \in E_{\hat{\boldsymbol{u}},\hat{\boldsymbol{\lambda}}}$,

$$r_{k+1} \leq \|\nabla_{\boldsymbol{u}} \mathcal{L}_A(\boldsymbol{u}^{k+1}(\omega), \boldsymbol{\lambda}^k(\omega); \mu_k(\omega))\|_{\mathcal{G}^N} + \frac{1}{\mu_k}(c_2 r_{k+1} + c_2 r_k),$$

for k large enough. Rearranging terms proves the claim. □

We are now ready to show the local rate of convergence.

Theorem 2.3. *Under the same assumptions as Lemma 2.3, assume further that*

$$\|\nabla_{\boldsymbol{u}} \mathcal{L}_A(\boldsymbol{u}^{k+1}, \boldsymbol{\lambda}^k; \mu_k)\|_{\mathcal{G}^N} = o(r_k).$$

Then

1) *If for all $q \in (0,1)$ there exists $\hat{\mu}_q > 0$ such that if $\mu_k \geq \hat{\mu}_q$ for k sufficiently large, then $(\boldsymbol{u}^k, \boldsymbol{\lambda}^k) \to (\hat{\boldsymbol{u}}, \hat{\boldsymbol{\lambda}})$ a.s. on $E_{\hat{\boldsymbol{u}},\hat{\boldsymbol{\lambda}}}$ at a linear rate.*

2) *If $\mu_k \to \infty$, then $(\boldsymbol{u}^k, \boldsymbol{\lambda}^k) \to (\hat{\boldsymbol{u}}, \hat{\boldsymbol{\lambda}})$ a.s. on $E_{\hat{\boldsymbol{u}},\hat{\boldsymbol{\lambda}}}$ at a superlinear rate.*

Proof. Note that for k large enough, we have $\boldsymbol{w}^k = \boldsymbol{\lambda}^k$ and Lemma 2.3 gives

$$\left(1 - \frac{c_2}{\mu_k}\right) r_{k+1} \leq \|\nabla_{\boldsymbol{u}} \mathcal{L}_A(\boldsymbol{u}^{k+1}, \boldsymbol{w}^k; \mu_k)\|_{\mathcal{G}^N} + \frac{c_2}{\mu_k} r_k = o(r_k) + \frac{c_2}{\mu_k} r_k.$$

Taking μ_k large enough gives $r_{k+1} \leq \frac{\mu_k}{\mu_k - c_2}\left(o(r_k) + \frac{c_2}{\mu_k} r_k\right)$. This implies

$$\frac{r_{k+1}}{r_k} \leq \frac{\mu_k}{\mu_k - c_2}\left(o(1) + \frac{c_2}{\mu_k}\right) = \frac{c_2}{\mu_k - c_2} + o(1).$$

Thanks to the error bound (23), we get the corresponding rates for $\{(\boldsymbol{u}^k, \boldsymbol{\lambda}^k)\}$. □

In practice, the assumption $\|\nabla_{\boldsymbol{u}} \mathcal{L}_A(\boldsymbol{u}^{k+1}, \boldsymbol{\lambda}^k; \mu_k)\|_{\mathcal{G}^N} = o(r_k)$ is difficult to implement since one can only work with estimates $\hat{f}_k \approx \mathbb{E}[L_A(\boldsymbol{u}^{k+1}, \boldsymbol{\lambda}^k, \boldsymbol{\xi}; \mu_k)] = \mathcal{L}_A(\boldsymbol{u}^{k+1}, \boldsymbol{\lambda}^k; \mu_k)$. However, we have a convergence rate guaranteed in expectation by (13), which can be used to choose appropriate sequences for N_k and m_k. A possible heuristic is shown in the following section.

3 Application and numerical results

In this section, we present an application to a two-dimensional fluid-mechanical problem to demonstrate the algorithm. We denote the hold-all domain as $D = D(\boldsymbol{u})$, which is partitioned into $N+1$ disjoint subdomains $D_1, \dots, D_{N+1}$, where D_{N+1} represents the subdomain in which fluid is allowed to flow, and the other sets are obstacles around which the fluid is supposed to flow. The subdomain boundaries are defined as $\partial D_1 = u_1, \dots, \partial D_N = u_N$, and $\partial D_{N+1} = \Gamma \cup u_1 \cup \cdots \cup u_N = \Gamma \cup \boldsymbol{u}$, where Γ is the outer boundary that is fixed and split into two disjoint parts $\mathit{\Gamma}_D$ and $\mathit{\Gamma}_N$ representing the Dirichlet and Neumann boundary, respectively.

The shape space we consider in the numerical experiments is the product space of plane unparametrized curves, i.e., $\mathcal{U}^N = B_e^N(S^1, \mathbb{R}^2)$. The shape space $B_e(S^1, \mathbb{R}^2)$ is defined as the orbit space of $\mathrm{Emb}(S^1, \mathbb{R}^2)$ under the action by composition from the right by the Lie group $\mathrm{Diff}(S^1)$, meaning $B_e(S^1, \mathbb{R}^2) := \mathrm{Emb}(S^1, \mathbb{R}^2)/\mathrm{Diff}(S^1)$ (cf., e.g., [23]). Here, $\mathrm{Emb}(S^1, \mathbb{R}^2)$ denotes the set of all embeddings from the unit circle S^1 into $\mathbb{R}^2$, and $\mathrm{Diff}(S^1)$ is the set of all diffeomorphisms from S^1 into itself. In [20], it is proven that the shape space $B_e(S^1, \mathbb{R}^2)$ is a smooth manifold; together with appropriate inner products, it is even a Riemannian manifold. In our numerical experiments, we choose the Steklov–Poincaré metric defined in [25]. Originally, it is defined as a mapping from Sobolev spaces. To define a metric on $B_e(S^1, \mathbb{R}^2)$, the Steklov–Poincaré metric is restricted to a mapping from the tangent spaces, i.e., $T_u B_e(S^1, \mathbb{R}^2) \times T_u B_e(S^1, \mathbb{R}^2) \to \mathbb{R}$, where $T_u B_e(S^1, \mathbb{R}^2) \cong \{h \colon h = \alpha \boldsymbol{n},\ \alpha \in \mathcal{C}^\infty(S^1)\}$. Of course, one can choose a different metric on the shape space to represent the shape gradient. We focus on the Steklov–Poincaré metric due to its advantages in combination with the computational mesh (cf. [28, 25]).

The physical system on D is described by the Stokes equations under uncertainty. Note that here, flow is modeled on the domain D instead of D_{N+1}. This is done (in view of the tracking-type functional) to produce a shape derivative on the entire domain. Let $V(D) = \{\boldsymbol{q} \in H^1(D, \mathbb{R}^2) \colon \boldsymbol{q}|_{\mathit{\Gamma}_D \cup \boldsymbol{u}} = \boldsymbol{0}\}$ denote the function space associated to the velocity for a fixed domain D. We neglect volume forces and consider a deterministic viscosity of the fluid. Inflow $\boldsymbol{g}$ on parts of the Dirichlet boundary is assumed to be uncertain and is modeled as a random field $\boldsymbol{g} \colon D \times \Xi \to \mathbb{R}^2$ with regularity $\boldsymbol{g} \in L^2_{\mathbb{P}}(\Xi, H^1(D, \mathbb{R}^2))$ and depending on $\boldsymbol{\xi} \colon \Omega \to \Xi \subset \mathbb{R}^m$. We will use the abbreviation $\boldsymbol{g}_{\boldsymbol{\xi}} = \boldsymbol{g}(\cdot, \boldsymbol{\xi})$. For each realization $\boldsymbol{\xi}$, consider Stokes flow in weak form: find $\boldsymbol{q}_{\boldsymbol{\xi}} \in H^1(D, \mathbb{R}^2)$ and $p_{\boldsymbol{\xi}} \in L^2(D)$ such that $\boldsymbol{q}_{\boldsymbol{\xi}} - \boldsymbol{g}_{\boldsymbol{\xi}} \in V(D)$ and

$$\int_D \nabla \boldsymbol{q}_{\boldsymbol{\xi}} : \nabla \boldsymbol{\varphi} - p_{\boldsymbol{\xi}} \operatorname{div} \boldsymbol{\varphi} \,\mathrm{d}\boldsymbol{x} = 0 \quad \forall \boldsymbol{\varphi} \in V(D), \tag{28a}$$

$$\int_D \psi \operatorname{div} \boldsymbol{q}_{\boldsymbol{\xi}} \,\mathrm{d}\boldsymbol{x} = 0 \quad \forall \psi \in L^2(D). \tag{28b}$$

Here, $\boldsymbol{A} : \boldsymbol{B} = \sum_{j=1}^d \sum_{k=1}^d A_{jk} B_{jk}$ for two matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{d \times d}$. The gradient and divergence operators ∇ and div act with respect to the spatial variable only with $\boldsymbol{\xi}$ acting as a parameter.

For each shape u_i, $i = 1, \dots, N$, we introduce one inequality constraint for a constrained volume, see equation (30a) and one inequality constraint for a constrained perimeter, see equation (30b). The volume of the domain D_i is given by $\mathrm{vol}(D_i) = \int_{D_i} 1 \,\mathrm{d}\boldsymbol{x}$ and the perimeter of u_i is given by $\mathrm{peri}(u_i) = \int_{u_i} 1 \,\mathrm{d}\boldsymbol{s}$. Now, we suppose there is a deterministic target velocity $\bar{\boldsymbol{q}}$ to be reached on the domain D. We would like to determine the optimal placement of shapes that come closest on average to this velocity field. More precisely, we solve the problem

$$\min_{\boldsymbol{u} \in B_e^N(S^1, \mathbb{R}^2)} \left\{ j(\boldsymbol{u}) = \int_\Omega \int_D \| \boldsymbol{q}_{\boldsymbol{\xi}(\omega)}(\boldsymbol{x}) + \boldsymbol{g}_{\boldsymbol{\xi}(\omega)}(\boldsymbol{x}) - \bar{\boldsymbol{q}}(\boldsymbol{x}) \|_2^2 \,\mathrm{d}\boldsymbol{x} \,\mathrm{d}\mathbb{P}(\omega) \right\} \tag{29}$$

subject to (28) and

$$\operatorname{vol}(D_i) \geq \underline{\mathcal{V}}_i \qquad \forall i = 1, \dots, N, \tag{30a}$$

$$\operatorname{peri}(u_i) \leq \overline{\mathcal{P}}_i \qquad \forall i = 1, \dots, N. \tag{30b}$$

We note that a deterministic model using a tracking-type functional in combination with Stokes flow has been studied in [7].

KKT conditions. In the following, we formulate the necessary optimality conditions to the model problem defined by (28)–(30). We define $\boldsymbol{h}\colon \mathcal{U} \to \mathbb{R}^{2N}$ by

$$\boldsymbol{h}(\boldsymbol{u}) = \begin{pmatrix} \boldsymbol{h}_V(\boldsymbol{u}) \\ \boldsymbol{h}_{\mathcal{P}}(\boldsymbol{u}) \end{pmatrix} = \begin{pmatrix} [\underline{\mathcal{V}}_i - \operatorname{vol}(D_i)]_{i\in\{1,\dots,N\}} \\ \left[\operatorname{peri}(u_i) - \overline{\mathcal{P}}_i\right]_{i\in\{1,\dots,N\}} \end{pmatrix},$$

as well as the set $\boldsymbol{K} := \{\boldsymbol{h} \in \mathbb{R}^{2N} : h_i \leq 0 \ \forall i = 1, \dots, 2N\}$ and the objective $J(\boldsymbol{u}, \boldsymbol{\xi}) := \int_D \|\boldsymbol{q}_{\boldsymbol{\xi}}(\boldsymbol{x}) + \boldsymbol{g}_{\boldsymbol{\xi}}(\boldsymbol{x}) - \bar{\boldsymbol{q}}(\boldsymbol{x})\|_2^2 \,\mathrm{d}\boldsymbol{x}$. The parametrized augmented Lagrangian is defined by

$$\begin{aligned} L_A(\boldsymbol{u}, \boldsymbol{\lambda}, \boldsymbol{\xi}; \mu) = J(\boldsymbol{u}, \boldsymbol{\xi}) + \int_D \nabla \boldsymbol{q}_{\boldsymbol{\xi}} : \nabla \boldsymbol{\varphi}_{\boldsymbol{\xi}} - p_{\boldsymbol{\xi}} \operatorname{div} \boldsymbol{\varphi}_{\boldsymbol{\xi}} + \psi_{\boldsymbol{\xi}} \operatorname{div} \boldsymbol{q}_{\boldsymbol{\xi}} \,\mathrm{d}\boldsymbol{x} \\ + \frac{\mu}{2} \operatorname{dist}_{\boldsymbol{K}} \left(\boldsymbol{h}(\boldsymbol{u}) + \frac{\boldsymbol{\lambda}}{\mu} \right)^2 - \frac{\|\boldsymbol{\lambda}\|_2^2}{2\mu}. \end{aligned} \tag{31}$$

Differentiating the Lagrangian (31) with respect to $(\boldsymbol{q}, p)$ and setting it to zero gives the weak form of the adjoint equation: find $\boldsymbol{\varphi}_{\boldsymbol{\xi}} \in V(D)$ and $\psi_{\boldsymbol{\xi}} \in L^2(D)$ such that

$$\int_D 2\tilde{\boldsymbol{\varphi}}^\top \left(\boldsymbol{q}_{\boldsymbol{\xi}} + \boldsymbol{g}_{\boldsymbol{\xi}} - \bar{\boldsymbol{q}} \right) + \nabla \boldsymbol{\varphi}_{\boldsymbol{\xi}} : \nabla \tilde{\boldsymbol{\varphi}} + \psi_{\boldsymbol{\xi}} \operatorname{div} \tilde{\boldsymbol{\varphi}} \,\mathrm{d}\boldsymbol{x} = 0 \quad \forall \tilde{\boldsymbol{\varphi}} \in V(D), \tag{32a}$$

$$\int_D \operatorname{div} \boldsymbol{\varphi}_{\boldsymbol{\xi}} \, \tilde{\psi} \,\mathrm{d}\boldsymbol{x} = 0 \quad \forall \tilde{\psi} \in L^2(D). \tag{32b}$$

We define the space $\mathcal{W}(D) = \{\boldsymbol{W} \in H^1(D, \mathbb{R}^2)\colon \boldsymbol{W}|_\Gamma = 0\}$. We have the shape derivative

$$\begin{aligned} &\mathrm{d}_{\boldsymbol{u}} L_A(\boldsymbol{u}, \boldsymbol{\lambda}, \boldsymbol{\xi}; \mu)\,[\boldsymbol{W}] \\ &= \int_D - \left(\nabla \boldsymbol{q}_{\boldsymbol{\xi}} \nabla \boldsymbol{W}\right) : \nabla \boldsymbol{\varphi}_{\boldsymbol{\xi}} - \left(\nabla \boldsymbol{\varphi}_{\boldsymbol{\xi}} \nabla \boldsymbol{W}\right) : \nabla \boldsymbol{q}_{\boldsymbol{\xi}} + \left(p_{\boldsymbol{\xi}} \nabla \boldsymbol{\varphi}_{\boldsymbol{\xi}}{}^\top - \psi_{\boldsymbol{\xi}} \nabla \boldsymbol{q}_{\boldsymbol{\xi}}{}^\top\right) : \nabla \boldsymbol{W} \\ &\qquad + \operatorname{div}(\boldsymbol{W}) \left(\|\boldsymbol{q}_{\boldsymbol{\xi}} + \boldsymbol{g}_{\boldsymbol{\xi}} - \bar{\boldsymbol{q}}\|_2^2 + \nabla \boldsymbol{q}_{\boldsymbol{\xi}} : \nabla \boldsymbol{\varphi}_{\boldsymbol{\xi}} - p_{\boldsymbol{\xi}} \operatorname{div} \boldsymbol{\varphi}_{\boldsymbol{\xi}} + \psi_{\boldsymbol{\xi}} \operatorname{div} \boldsymbol{q}_{\boldsymbol{\xi}} \right) \mathrm{d}\boldsymbol{x} \\ &\quad + \mu \left(\left(\boldsymbol{h}(\boldsymbol{u}) + \frac{\boldsymbol{\lambda}}{\mu} \right) - \pi_{\boldsymbol{K}} \left(\boldsymbol{h}(\boldsymbol{u}) + \frac{\boldsymbol{\lambda}}{\mu} \right) \right)^\top \\ &\quad \begin{pmatrix} \left[\int_{D_i} \operatorname{div}(\boldsymbol{W}) \,\mathrm{d}\boldsymbol{x} \right]_{i\in\{1,\dots,N\}} \\ \left[\int_{u_i} \operatorname{div}(\boldsymbol{W}) - \boldsymbol{n}^\top \nabla \boldsymbol{W} \boldsymbol{n} \,\mathrm{d}\boldsymbol{s} \right]_{i\in\{1,\dots,N\}} \end{pmatrix}, \end{aligned} \tag{33}$$

where $\left(\boldsymbol{q}_{\boldsymbol{\xi}}, p_{\boldsymbol{\xi}}\right)$ and $\left(\boldsymbol{\varphi}_{\boldsymbol{\xi}}, \psi_{\boldsymbol{\xi}}\right)$ solve the state equation (28) and adjoint equation (32), respectively. The shape derivative is needed to represent the gradient with respect to the metric under consideration (cf., e.g., [11]). As described in [11], we can use the multi-shape derivative in an "all-at-once"-approach to compute the multi-shape gradient with respect to the Steklov–Poincaré metric and the mesh deformation $\boldsymbol{V} = \boldsymbol{V}_{\boldsymbol{\xi}}$ all at once by solving

$$a(\boldsymbol{V}, \boldsymbol{W}) = \mathrm{d}_{\boldsymbol{u}} L_A(\boldsymbol{u}, \boldsymbol{\lambda}, \boldsymbol{\xi}; \mu)[\boldsymbol{W}] \quad \forall \boldsymbol{W} \in \mathcal{W}(D) \cap \mathcal{C}^\infty(D, \mathbb{R}^2), \tag{34}$$

where a is a coercive and symmetric bilinear form. The mesh deformation $\boldsymbol{V}$ calculated from (34) can be viewed as an extension of the multi-shape gradient $\boldsymbol{v}$ with respect to the Steklov–Poincaré metric to the hold-all domain D (for details we refer the reader to [11]).

The bilinear form that describes linear elasticity is a common choice for a due to the advantageous effect on the computational mesh (cf. [28, 30]), and is selected for the following numerical studies. The Lamé parameters are chosen as $\hat{\lambda} = 0$ and $\hat{\mu}$ smoothly decreasing from 33 on $\boldsymbol{u}$ to 10 on Γ, as obtained by the solution of Poisson's equation on D.

To update the shapes according to Algorithm 1, we need to compute the multi-exponential map. This computation is prohibitively expensive in most applications because a calculus of variations problem must be solved or the Christoffel symbols need be known. Therefore, we approximate it using a multi-retraction

$$\mathcal{R}^N_{\boldsymbol{z}^{k,j}} \colon T_{\boldsymbol{z}^{k,j}}\mathcal{U}^N \to \mathcal{U}^N,\ \boldsymbol{v} = (v_1, \dots, v_N) \mapsto (\mathcal{R}_{z_1^{k,j}} v_1, \dots, \mathcal{R}_{z_N^{k,j}} v_N)$$

to update the shape vector $\boldsymbol{z}^{k,j} = (z_1^{k,j}, \dots, z_N^{k,j})$ in each pair (j,k). For each shape $z_i^{k,j}$ we use the retraction in [11, 10, 26]: $\mathcal{R}_{z_i^{k,j}} \colon T_{z_i^{k,j}}\mathcal{U}^i \to \mathcal{U}^i,\ v_i \mapsto z_i^{k,j} + v_i$ for all $i = 1, \dots, N$.

Numerical results. All numerical simulations were performed on the HPC cluster HSUper* using the FEniCS toolbox, version 2019.1.0 [2] and Python 3.9.12. The hold-all domain is chosen as $D = (0,1)^2$. We choose $N = 3$ shapes inside the hold-all domain, which can be seen on the left-hand side of Figure 1. The computational mesh is generated with Gmsh 4.8.4 [13], which yields 265 line elements for the outer boundary and the interfaces, and 3803 triangular elements as the discretization of D. Additionally, a new mesh was automatically generated if the mesh quality† fell below a threshold of 40%. The target velocity is shown in Figure 1 on the right, together with the shapes to obtain the target velocity. We used `numpy.random` from `numpy` 1.22.4 for the generation of all random values. The different seeds 964113, 454612, 421507 and 107785 were chosen. Parallelization of multiple realizations was performed via MPI using `mpi4py` version 3.1.2. Standard Taylor–Hood elements are used.

The values of the geometrical constraints were chosen in accordance with the shapes of the target velocity. The volumes of D_1, D_2 and D_3 were constrained to be at or above 0.035295, 0.025397 and 0.036967, and the perimeters of u_1, u_2 and u_3 to be at or below 0.72630, 0.56521 and 0.69796, respectively. The augmented Lagrangian parameters in Algorithm 1 were initialized to $\boldsymbol{\lambda}^1 = \boldsymbol{0}$, $\mu^1 = 10$, $\gamma = 10$, and $\tau = 0.9$. The ball for the projection of Lagrange multipliers was chosen to be $B = [-100, 100]^{2N}$.

We chose homogenous Dirichlet boundary conditions for the velocity on the top and bottom boundary and on $\boldsymbol{u}$ (see Figure 1, left). The inflow profile on the left boundary is modeled as an inhomogenous Dirichlet boundary with $\boldsymbol{g}_{\boldsymbol{\xi}}(\boldsymbol{x}) = (\kappa(\boldsymbol{x}, \boldsymbol{\xi}), 0)^\top$. The horizontal component is given by the truncated Karhunen-Loève expansion

$$\kappa(\boldsymbol{x}, \boldsymbol{\xi}) = -4x_2(x_2 - 1) + \sum_{\ell=1}^{100} \ell^{-\eta - 1/2} \sin(2\pi\ell(x_2 - 1/2))\xi_\ell,$$

where $\eta = 3.5$ and $\xi_\ell \sim U\left[-\frac{1}{2}, \frac{1}{2}\right]$ ($U[a,b]$ being the uniform distribution on the interval $[a,b]$). On the right boundary, a homogenous Neumann boundary condition is imposed. The step size is

*Further information about the technical specifications can be found at `https://www.hsu-hh.de/hpc/en/hsuper/`.

†The mesh quality is measured with the FEniCS function `MeshQuality.radius_ratio_min_max`.

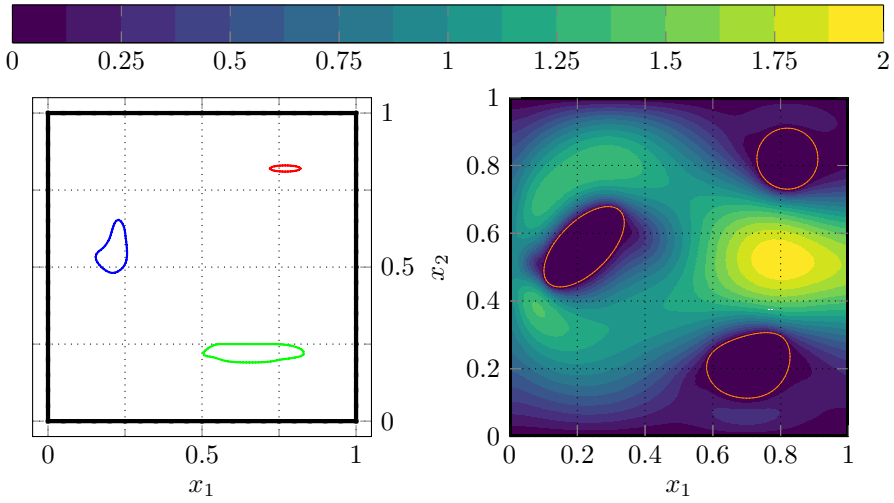
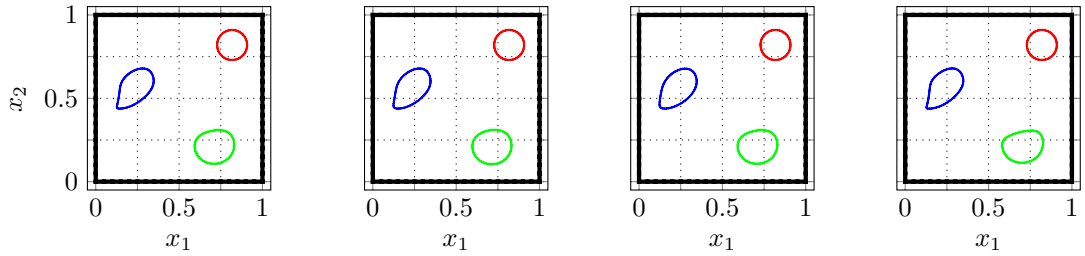


Figure 1: Shapes u_1 (blue), u_2 (red) and u_3 (green) at the start of the stochastic optimization (left) and the magnitude of the target fluid velocity $|\boldsymbol{q}|$ together with the shapes (in orange) used to obtain the target velocity (right).



(a) Seed 964113. (b) Seed 454612. (c) Seed 421507. (d) Seed 107785.

Figure 2: Shapes u_1 (blue), u_2 (red) and u_3 (green) at the end of the stochastic optimization with different seeds.

chosen as $t_k = \frac{20}{\mu_k}$, the scaling of which is obtained by tuning (to avoid deterioration of the mesh, especially in the first steps of the inner loop procedure). The maximum number of inner loop iterations is initialized to $N_1 = 5$ and is updated using $T_k = \frac{1}{\sqrt{k}}$ to $N_k = \lceil \frac{\gamma N_{k-1}}{T_k} \rceil$ if $\mu_k > \mu_{k-1}$, otherwise $N_k = \lceil \frac{N_{k-1}}{T_k} \rceil$. The batch size is increased according to $m_k = \lceil \frac{m_{k-1}}{T_k} \rceil$ with $m_1 = 25$. Each inner loop k requires $m_k \cdot R_k$ solutions of the state equation, the adjoint equation, the Poisson equation for the Lamé parameter, and the deformation equation, which becomes computationally expensive for high k and potentially for higher μ_k.

The obtained shapes for the four different seeds are shown in Figure 2. The red shape u_2 looks basically identical for the four different seeds, however u_1 (blue) shows a difference in the bottom-left and u_3 has a different top part. We investigate the optimization with the random seed 421507 further. The remesher is activated at the stochastic gradient step 6, 13, 18, 25, 40 and 86. In Figure 3, the numerical results for objective functional estimate $\hat{j} = \frac{1}{m_k} \sum_{i=1}^{m_k} J(\boldsymbol{z}^{k,j}, \boldsymbol{\xi}^{k,j,i})$ and the estimate of the H^1 norm of the mesh deformation $\widehat{\boldsymbol{V}} = \frac{1}{m_k} \sum_{i=1}^{m_k} \boldsymbol{V}_{\boldsymbol{\xi}^{k,j,i}}$ over cumulative inner iterations is provided. Here, even for a comparatively low number of samples per inner iteration, we see a strong decrease in objective functional values initially. The points where the inner loop is stopped due to reaching R_k are denoted by the red vertical dashed lines in the right-hand side plot. At the later stages of the optimization the batch size is increased up to $m_9 = 15525$ for $k = 9$, which requires over 60000

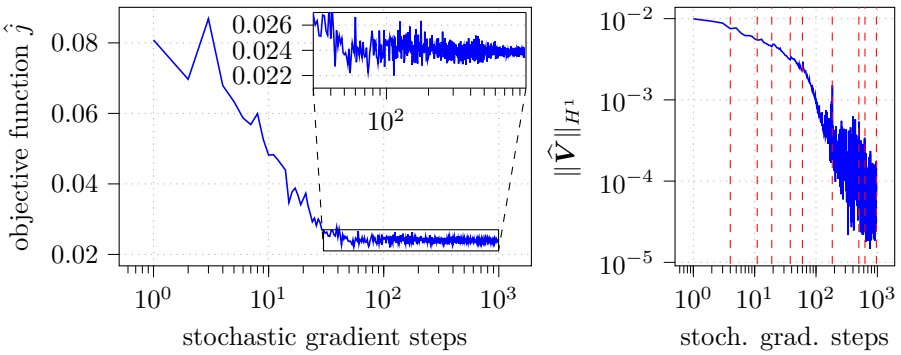


Figure 3: Stochastic optimization with seed 421507 (Figure 2c): objective functional (left) and H^1 norm of the mesh deformation (right) as a function of cumulative stochastic gradient steps. The changes of augmented Lagrange parameters are indicated with a red, dashed, vertical line.

PDE solves per stochastic gradient step. This yields an increasingly accurate approximation of the mesh deformation and the objective functional value as evidenced by the decreasing variance.

We provide the numerical results at the end of each inner loop for the four different seeds in table 1. Here, one can clearly see a dependence of the algorithm's performance based on the seed chosen, leading to different sequences of RSG iterations and penalty parameters. A further analysis of this dependence will be the subject of a future study.

References

[1] P. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

[2] M. S. Alnaes, J. Blechta, J. Hake, A. Johansson, B. Kehlet, A. Logg, C. Richardson, J. Ring, M. E. Rognes, and G. N. Wells. The FEniCS project version 1.5. *Archive of Numerical Software*, 3, 2015.

[3] H. Bauschke and P. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer International Publishing, 2017.

[4] R. Bergmann and R. Herzog. Intrinsic formulation of KKT conditions and constraint qualifications on smooth manifolds. *SIAM Journal on Optimization*, 29(4):2423–2444, 2019.

[5] D. Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic Press, 1982.

[6] E. Birgin and J. Martínez. *Practical augmented Lagrangian methods for constrained optimization*. SIAM, 2014.

[7] S. Blauth, C. Leithäuser, and R. Pinnau. Shape sensitivity analysis for a microchannel cooling system. *Journal of Mathematical Analysis and Applications*, 492(2):124476, 2020.

[8] M. Cheney, D. Isaacson, and J. Newell. Electrical impedance tomography. *SIAM Review*, 41(1):85–101, 1999.

[9] A. Dener, M. Miller, R. Churchill, T. Munson, and C.-S. Chang. Training neural networks under physical constraints using a stochastic augmented Lagrangian approach. arXiv preprint, 2020.

Table 1: Numerical values at the end of each inner loop for four different seeds. One additional iteration compared to the other seeds is performed for seed 421507 before computational effort prohibits further iterations.

k	R_k	μ_k	$\|\widehat{\boldsymbol{V}}^k\|_{H^1}$
1	2	10	$9.42 \cdot 10^{-3}$
2	4	10	$6.55 \cdot 10^{-3}$
3	137	100	$6.07 \cdot 10^{-3}$
4	241	100	$2.18 \cdot 10^{-3}$
5	305	100	$5.24 \cdot 10^{-4}$
6	186	100	$2.77 \cdot 10^{-4}$
7	539	100	$2.29 \cdot 10^{-4}$
8	7580	100	$1.30 \cdot 10^{-4}$

k	R_k	μ_k	$\|\widehat{\boldsymbol{V}}^k\|_{H^1}$
1	3	10	$9.29 \cdot 10^{-3}$
2	3	10	$6.63 \cdot 10^{-3}$
3	69	100	$8.43 \cdot 10^{-3}$
4	245	100	$2.21 \cdot 10^{-3}$
5	373	100	$9.28 \cdot 10^{-4}$
6	1271	100	$4.28 \cdot 10^{-4}$
7	3856	100	$2.25 \cdot 10^{-4}$
8	233	100	$2.65 \cdot 10^{-4}$

Seed 964113 (left) and seed 454612 (right)

k	R_k	μ_k	$\|\widehat{\boldsymbol{V}}^k\|_{H^1}$
1	4	10	$7.53 \cdot 10^{-3}$
2	7	10	$5.46 \cdot 10^{-3}$
3	8	10	$4.52 \cdot 10^{-3}$
4	19	10	$3.11 \cdot 10^{-3}$
5	22	10	$2.38 \cdot 10^{-3}$
6	124	10	$9.51 \cdot 10^{-4}$
7	315	10	$1.07 \cdot 10^{-4}$
8	129	10	$1.72 \cdot 10^{-4}$
9	336	10	$9.53 \cdot 10^{-5}$

k	R_k	μ_k	$\|\widehat{\boldsymbol{V}}^k\|_{H^1}$
1	1	10	$1.09 \cdot 10^{-2}$
2	2	10	$8.48 \cdot 10^{-3}$
3	85	100	$8.33 \cdot 10^{-3}$
4	139	100	$5.00 \cdot 10^{-3}$
5	174	100	$1.27 \cdot 10^{-3}$
6	2002	1000	$8.01 \cdot 10^{-4}$
7	21378	1000	$1.08 \cdot 10^{-4}$
8	45302*	1000	$1.00 \cdot 10^{-4}$

Seed 421507 (left) and seed 107785 (right)

[10] C. Geiersbach, E. Loayza-Romero, and K. Welker. Stochastic approximation for optimization in shape spaces. *SIAM Journal on Optimization*, 31(1):348–376, 2021.

[11] C. Geiersbach, E. Loayza-Romero, and K. Welker. PDE-constrained shape optimization: towards product shape spaces and stochastic models. In *Handbook of Mathematical Models and Algorithms in Computer Vision and Imaging*. Springer, 2022. Accepted for publication.

[12] C. Geiersbach and T. Scarinci. A stochastic gradient method for a class of nonlinear PDE-constrained optimal control problems under uncertainty. *arXiv preprint*, 2021.

[13] C. Geuzaine and J.-F. Remacle. Gmsh: A 3-D finite element mesh generator with built-in pre- and post-processing facilities. *International Journal for Numerical Methods in Engineering*, 79(11):1309–1331, 2009.

[14] S. Ghadimi and G. Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

*This iteration had to be stopped after 16519 iterations (37898 cumulative inner iterations) due to the maximum run time of 24 h on the cluster.

[15] S. Ghadimi, G. Lan, and H. Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Applications of Management Science: in Productivity, Finance, and Operations*, 155(1-2):267–305, 2016.

[16] C. Kanzow and D. Steck. On error bounds and multiplier methods for variational problems in Banach spaces. *SIAM Journal on Control and Optimization*, 56(3):1716–1738, 2018.

[17] C. Kanzow and D. Steck. Improved local convergence results for augmented Lagrangian methods in C^2-cone reducible constrained optimization. *Applications of Management Science: in Productivity, Finance, and Operations*, 177(1):425–438, 2019.

[18] C. Kanzow, D. Steck, and D. Wachsmuth. An augmented Lagrangian method for optimization problems in Banach spaces. *SIAM Journal on Control and Optimization*, 56(1):272–291, 2018.

[19] V. Karl and D. Wachsmuth. An augmented Lagrange method for elliptic state constrained optimal control problems. *Computational Optimization and Applications*, 69(3):857–880, 2018.

[20] A. Kriegl and P. Michor. *The convient setting of global analysis*, volume 53 of *Mathematical Surveys and Monographs*. American Mathematical Society, 1997.

[21] O. Kwon, E. J. Woo, J. Yoon, and J. Seo. Magnetic resonance electrical impedance tomography (MREIT): Simulation study of J-substitution algorithm. *IEEE Transactions on Biomedical Engineering*, 49(2):160–167, 2002.

[22] A. Laurain and K. Sturm. Distributed shape derivative via averaged adjoint method and applications. *ESAIM: Mathematical Modelling and Numerical Analysis*, 50(4):1241–1267, 2016.

[23] P. Michor and D. Mumford. Riemannian geometries on spaces of plane curves. *J. Eur. Math. Soc. (JEMS)*, 8(1):1–48, 2006.

[24] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[25] V. Schulz, M. Siebenborn, and K. Welker. Efficient PDE constrained shape optimization based on Steklov-Poincaré-type metrics. *SIAM Journal on Optimization*, 26(4):2800–2819, 2016.

[26] V. Schulz and K. Welker. On optimization transfer operators in shape spaces. In *Shape Optimization, Homogenization and Optimal Control*, pages 259–275. Springer, 2018.

[27] M. Siebenborn and A. Vogel. A shape optimization algorithm for cellular composites. *PINT Computing and Visualization in Science*, 2021.

[28] M. Siebenborn and K. Welker. Algorithmic aspects of multigrid methods for optimization in shape spaces. *SIAM Journal on Scientific Computing*, 39(6):B1156–B1177, 2017.

[29] D. Steck. *Lagrange Multiplier Methods for Constrained Optimization and Variational Problems in Banach Spaces*. PhD thesis, Universität Würzburg, 2018.

[30] K. Welker. *Efficient PDE constrained shape optimization in shape spaces*. PhD thesis, Universität Trier, 2016.

[31] W. Yang, L.-H. Zhang, and R. Song. Optimality conditions for the nonlinear programming problems on Riemannian manifolds. *Pacific Journal of Optimization*, 10(2):415–434, 2014.