Noname manuscript No.
(will be inserted by the editor)

Riemannian Stochastic Fixed Point Optimization Algorithm

Hideaki Iiduka · Hiroyuki Sakai

Received: date / Accepted: date

Abstract This paper considers a stochastic optimization problem over the fixed point sets of quasinonexpansive mappings on Riemannian manifolds. The problem enables us to consider Riemannian hierarchical optimization problems over complicated sets, such as the intersection of many closed convex sets, the set of all minimizers of a nonsmooth convex function, and the intersection of sublevel sets of nonsmooth convex functions. We focus on adaptive learning rate optimization algorithms, which adapt step-sizes (referred to as learning rates in the machine learning field) to find optimal solutions quickly. We then propose a Riemannian stochastic fixed point optimization algorithm, which combines fixed point approximation methods on Riemannian manifolds with the adaptive learning rate optimization algorithms. We also give convergence analyses of the proposed algorithm for nonsmooth convex and smooth nonconvex optimization. The analysis results indicate that, with small constant step-sizes, the proposed algorithm approximates a solution to the problem. Consideration of the case in which step-size sequences are diminishing demonstrates that the proposed algorithm solves the problem with a guaranteed convergence rate. This paper also provides numerical comparisons that demonstrate the effectiveness of the proposed algorithms with formulas based on the adaptive learning rate optimization algorithms, such as Adam and AMSGrad.

Keywords adaptive learning rate optimization algorithm · fixed point · hierarchical optimization · quasinonexpansive mapping · Riemannian

This work was supported by JSPS KAKENHI Grant Number JP18K11184.

H. Iiduka
Department of Computer Science, Meiji University 1-1-1 Higashimita, Tama-ku, Kawasaki-shi, Kanagawa 214-8571, Japan
E-mail: iiduka@cs.meiji.ac.jp
H. Sakai
Department of Computer Science, Meiji University 1-1-1 Higashimita, Tama-ku, Kawasaki-shi, Kanagawa 214-8571, Japan
E-mail: sakai0815@cs.meiji.ac.jp

stochastic fixed point optimization algorithm · Riemannian stochastic optimization

Mathematics Subject Classification (2000) 65K05 · 90C15 · 90C25 · 90C26

1 Introduction

In light of developments in machine learning and image/signal processing (see, e.g., [5,15,26,29] and references therein), Riemannian optimization has attracted a great deal of attention. Useful iterative algorithms thus have been presented for Riemannian optimization. For example, nonlinear Riemannian conjugate gradient methods have been widely studied in [15,28,31,32] for unconstrained optimization. First-order methods [6,36] and proximal point algorithms [11,22] have been reported for unconstrained/constrained Riemannian optimization. Riemannian stochastic gradient methods were proposed in [7, 17,33] for Riemannian stochastic optimization.

For training deep neural networks, *adaptive learning rate optimization algorithms* based on using stochastic subgradients and exponential moving averages have a strong presence since they have fast convergence for stochastic optimization in Euclidean space. For example, AdaGrad [10] and RMSProp [34] take advantage of efficient learning rates (referred to as step-sizes in the field of optimization) derived from element-wise squared stochastic gradients. Adam [18] and AMSGrad [27] are also useful algorithms using exponential moving averages of stochastic gradients and of element-wise squared stochastic gradients.

Recently, Riemannian AMSGrad (RAMSGrad) was studied in [5], which is a modification of AMSGrad for Euclidean space to be applicable to a product of Riemannian manifolds. RAMSGrad uses the metric projection onto a constraint convex set so as to satisfy that the sequence generated by RAMSGrad belongs to the constraint set. Accordingly, RAMSGrad can be applied to only Riemannian convex optimization with *simple* constraints in the sense that the metric projection can be easily computed.

In contrast to [5], this paper tries to consider a Riemannian optimization problem with *complicated* constraints, such as the intersection of many convex sets [1,6,36], the set of minimizers of a convex function [11,22], and the intersection of sublevel sets of convex functions [36]. The problem is a *hierarchical constrained optimization problem* with three stages, as follows. The first stage is to find points in Riemannian manifolds (e.g., to find points in nonconvex constraints in Euclidean space). The second stage is to find fixed points of quasinonexpansive mappings on Riemannian manifolds. Complicated convex sets, such as those mentioned above, can be expressed as the fixed point set of a quasinonexpansive mapping on a Riemannian manifold (Proposition 2.2). The third stage is to optimize an objective function over the second stage. For example, the third stage includes the case of trying to find a stationary

point of a smooth nonconvex function over the set of minimizers of a convex function over a Riemannian manifold.

The reason why the above problem should be considered is to enable us to resolve unsolved optimization problems on Riemannian manifolds. For example, in the natural language processing for hierarchical representations of symbolic data, embeddings into a Poincaré ball perform better than embeddings into a Euclidean space [26]. This implies that a Riemannian optimization problem should be considered for natural language processing. As seen above, we expect to gain new insights from re-considering several problems with complicated constraints in the Hilbert/Euclidean space setting as Riemannian optimization. In addition, a classifier ensemble problem with sparsity leaning can be expressed as a Euclidean convex optimization problem over the sublevel set of a convex function [16]. Since the results in this paper enable us to consider a Riemannian optimization problem over the sublevel set of a convex function, there is a possibility that the results will lead to new Riemannian learning methods which can outperform the existing methods in [16] by a wide margin.

We first define *quasinonexpansive mappings* of which *fixed point sets* are equal to complicated constraint sets. Thanks to the previously reported results in [11,21,22], we can define quasinonexpansive mappings for the cases where the constraint sets are those mentioned above: the intersection of many closed convex sets, the set of all minimizers of a nonsmooth convex function, and the intersection of sublevel sets of convex functions (Proposition 2.3). Accordingly, the Riemannian optimization problem with such complicated constraints can be expressed as a Riemannian optimization problem over the fixed point sets of quasinonexpansive mappings (Problem 1). Next, we combine the ideas of adaptive learning rate optimization algorithms (see the second and third paragraphs of this section) with the fixed point methods [21]. We then propose a *Riemannian stochastic fixed point optimization algorithm* (Algorithm 1) for solving the problem.

The intellectual contribution of this paper is that the proposed methodology enables one to deal with Riemannian optimization over the fixed point sets of quasinonexpansive mappings, especially in contrast to recent papers [5,29] that discussed Riemannian convex optimization over simple constraints. To clarify this contribution, let us consider the case where a constraint set is the intersection of many closed convex sets on a Riemannian manifold (Proposition 2.3(ii), Example 1). Even if the metric projection onto each closed convex set can be easily computed within a finite number of arithmetic operations, the metric projection onto the intersection of many closed convex sets would not be implemented in practice. This is because, for each iteration, we must solve a subproblem of minimizing the distance function over the intersection to find the nearest point to the intersection. Meanwhile, we can use a *computable* mapping which consists of the product of the metric projections (see Example 1 for the details), since the metric projection onto each closed convex set can be easily implemented. This computable mapping satisfies the quasinonexpansivity condition and that the fixed point set of this mapping coincides with the intersection of many closed convex sets. Therefore, the proposed algorithm (Al-

gorithm 1) using this computable quasinonexpansive mapping can be applied to Riemannian optimization over the intersection of many closed convex sets, in contrast to [5,29], which discussed Riemannian convex optimization over simple constraints. This paper also gives other examples of Problem 1, namely, Riemannian optimization over the set of minimizers of a convex function (Example 2) and Riemannian optimization over the intersection of sublevel sets of convex functions (Example 3).

The theoretical contribution of this paper is its analysis of the proposed algorithm (Algorithm 1) for solving the Riemannian optimization problem over the fixed point sets of quasinonexpansive mappings (Problem 1). The analysis indicates that the proposed algorithm with small constant step-sizes can approximate a solution to the main problem (Theorems 1 and 3). The analysis also shows that the proposed algorithm with diminishing step-sizes can solve the main problem with a guaranteed convergence rate (Theorems 2 and 4, and Corollary 1).

The practical contribution of this paper is a presentation of numerical results demonstrating that the proposed algorithm can be applied to Riemannian optimization over fixed point constraints. In this paper, we consider two cases for the constraint conditions. The first case is a consistent case such that the intersection of finite closed balls on the Poincaré disk is nonempty (Subsection 6.2). The second case is an inconsistent case such that the intersection is empty (Subsection 6.3). For the second case, we define a *generalized convex feasible set* as a subset of the absolute constrained set with the elements closest to the subsidiary constraint set. Numerical results show that the proposed algorithms with formulas based on Adam and AMSGrad perform well.

2 Mathematical Preliminaries

Let $\mathbb{N}$ be the set of all positive integers including zero, $\mathbb{R}^I$ be an I-dimensional Euclidean space, $\mathbb{R}^I_+ := \{(x_i)_{i=1}^I \in \mathbb{R}^I \colon x_i \geq 0 \ (i = 1, 2, \ldots, I)\}$, and $\mathbb{R}^I_{++} := \{(x_i)_{i=1}^I \in \mathbb{R}^I \colon x_i > 0 \ (i = 1, 2, \ldots, I)\}$. Let $\mathbb{E}[X]$ denote the expectation of random variable X. Unless stated otherwise, all relations between random variables are supported to hold almost surely.

2.1 Riemannian manifold and Hadamard manifold

Let M be a connected m-dimensional smooth manifold. Let T_xM be the tangent space of M at $x \in M$ and $TM = \bigcup_{x \in M} T_xM$ be the tangent bundle of M. A Riemannian metric at $x \in M$ is denoted by $\langle \cdot, \cdot \rangle_x \colon T_xM \times T_xM \to \mathbb{R}$ and its induced norm is defined for all $u \in T_xM$ by $\|u\|_x := \sqrt{\langle u, u \rangle_x}$. Manifold M endowed with Riemannian metric $\langle \cdot, \cdot \rangle := (\langle \cdot, \cdot \rangle_x)_{x \in M}$ is called a Riemannian manifold.

Given a piecewise smooth curve $\gamma \colon [a, b] \to M$ joining p to q (i.e., $\gamma(a) = p$ and $\gamma(b) = q$), the length $L(\gamma)$ of γ is defined by $L(\gamma) := \int_a^b \|\dot{\gamma}(t)\|_{\gamma(t)} \mathrm{d}t$,

where $\dot{\gamma}$ denotes the derivative of γ. The distance function $\mathrm{d}\colon M \times M \to \mathbb{R}_+$ is defined for all $p, q \in M$ by the minimal length over the set of all such curves joining p to q.

A complete, simply connected Riemannian manifold of nonpositive sectional curvature is called an Hadamard manifold. An m-dimensional Hadamard manifold M is diffeomorphic to the Euclidean space $\mathbb{R}^m$ [30, Chapter V, Corollary 3.5]. An exponential mapping at a point x in an Hadamard manifold M is denoted by $\exp_x\colon T_xM \to M$. The mapping $\exp_x$ is well-defined on T_xM, which is guaranteed by the Hopf-Rinow theorem [30, Chapter III, Theorem 1.1]. The mapping $\exp_x$ maps $u \in T_xM$ to $y := \exp_x(u) \in M$ such that there exists a geodesic $\gamma\colon [a, b] \to M$ satisfying $\gamma(a) = x$, $\gamma(b) = y$, and $\dot{\gamma}(a) = u$. The Hadamard-Cartan theorem [30, Chapter V, Theorem 4.1] guarantees that $\exp_x$ is diffeomorphic, that is, there exists an inverse mapping $\exp_x^{-1}\colon M \to T_xM$. For all $x, y \in M$, $\varphi_{x\to y}$ denotes an isometry from T_xM to T_yM.

Let M^i be an m^i-dimensional Hadamard space and M be the Cartesian product of the M^is, i.e., $M := M^1 \times M^2 \times \cdots \times M^I$. The tangent space of M at $x = (x^1, x^2, \ldots, x^I) \in M$ is defined by $T_xM := T_{x^1}M^1 \oplus T_{x^2}M^2 \oplus \cdots \oplus T_{x^I}M^I$, where $\oplus$ stands for the direct sum of vector spaces. For all $x = (x_i)_{i=1}^I \in M$, we define $\psi \in T_xM$ by $\psi = (\psi^i)_{i=1}^I = (\psi^1, \psi^2, \ldots, \psi^I)$, where $\psi^i \in T_{x^i}M^i$. An exponential mapping at a point $x^i \in M^i$ is denoted by $\exp_{x^i}^i$, and an isometry from $T_{x^i}M^i$ to $T_{y^i}M^i$ is denoted by $\varphi^i_{x^i\to y^i}$.

2.2 Convexity, monotonicity, and related mappings

Let M be an Hadamard manifold. A set $C \subset M$ is referred to as a convex set (see, e.g., [11, Subsection 3.1] and references therein) if, for any pair of points in C, the geodesic joining those two points is contained in C. Suppose that $C \subset M$ is nonempty, closed, and convex, and $x \in M$. Then there exists a unique point [11, Corollary 3.1], denoted by $P_C(x)$, such that

$$P_C(x) \in C \text{ and } \mathrm{d}(x, P_C(x)) = \inf_{y\in C} \mathrm{d}(x, y) =: \mathrm{d}(x, C).$$

We call P_C the *metric projection* onto C.

A function $f\colon M \to \mathbb{R}$ is said to be convex (see, e.g., [11, Subsection 3.2] and references therein) if, for any geodesic γ of M, $f \circ \gamma\colon \mathbb{R} \to \mathbb{R}$ is convex. Accordingly, any convex function on M is continuous. Suppose that $f\colon M \to \mathbb{R}$ is convex. Theorem 3.3 in [11] guarantees that, for all $x \in M$, there exists $u_x \in T_xM$ such that, for all $y \in M$,

$$f(y) \geq f(x) + \langle u_x, \exp_x^{-1}(y)\rangle_x.$$

The tangent vector u_x is called a *subgradient* of f at x. When f is smooth, the vector u_x is called the Riemannian gradient of f at x and is denoted by

$\mathrm{grad} f(x)$. The *subdifferential vector field* $\partial f\colon M \rightrightarrows TM$ of a convex function $f\colon M \to \mathbb{R}$ is defined by the set of all subgradients of f, i.e., for all $x \in M$,

$$\partial f(x) := \left\{u \in T_xM\colon f(y) \geq f(x) + \langle u, \exp_x^{-1}(y)\rangle_x \ (y \in M)\right\} \neq \emptyset.$$

The *subgradient projection* $P_{f,\lambda}$ relative to a convex function $f\colon M \to \mathbb{R}$ and $\lambda > 0$ is defined for all $x \in M$ by

$$P_{f,\lambda}(x) := \begin{cases} x & (x \in \mathrm{lev}_{\leq 0}(f) := \{x \in M\colon f(x) \leq 0\}), \\ \exp_x\left(-\lambda \dfrac{f(x)}{\|u_x\|_x} u_x\right) & (x \notin \mathrm{lev}_{\leq 0}(f)), \end{cases}$$

where u_x is any tangent vector in $\partial f(x)$. The results in [2, Lemma 3.1], [3, Proposition 2.3], and [35, Subchapter 4.3] provide the definition and properties of the subgradient projection under the Hilbert space setting.

Let $A\colon M \rightrightarrows TM$ be a set-valued vector field such that, for all $x \in D(A) := \{x \in M\colon A(x) \neq \emptyset\}$, $A(x) \subset T_xM$. A is said to be *monotone* (see, e.g., [22, Definition 2] and references therein) if, for all $x, y \in D(A)$, all $u \in A(x)$, and all $v \in A(y)$, $\langle u, \exp_x^{-1}(y)\rangle_x \leq \langle v, -\exp_y^{-1}(x)\rangle_y$. A is said to be *maximal* (see, e.g., [22, Definition 2] and references therein) if A is monotone and the following holds: for all $x \in M$ and all $u \in T_xM$, $\langle u, \exp_x^{-1}(y)\rangle_x \leq \langle v, -\exp_y^{-1}(x)\rangle_y$ $(y \in D(A), v \in A(y))$ implies that $u \in A(x)$. The subdifferential vector field ∂f of a convex function $f\colon M \to \mathbb{R}$ with $D(f) := \{x \in M\colon f(x) < +\infty\} = M$ is maximal monotone [20, Theorem 5.1]. We call the set of zeros of a set-valued vector field $A\colon M \rightrightarrows TM$ the zero point set, which is defined by $\mathrm{zer}(A) := \{x \in D(A)\colon 0 \in A(x)\}$.

Let $\lambda > 0$. The *resolvent* $J_\lambda\colon M \rightrightarrows M$ [22, Definition 6] of a set-valued vector field $A\colon M \rightrightarrows TM$ is defined for all $x \in M$ by

$$J_\lambda(x) := \{z \in M\colon x \in \exp_z(\lambda A(z))\}.$$

J_λ is single-valued when A is monotone [22, Theorem 4]. The *Moreau-Yosida regularization* $R_\lambda^f\colon M \rightrightarrows M$ [11, (20)], [22, (60)] of a convex function $f\colon M \to \mathbb{R}$ is defined for all $x \in M$ by

$$R_\lambda^f(x) := \operatorname*{argmin}_{y \in M}\left\{f(y) + \frac{1}{2\lambda}\mathrm{d}(x,y)^2\right\}.$$

R_λ^f is single-valued and $D(R_\lambda^f) = M$ [11, Lemma 4.2].

2.3 Nonexpansivity and fixed point set

Let C be a nonempty subset of a Riemannian manifold M with the distance function d and let $T\colon C \to M$ be a mapping. The *fixed point set* of T is defined by

$$\mathrm{Fix}(T) := \{x \in C\colon T(x) = x\}.$$

T is said to be *firmly nonexpansive* [22, Definition 1], [12, Subchapter 1.11] if, for all $x, y \in C$, the function $\Phi\colon [0,1] \to \mathbb{R}_+$ defined by

$$\Phi(t) := \mathrm{d}\left(\exp_x[t \exp_x^{-1}(T(x))], \exp_y[t \exp_y^{-1}(T(y))]\right) \text{ is decreasing.} \tag{1}$$

T is said to be *nonexpansive* if

$$\mathrm{d}(T(x), T(y)) \leq \mathrm{d}(x, y) \quad (x, y \in C). \tag{2}$$

T is said to be *quasinonexpansive* if

$$\mathrm{d}(T(x), y) \leq \mathrm{d}(x, y) \quad (x \in C, y \in \mathrm{Fix}(T)). \tag{3}$$

T is said to be *strictly quasinonexpansive* if

$$\mathrm{d}(T(x), y) < \mathrm{d}(x, y) \quad (x \in C \backslash \mathrm{Fix}(T), y \in \mathrm{Fix}(T)). \tag{4}$$

Finally, T is said to be *firmly quasinonexpansive* if

$$\mathrm{d}(T(x), y)^2 + \mathrm{d}(T(x), x)^2 \leq \mathrm{d}(x, y)^2 \quad (x \in C, y \in \mathrm{Fix}(T)). \tag{5}$$

The following proposition is true.

Proposition 2.1 *Suppose that C is a nonempty, closed convex set of an Hadamard manifold M and $T\colon C \to C$ is quasinonexpansive.*

(i) [8, Theorem 1.3] *If* $\mathrm{Fix}(T)$ *is nonempty, then* $\mathrm{Fix}(T)$ *is closed and convex;*
(ii) [19, Theorem 13] *If C is bounded and T is nonexpansive, then* $\mathrm{Fix}(T)$ *is nonempty.*

The relationships between the above mappings are given in the following proposition (the proof is given in Supplementary Material).

Proposition 2.2 *Let C be a nonempty subset of an m-dimensional Hadamard manifold M with the distance function* d *and let $T\colon C \to M$ be a mapping. Then,*

(i) (1) *implies* (2)*, and* (2) *implies* (3)*;*
(ii) (1) *implies* (5)*,* (5) *implies* (4)*, and* (4) *implies* (3)*.*

Let $T\colon C \to M$ be quasinonexpansive and $\alpha \in (0,1)$. Here, we define $S_\alpha\colon C \to M$ as follows: for all $x \in M$,

$$S_\alpha(x) := \exp_x[(1-\alpha) \exp_x^{-1}(T(x))].$$

Then, the condition $\mathrm{Fix}(T) = \mathrm{Fix}(S_\alpha)$ holds from the facts that $\exp_x$ is bijective and $\exp_x(0_x) = x$, where 0_x denotes the zero element of T_xM. Moreover, the discussion in [21, p.553] guarantees that, for all $x \in C \backslash \mathrm{Fix}(T)$ and all $y \in \mathrm{Fix}(T)$,

$$\mathrm{d}(S_\alpha(x), y)^2 \leq \mathrm{d}(x, y)^2 - \alpha(1-\alpha)\mathrm{d}(T(x), x)^2 < \mathrm{d}(x, y)^2, \tag{6}$$

that is, S_α is strictly quasinonexpansive.

The following proposition suggests some examples of quasinonexpansive mappings (the proof is given in Supplementary Material).

Proposition 2.3 *Let M be an m-dimensional Hadamard manifold, C_j $(j = 1, \ldots, J)$ be a nonempty, closed convex subset of M, and $\lambda > 0$. Suppose that $P_j := P_{C_j}$ is the metric projection onto C_j, $A\colon M \rightrightarrows TM$ is monotone, $g\colon M \to \mathbb{R}$ is convex, and $P_{g,\lambda}$ is the subgradient projection. Then, the following hold:*

(i) *The metric projection P_j is firmly nonexpansive with $\mathrm{Fix}(P_j) = C_j$;*
(ii) *Under $\bigcap_{j=1}^{J} C_j \neq \emptyset$, the mapping $T := P_1 P_2 \cdots P_J$ is nonexpansive with $\mathrm{Fix}(T) = \bigcap_{i=1}^{J} C_i$;*
(iii) *The resolvent J_λ of A is firmly nonexpansive with $\mathrm{Fix}(J_\lambda) = \mathrm{zer}(A)$;*
(iv) *The Moreau-Yosida regularization R_λ^g of g is firmly nonexpansive with $\mathrm{Fix}(R_\lambda^g) = \mathrm{argmin}_{x \in M}\, g(x)$;*
(v) *The subgradient projection $P_{g,\lambda}$ satisfies that $\mathrm{Fix}(P_{g,\lambda}) = \mathrm{lev}_{\leq 0}(g)$.*

Moreover, suppose that M has its sectional curvature lower-bounded by $\kappa \leq 0$, that C has a diameter bounded by D, and that $h_j\colon C \to \mathbb{R}$ $(j = 1, 2, \ldots, J)$ is convex with $D(h_j) = C$. Then, the following also hold:

(vi) *The subgradient projection $P_{h_j,\lambda}$ with $\lambda \in (0, 2/\zeta)$ is strictly quasinonexpansive with $\mathrm{Fix}(P_{h_j,\lambda}) = \mathrm{lev}_{\leq 0}(h_j)$, where ζ is a positive number depending on κ and D;*
(vii) *Under $\bigcap_{j=1}^{J} \mathrm{lev}_{\leq 0}(h_j) \neq \emptyset$, the mapping $T := P_{h_1,\lambda} P_{h_2,\lambda} \cdots P_{h_J,\lambda}$ with $\lambda \in (0, 2/\zeta)$ is strictly quasinonexpansive with $\mathrm{Fix}(T) = \bigcap_{j=1}^{J} \mathrm{lev}_{\leq 0}(h_j)$.*

3 Stochastic Optimization over Fixed Point Set on Riemannian Manifold

This paper considers the following problem.

Problem 1 Let M^i $(i \in \mathcal{I} := \{1, 2, \ldots, I\})$ be an m^i-dimensional Hadamard manifold with sectional curvature lower-bounded by $\kappa^i \leq 0$ and distance function d^i and M be the Cartesian product of the M^is, i.e., $M := M^1 \times M^2 \times \cdots \times M^I$. Assume that

(A1) $T^i\colon M^i \to M^i$ $(i \in \mathcal{I})$ is quasinonexpansive with $\mathrm{Fix}(T^i) \neq \emptyset$ $(i \in \mathcal{I})$, and $X := \mathrm{Fix}(T^1) \times \mathrm{Fix}(T^2) \times \cdots \times \mathrm{Fix}(T^I)$;
(A2) A function $f\colon M \to \mathbb{R}$ is defined for all $x \in M$ by $f(x) := \mathbb{E}[F(x, \xi)]$, where $F(\cdot, \xi)\colon M \to \mathbb{R}$ and ξ is a random vector whose probability distribution P is supported on a set $\Xi \subset \mathbb{R}^M$.

Then, we would like to find a point $x_\star$ in $X_\star$ defined by

$$X_\star := \left\{ x_\star \in X\colon \left\langle \exp_{x_\star}^{-1}(x), \mathsf{g}(x_\star) \right\rangle_{x_\star} \geq 0 \ (x \in X) \right\},$$

where $\mathsf{g}(x) = (\mathsf{g}^i(x))_{i \in \mathcal{I}}$ denotes the (sub)gradient of f.

The relationship between Problem 1 and the problem of minimizing f over X is expressed by the following proposition.

Proposition 3.1 *Suppose that Assumptions (A1) and (A2) hold.*

(i) *If $f\colon M \to \mathbb{R}$ is smooth, then*

$$X_\star \supset \operatorname*{argmin}_{x\in X} f(x) := \left\{ x_\star \in X\colon f(x_\star) = f_\star := \inf_{x\in X} f(x) \right\};$$

(ii) *If $f\colon M \to \mathbb{R}$ is convex, then*

$$X_\star = \operatorname*{argmin}_{x\in X} f(x).$$

Proposition 2.1(i) and Proposition 3.1 in [23] imply Proposition 3.1(i), which in turn implies that Problem 1 when f is smooth and nonconvex is a stationary point problem associated with the nonconvex optimization problem to minimize f over X. Meanwhile, when f is nonsmooth and convex, from the definition of the subdifferential vector field ∂f, we can prove Proposition 3.1(ii), i.e., that Problem 1 coincides with the nonconvex optimization problem to minimize f over X.

Proposition 2.3(i) and (ii) suggest the following example of Problem 1.

Example 1 (Optimization over the intersection of convex sets) Let C_j^i $(i \in \mathcal{I}, j \in \mathcal{J}^i := \{1, 2, \dots, J^i\})$ be a nonempty, closed convex subset of M^i with $\bigcap_{j\in\mathcal{J}^i} C_j^i \neq \emptyset$ and P_j^i $(j \in \mathcal{J}^i)$ be the metric projection onto C_j^i. Then, Problem 1 with a mapping $T^i := P_1^i P_2^i \cdots P_{J^i}^i$ $(i \in \mathcal{I})$ is to find a point $x_\star$ in $X_\star$ with

$$X = \bigcap_{j\in\mathcal{J}^1} C_j^1 \times \bigcap_{j\in\mathcal{J}^2} C_j^2 \times \cdots \times \bigcap_{j\in\mathcal{J}^I} C_j^I.$$

Let us compare the convex optimization problem considered in [5, Section 4] with Example 1. Example 1 when $J^i = 1$ $(i \in \mathcal{I})$ and f is a convex function coincides with the problem in [5, Section 4] that is to

$$\text{minimize } f(x) \text{ subject to } x \in C^1 \times C^2 \times \cdots \times C^I, \tag{7}$$

where $C^i := C_1^i$ $(i \in \mathcal{I})$ is simple in the sense that $P^i := P_1^i$ can be easily computed. Meanwhile, Example 1 has three stages as follows: The first stage is to find points of M. The second stage is to find points of complicated sets $\bigcap_{j\in\mathcal{J}^i} C_j^i$ $(i \in \mathcal{I})$, which are each the intersection of many convex sets. The problem in the second stage is called a *convex feasibility problem* [1], [4, p.99], [6,36]. The third stage is to minimize a function over the second stage. Hence, Problem 1 includes optimization problems with complicated constraints, as seen in Example 1.

From Proposition 2.3(iii) and (iv), we also have the following.

Example 2 (Optimization over the zero point sets) Let $A^i\colon M^i \rightrightarrows TM^i$ $(i \in \mathcal{I})$ be a monotone set-valued vector field with $\operatorname{zer}(A^i) \neq \emptyset$ and $J_{\lambda^i}^i\colon M^i \to M^i$

$(i \in \mathcal{I})$ be the resolvent of A^i with $\lambda^i > 0$. Then, Problem 1 with a mapping $T^i := J^i_{\lambda^i}$ $(i \in \mathcal{I})$ is to find a point $x_\star$ in $X_\star$ with

$$X = \mathrm{zer}\left(A^1\right) \times \mathrm{zer}\left(A^2\right) \times \cdots \times \mathrm{zer}\left(A^I\right).$$

In the case where $A^i := \partial g^i$ $(i \in \mathcal{I})$, where $g^i \colon M \to \mathbb{R}$ is convex, the problem is to find a point in $X_\star$ with

$$X = \operatorname*{argmin}_{x^1 \in M^1} g^1(x^1) \times \operatorname*{argmin}_{x^2 \in M^2} g^2(x^2) \times \cdots \times \operatorname*{argmin}_{x^I \in M^I} g^I(x^I).$$

References [11] and [22] presented proximal point algorithms which use the resolvents of a monotone vector field A for finding a zero of A,

$$x^* \in \mathrm{zer}(A).$$

Thanks to the results in [11] and [22] for the resolvents and Moreau-Yosida regularizations, Problem 1 includes Example 2 that is to minimize not only convex functions g^i (using the resolvents of ∂g^i) but also a function f over the sets of minimizers of the g^is.

Proposition 2.3 (v)–(vii) suggest the following example:

Example 3 (Optimization over the sublevel sets of convex functions) Let C^i be a nonempty, closed convex subset of M^i which has a diameter bounded by D^i and $g^i_j \colon C^i \to \mathbb{R}$ $(i \in \mathcal{I}, j \in \mathcal{J}^i)$ be a convex function with $D(g^i_j) = C^i$ $(j \in \mathcal{J}^i)$ and $\bigcap_{j \in \mathcal{J}^i} \mathrm{lev}_{\leq 0}(g^i_j) \neq \emptyset$. Then, Problem 1 with a mapping $T^i := P_{g^i_1, \lambda^i} P_{g^i_2, \lambda^i} \cdots P_{g^i_{J^i}, \lambda^i}$ $(i \in \mathcal{I})$ is to find a point $x_\star$ in $X_\star$ with

$$X = \bigcap_{j \in \mathcal{J}^1} \mathrm{lev}_{\leq 0}\left(g^1_j\right) \times \bigcap_{j \in \mathcal{J}^2} \mathrm{lev}_{\leq 0}\left(g^2_j\right) \times \cdots \times \bigcap_{j \in \mathcal{J}^I} \mathrm{lev}_{\leq 0}\left(g^I_j\right),$$

where $\lambda^i \in (0, 2/\zeta^i)$ and $\zeta^i := \sqrt{|\kappa^i|} D^i / \tanh(\sqrt{|\kappa^i|} D^i)$ $(i \in \mathcal{I})$.

References [6] and [36] proposed Riemannian subgradient algorithms for finding a point x^* in the intersection of sublevel sets of convex functions g_j $(j = 1, 2, \ldots, J)$ defined on a Riemannian manifold, i.e.,

$$x^* \in \bigcap_{j=1}^{J} \mathrm{lev}_{\leq 0}\left(g_j\right).$$

Algorithm 3.1 in [36] converges linearly to x^* without assuming that the domain of g_j has a bounded diameter. Meanwhile, under the assumption that the domain of g_j has a bounded diameter, Example 3 enables us to consider the three-stage Riemannian optimization problem such that the first stage is to find points in M, the second stage is to find points in sublevel sets of convex functions, and the third stage is to minimize a function over the second stage.

This section ends with a statement of the conditions for being able to solve Problem 1 (see, e.g., [25, (A1), (A2), (2.5)]).

(C1) There is an independent and identically distributed sample $\xi_0, \xi_1, \ldots$ of realizations of the random vector ξ;

(C2) There is an oracle which, for a given input point $(x, \xi) \in M \times \Xi$, returns a stochastic (sub)gradient $\mathsf{G}(x, \xi) = (\mathsf{G}^i(x, \xi))_{i \in \mathcal{I}}$ such that

$$\mathsf{g}(x) = (\mathsf{g}^i(x))_{i \in \mathcal{I}} := \mathbb{E}[\mathsf{G}(x, \xi)] \begin{cases} \in \partial f(x) \ (f \text{ is nonsmooth and convex}), \\ = \mathrm{grad} f(x) \ (f \text{ is smooth and nonconvex}); \end{cases}$$

(C3) For all $i \in \mathcal{I}$, there exists a positive number B^i such that, for all $x \in M$, $\mathbb{E}[\|\mathsf{G}^i(x, \xi)\|_{x^i}^2] \leq {B^i}^2$.

4 Riemannian Stochastic Fixed Point Optimization Algorithm

Let $i \in \mathcal{I}$. Given a quasinonexpansive mapping $T^i \colon M^i \to M^i$ in Problem 1 and $\alpha^i \in (0, 1)$, we define $S^i_{\alpha^i} \colon M^i \to M^i$ for all $x^i \in M^i$ by

$$S^i_{\alpha^i}(x^i) := \exp^i_{x^i} \left[\left(1 - \alpha^i\right) \left(\exp^i_{x^i}\right)^{-1} \left(T^i(x^i)\right) \right]. \tag{8}$$

The discussion in Subsection 2.3 (see (6)) ensures that $S^i_{\alpha^i}$ is strictly quasinonexpansive with $\mathrm{Fix}(T^i) = \mathrm{Fix}(S^i_{\alpha^i})$. Moreover, we define

$$Q^i_{\alpha^i} := P^i S^i_{\alpha^i}, \tag{9}$$

where P^i is the metric projection onto a nonempty, closed convex set C^i satisfying

$$C^i \supset \mathrm{Fix}\left(S^i_{\alpha^i}\right) = \mathrm{Fix}\left(T^i\right). \tag{10}$$

Algorithm 1 is the proposed algorithm for solving Problem 1. The tangent vectors m_n and $\hat{m}_n$ generated by steps 3 and 4 in Algorithm 1 are based on so-called momentum terms [13, Subchapter 8.3.2]. Step 8 in Algorithm 1 is expressed as

$$x^i_{n+1} := Q^i_{\alpha^i} \left[\exp^i_{x^i_n} \left(-\frac{\alpha_n}{(1 - \hat{\beta}^{n+1})\mathsf{h}^i_n} \hat{m}^i_n \right) \right],$$

which implies that Algorithm 1 adapts the step-size $\alpha_n / ((1 - \hat{\beta}^{n+1})\mathsf{h}^i_n)$ for each $i \in \mathcal{I}$ and each $n \in \mathbb{N}$. Hence, we can see that Algorithm 1 is based on so-called *adaptive learning rate optimization algorithms*, such as AdaGrad [10], Adam [18], and AMSGrad [27] defined on Euclidean space and RAMSGrad [5] defined on a Riemannian manifold. Examples of h^i_n are included in Examples 4 and 5.

The following conditions are assumed to analyze Algorithm 1.

Assumption 4.1 *The sequence* $(\mathsf{H}_n)_{n \in \mathbb{N}} := ((\mathsf{h}^i_n)_{i \in \mathcal{I}})_{n \in \mathbb{N}}$ *and a nonempty, closed convex set* $C^i \supset \mathrm{Fix}(T^i)$ *in Algorithm 1 satisfy the following conditions:*

Algorithm 1 Riemannian stochastic fixed point optimization algorithm

Require: $(\alpha_n)_{n\in\mathbb{N}} \subset (0,1)$, $(\alpha^i)_{i\in\mathcal{I}} \subset (0,1)$, $(\beta_n)_{n\in\mathbb{N}} \subset [0,1)$, $\hat{\beta} \in [0,1)$
$n \leftarrow 0$, $x_0 \in M$, $\tau_{-1} = m_{-1} \in T_{x_0}M$, $(\mathsf{h}_0^i)_{i\in\mathcal{I}} \subset \mathbb{R}^I_{++}$
loop
  $m_n := \beta_n \tau_{n-1} + (1-\beta_n)\mathsf{G}(x_n, \xi_n)$
  $\hat{m}_n := \left(1-\hat{\beta}^{n+1}\right)^{-1} m_n$
  $(\mathsf{h}_n^i)_{i\in\mathcal{I}} \subset \mathbb{R}^I_{++}$
  for $i = 1, 2, \ldots, I$ **do**
    $\mathsf{d}_n^i = -\dfrac{\hat{m}_n^i}{\mathsf{h}_n^i}$
    $x_{n+1}^i := Q_{\alpha^i}^i \left[\exp_{x_n^i}^i \left(\alpha_n \mathsf{d}_n^i\right)\right]$
    $\tau_n^i := \varphi_{x_n^i \to x_{n+1}^i}^i (m_n^i)$
    $n \leftarrow n+1$
  end for
end loop

(A3) *For all $i \in \mathcal{I}$, C^i has a diameter bounded by D^i;*
(A4) *For all $n \in \mathbb{N}$ and all $i \in \mathcal{I}$, almost surely $\mathsf{h}_{n+1}^i \geq \mathsf{h}_n^i$;*
(A5) *For all $i \in \mathcal{I}$, there exists a positive number $\hat{B}^i$ such that, for all $n \in \mathbb{N}$, $\mathbb{E}[\mathsf{h}_n^i] \leq \hat{B}^i$.*

Assumption (A3) will be needed to analyze Algorithm 1 since the previously reported results were analyzed under Assumption (A3) (see, e.g., [25, p.1574] and [27, p.2] for convex stochastic optimization on Euclidean space, and see, e.g., [5, Section 4], [24, Subsection 3.2], and [38, Subsection 3.2] for convex stochastic optimization on a Riemannian manifold). For example, let us consider the Poincaré model of a hyperbolic space defined by a manifold $\mathcal{D}^{m^i} := \{x^i \in \mathbb{R}^{m^i} : \|x^i\| < 1\}$ equipped with the Riemannian metric $\rho_{x^i} := (1/(1-\|x^i\|^2)^2)\rho^{\mathrm{E}}_{x^i}$, where $\|\cdot\|$ is the Euclidean norm, $x^i \in \mathcal{D}^{m^i}$, and $\rho^{\mathrm{E}}_{x^i}$ is the Euclidean metric tensor (the Poincaré embedding has been used for natural language processing [5, Section 5], [29, Section 4]). In [29, Section 4], $C^i := \{x^i \in \mathcal{D}^{m^i} : \|x^i\| \leq 1 - 10^{-5}\}$, which has a bounded diameter, was used to evaluate the performance of RAMSGrad for natural language processing. In the case of Example 1, Assumption (A3) is satisfied when at least one of C_j^i ($j \in \mathcal{J}^i$) has a bounded diameter. In Example 3, since $\mathrm{Fix}(T^i) = \bigcap_{j\in\mathcal{J}^i} \mathrm{lev}_{\leq 0}(g_j^i) \subset C^i$ and C^i has a bounded diameter, Assumption (A3) is satisfied. Assumption (A3) implies that, for all $i \in \mathcal{I}$,

$$D^i := \sup\left\{\mathrm{d}^i\left(x^i, y^i\right) : (x^i)_{i\in\mathcal{I}}, (y^i)_{i\in\mathcal{I}} \in X\right\} < +\infty, \tag{11}$$

where $\mathrm{d}^i : M^i \times M^i \to \mathbb{R}_+$ is the distance function of M^i.

Under Assumption (A3), we provide some examples of $(\mathsf{H}_n)_{n\in\mathbb{N}}$ satisfying Assumptions (A4) and (A5). The following examples are based on adaptive learning rate optimization algorithms, such as Adam [18] and AMSGrad [27], defined on Euclidean space.

Example 4 (H_n *based on Adam [18])* Let us define h_n^i and v_n^i for all $i \in \mathcal{I}$ and all $n \in \mathbb{N}$ by

$$\begin{aligned} v_n^i &:= \bar{\beta} v_{n-1}^i + (1-\bar{\beta}) \left\| \mathsf{G}^i(x_n, \xi_n) \right\|_{x_n^i}^2, \\ \bar{v}_n^i &:= \frac{v_n^i}{1-\bar{\beta}^{n+1}},\ \hat{v}_n^i := \max\left\{\hat{v}_{n-1}^i, \bar{v}_n^i\right\}, \\ \mathsf{h}_n^i &:= \sqrt{\hat{v}_n^i}, \end{aligned} \tag{12}$$

where $v_{-1}^i, \hat{v}_{-1}^i \in \mathbb{R}_+$ and $\bar{\beta} \in [0,1)$. From (12), H_n satisfies Assumption (A4). Moreover, (8), (9), and (10) mean that $(x_n^i)_{n\in\mathbb{N}} \subset C^i$, which, together with Assumption (A3), implies that $(\|\mathsf{G}^i(x_n, \xi_n)\|_{x_n^i})_{n\in\mathbb{N}}$ is almost surely bounded [14, Lemma 3.3]. For all $i \in \mathcal{I}$, we define

$$U^i := \max\left\{v_{-1}^i, \sup\left\{\left\|\mathsf{G}^i(x_n, \xi_n)\right\|_{x_n^i}^2 : n \in \mathbb{N}\right\}\right\} < +\infty.$$

Induction, together with the definitions of v_n^i, $\bar{v}_n^i$, and $\bar{\beta} \in [0,1)$, implies that, for all $n \in \mathbb{N}$, $v_n^i \leq U^i$ and $\bar{v}_n^i \leq U^i/(1-\bar{\beta})$. Accordingly, induction ensures that

$$\mathbb{E}\left[\mathsf{h}_n^i\right] \leq \sqrt{\max\left\{\hat{v}_{-1}^i, \frac{U^i}{1-\bar{\beta}}\right\}},$$

which implies that Assumption (A5) holds.

Example 5 (H_n *based on AMSGrad [5, 27])* Let us define h_n^i and v_n^i for all $i \in \mathcal{I}$ and all $n \in \mathbb{N}$ by

$$\begin{aligned} v_n^i &:= \bar{\beta} v_{n-1}^i + (1-\bar{\beta}) \left\| \mathsf{G}^i(x_n, \xi_n) \right\|_{x_n^i}^2, \\ \hat{v}_n^i &:= \max\left\{\hat{v}_{n-1}^i, v_n^i\right\}, \\ \mathsf{h}_n^i &:= \sqrt{\hat{v}_n^i}, \end{aligned} \tag{13}$$

where $v_{-1}^i, \hat{v}_{-1}^i \in \mathbb{R}_+$ and $\bar{\beta} \in [0,1)$. The same discussion as in Example 4 ensures that h_n^i defined by (13) satisfies Assumptions (A4) and (A5), i.e.,

$$\mathbb{E}\left[\mathsf{h}_n^i\right] \leq \sqrt{\max\left\{\hat{v}_{-1}^i, U^i\right\}}.$$

5 Convergence analyses of Algorithm 1

5.1 Nonsmooth convex optimization

This subsection considers Problem 1 when f is nonsmooth and convex. The following is a convergence analysis of Algorithm 1 with constant step-sizes (the proof of the following theorem is given in Supplementary Material).

Theorem 1 *Suppose that Assumptions (A1)–(A5) and Conditions (C1)–(C3) hold. Then, Algorithm 1 with $\alpha_n := \alpha$ and $\beta_n := \beta$ satisfies that, for all $i \in \mathcal{I}$,*

$$\limsup_{n\to+\infty} \mathbb{E}\left[\mathrm{d}^i\left(y_n^i, x_n^i\right)^2\right] \leq \frac{\tilde{B^i}^2}{(1-\hat{\beta})^2(\mathsf{h}_0^i)^2}\alpha^2, \tag{14}$$

$$\liminf_{n\to+\infty} \mathbb{E}\left[\mathrm{d}^i\left(T^i(y_n^i), y_n^i\right)^2\right] \leq \frac{1}{\hat{\alpha}^i}\left\{\frac{2\tilde{B}^i D^i}{(1-\hat{\beta})\mathsf{h}_0^i}\alpha + \frac{\zeta^i \tilde{B^i}^2}{(1-\hat{\beta})^2(\mathsf{h}_0^i)^2}\alpha^2\right\}, \tag{15}$$

and

$$\liminf_{n\to+\infty} \mathbb{E}\left[f(x_n) - f^\star\right] \leq \frac{\sum_{i\in\mathcal{I}} \zeta^i \tilde{B^i}^2 (\mathsf{h}_0^i)^{-1}}{2(1-\beta)(1-\hat{\beta})}\alpha + \frac{\sum_{i\in\mathcal{I}} \tilde{B}^i D^i}{(1-\beta)(1-\hat{\beta})}\beta, \tag{16}$$

where $\hat{\alpha}_i := \alpha^i(1-\alpha^i)$ and $\tilde{B^i}^2 := \max\{\|\tau_{-1}\|^2_{x_0^i}, {B^i}^2\}$. Moreover, for all $i \in \mathcal{I}$ and all $n \geq 1$,

$$\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^n \mathrm{d}^i\left(T^i(y_k^i), y_k^i\right)^2\right] \leq \frac{D^i}{\hat{\alpha}_i}\frac{1}{n} + \frac{2\tilde{B}^i D^i}{\hat{\alpha}_i \hat{\mathsf{h}}_0^i}\alpha + \frac{\zeta^i \tilde{B^i}^2}{\hat{\alpha}_i(\hat{\mathsf{h}}_0^i)^2}\alpha^2, \tag{17}$$

where $\hat{\mathsf{h}}_0^i := (1-\hat{\beta})\mathsf{h}_0^i$. Let us define $\bar{x}_n$ for all $n \geq 1$ by

$$\bar{x}_n := \exp_{\bar{x}_{n-1}}\left(\frac{1}{n}\exp^{-1}_{\bar{x}_{n-1}}(x_n)\right), \tag{18}$$

where $\bar{x}_0 := x_0$. Then, for all $n \geq 1$,

$$\mathbb{E}\left[f(\bar{x}_n) - f_\star\right] \leq \frac{\sum_{i\in\mathcal{I}} \hat{B}^i {D^i}^2}{2(1-\beta_1)}\frac{1}{\alpha n} + \frac{\sum_{i\in\mathcal{I}} \zeta^i \tilde{B^i}^2 (\mathsf{h}_0^i)^{-1}}{2(1-\hat{\beta})(1-\beta_1)}\alpha + \frac{\sum_{i\in\mathcal{I}} \tilde{B}^i D^i}{1-\beta_1}\beta. \tag{19}$$

If (A1)' *$T^i \colon M^i \to M^i$ $(i \in \mathcal{I})$ is nonexpansive with $\mathrm{Fix}(T^i) \neq \emptyset$, then*

$$\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^n \mathrm{d}^i\left(T^i(x_k^i), x_k^i\right)^2\right] \leq \frac{2D^i}{\hat{\alpha}^i}\frac{1}{n} + \frac{4\tilde{B}^i D^i}{\hat{\alpha}^i \hat{\mathsf{h}}_0^i}\alpha + \frac{2\tilde{B^i}^2}{(\hat{\mathsf{h}}_0^i)^2}\left\{\frac{\zeta^i}{\hat{\alpha}^i} + \frac{4}{(1-\hat{\beta})^2}\right\}\alpha^2. \tag{20}$$

The following is a convergence analysis of Algorithm 1 with diminishing step-sizes (the proof of the theorem is given in Supplementary Material).

Theorem 2 *Suppose that Assumptions (A1)–(A5) and Conditions (C1)–(C3) hold and assume that $(\alpha_n)_{n\in\mathbb{N}}$ is monotone decreasing and $(\alpha_n)_{n\in\mathbb{N}}$ and $(\beta_n)_{n\in\mathbb{N}}$ satisfy that*

$$\sum_{n=0}^{+\infty} \alpha_n = +\infty, \sum_{n=0}^{+\infty} \alpha_n^2 < +\infty, \text{ and } \sum_{n=0}^{+\infty} \alpha_n \beta_n < +\infty. \tag{21}$$

Then, Algorithm 1 satisfies that, for all $i \in \mathcal{I}$,

$$\lim_{n\to+\infty} \mathbb{E}\left[\mathrm{d}^i\left(y_n^i, x_n^i\right)^2\right] = 0, \ \liminf_{n\to+\infty} \mathbb{E}\left[\mathrm{d}^i\left(T^i(y_n^i), y_n^i\right)^2\right] = 0, \tag{22}$$

and

$$\liminf_{n\to+\infty} \mathbb{E}\left[f(x_n) - f_\star\right] \leq 0.$$

Suppose that Assumptions (A1)–(A5) and Conditions (C1)–(C3) hold and assume that $(\alpha_n(1-\beta_n))_{n\in\mathbb{N}}$ *and* $(\beta_n)_{n\in\mathbb{N}}$ *are monotone decreasing and satisfy the following:*

$$\lim_{n\to+\infty} \frac{1}{n\alpha_n} = 0, \ \lim_{n\to+\infty} \frac{1}{n}\sum_{k=1}^n \alpha_k = 0, \ \textit{and} \ \lim_{n\to+\infty} \frac{1}{n}\sum_{k=1}^n \beta_k = 0. \tag{23}$$

Then, Algorithm 1 satisfies that

$$\lim_{n\to+\infty} \mathbb{E}\left[\frac{1}{n}\sum_{k=1}^n \sum_{i\in\mathcal{I}} \mathrm{d}^i\left(T^i(y_k^i), y_k^i\right)^2\right] = 0 \tag{24}$$

and

$$\limsup_{n\to+\infty} \mathbb{E}\left[f(\bar{x}_n) - f_\star\right] \leq 0 \tag{25}$$

with the rate of convergence expressed as follows:

$$\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^n \mathrm{d}^i\left(T^i(y_k^i), y_k^i\right)^2\right] \leq \frac{1}{\hat{\alpha}^i}\left\{\frac{D^i}{n} + \frac{2\tilde{B}^i D^i}{\hat{\mathsf{h}}_0^i}\frac{1}{n}\sum_{k=1}^n \alpha_k + \frac{\zeta^i {\tilde{B}^i}^2}{(\hat{\mathsf{h}}_0^i)^2}\frac{1}{n}\sum_{k=1}^n \alpha_k^2\right\},$$
$$\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^n \mathrm{d}^i\left(y_k^i, x_k^i\right)^2\right] \leq \frac{{\tilde{B}^i}^2}{(1-\hat{\beta})^2(\hat{\mathsf{h}}_0^i)^2}\frac{1}{n}\sum_{k=1}^n \alpha_k^2,$$

and

$$\mathbb{E}\left[f(\bar{x}_n) - f_\star\right] \leq \frac{\sum_{i\in\mathcal{I}} \hat{B}^i {D^i}^2}{2(1-\beta_1)}\frac{1}{n\alpha_n} + \frac{\sum_{i\in\mathcal{I}} \zeta^i {\tilde{B}^i}^2 (\mathsf{h}_0^i)^{-1}}{2(1-\hat{\beta})(1-\beta_1)}\frac{1}{n}\sum_{k=1}^n \alpha_k + \frac{\sum_{i\in\mathcal{I}} \tilde{B}^i D^i}{1-\beta_1}\frac{1}{n}\sum_{k=1}^n \beta_k,$$

where $\bar{x}_n$ *is defined by* (18)*. Under Assumption (A1)', we have*

$$\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^n \mathrm{d}^i\left(T^i(x_k^i), x_k^i\right)^2\right] \leq \frac{2}{\hat{\alpha}^i}\frac{D^i}{n} + \frac{4\tilde{B}^i D^i}{\hat{\alpha}^i \hat{\mathsf{h}}_0^i}\frac{1}{n}\sum_{k=1}^n \alpha_k + \frac{2{\tilde{B}^i}^2}{(\hat{\mathsf{h}}_0^i)^2}\left\{\frac{\zeta^i}{\hat{\alpha}^i} + \frac{4}{(1-\hat{\beta})^2}\right\}\frac{1}{n}\sum_{k=1}^n \alpha_k^2.$$

Theorem 2 yields the following corollary.

Corollary 1 *Suppose that Assumptions (A1)–(A5) and Conditions (C1)–(C3) hold. Then, Algorithm 1 with* $\alpha_n := 1/n^\eta$ *(*$\eta \in (1/2, 1], n \geq 1$*) and* $(\beta_n)_{n\in\mathbb{N}}$ *such that* $\sum_{n=1}^{+\infty} \alpha_n\beta_n < +\infty$[1] *satisfies that, for all* $i \in \mathcal{I}$, $\mathbb{E}[\mathrm{d}^i(y_n^i, x_n^i)^2] = \mathcal{O}(n^{-2\eta})$,

$$\liminf_{n\to+\infty} \mathbb{E}\left[\mathrm{d}^i\left(T^i(y_n^i), y_n^i\right)^2\right] = 0, \text{ and } \liminf_{n\to+\infty} \mathbb{E}\left[f(x_n) - f_\star\right] \leq 0. \tag{26}$$

Moreover, Algorithm 1 with $\alpha_n := 1/n^\eta$ *(*$\eta \in [1/2, 1)$*) and* $(\beta_n)_{n\in\mathbb{N}}$ *such that* $\sum_{n=1}^{+\infty} \beta_n < +\infty$[2] *satisfies that, for all* $n \geq 1$,

$$\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n}\sum_{i\in\mathcal{I}} \mathrm{d}^i\left(T^i(y_k^i), y_k^i\right)^2\right] = \mathcal{O}\left(\frac{1}{n^\eta}\right)$$

and

$$\mathbb{E}\left[f(\bar{x}_n) - f_\star\right] \leq \mathcal{O}\left(\frac{1}{n^{1-\eta}}\right), \tag{27}$$

where $\bar{x}_n$ *is defined by* (18)*. Under Assumption (A1)', we have*

$$\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n}\sum_{i\in\mathcal{I}} \mathrm{d}^i\left(T^i(x_k^i), x_k^i\right)^2\right] = \mathcal{O}\left(\frac{1}{n^\eta}\right).$$

5.2 Smooth nonconvex optimization

This subsection considers Problem 1 when f is smooth and nonconvex. The following is a convergence analysis of Algorithm 1 with constant step-sizes (the proof of the theorem is given in Supplementary Material).

Theorem 3 *Suppose that Assumptions (A1)–(A5) and Conditions (C1)–(C3) hold. Then, Algorithm 1 with* $\alpha_n := \alpha$ *and* $\beta_n := \beta$ *satisfies that, for all* $i \in \mathcal{I}$, (14) *and* (15) *hold, and*

$$\limsup_{n\to+\infty} \mathbb{E}\left[\left\langle \exp_{x_n}^{-1}(x), \mathrm{grad} f(x_n)\right\rangle_{x_n}\right] \geq -\frac{\sum_{i\in\mathcal{I}} \zeta^i \tilde{B^i}^2}{2\gamma \mathsf{h}_0^i}\alpha - \frac{\sum_{i\in\mathcal{I}} \tilde{B^i} D^i}{\gamma}\beta, \tag{28}$$

[1] The step-sizes $\beta_n := \lambda^n$ and $\alpha_n := 1/n^\eta$ $(n \geq 1, \lambda \in (0,1), \eta \in (1/2, 1])$ satisfy $\sum_{n=1}^{+\infty} \alpha_n = +\infty$, $\sum_{n=1}^{+\infty} \alpha_n^2 < +\infty$, and $\sum_{n=1}^{+\infty} \alpha_n\beta_n < +\infty$.

[2] The step-sizes $\beta_n := 1/2^n$ and $\alpha_n := 1/n^\eta$ $(n \geq 1, \eta \in [1/2, 1))$ are used to implement adaptive learning rate optimization algorithms, such as Adam [18], AMSGrad [27], and RAMSGrad [5]. These step-sizes satisfy $\sum_{n=1}^{+\infty} \beta_n = 1$ and $(\alpha_n(1-\beta_n))_{n\in\mathbb{N}}$ is monotone decreasing.

where $\gamma := (1-\beta)(1-\hat{\beta})$. *Moreover, for all* $i \in \mathcal{I}$ *and all* $n \geq 1$, (17) *holds and*

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n}\left\langle \exp_{x_k}^{-1}(x), \operatorname{grad} f(x_k)\right\rangle_{x_k}\right]\\ &\geq -\frac{\sum_{i\in\mathcal{I}}\hat{B}^i {D^i}^2}{2(1-\beta_1)}\frac{1}{\alpha n} - \frac{\sum_{i\in\mathcal{I}}\zeta^i {\tilde{B}^i}^2 (\mathsf{h}_0^i)^{-1}}{2(1-\hat{\beta})(1-\beta_1)}\alpha - \frac{\sum_{i\in\mathcal{I}}\tilde{B}^i D^i}{1-\beta_1}\beta.\end{aligned} \tag{29}$$

Under Assumption (A1)', (20) *holds.*

The following is a convergence analysis of Algorithm 1 with diminishing step-sizes (the proof of the theorem is given in Supplementary Material).

Theorem 4 *Suppose that Assumptions (A1)–(A5) and Conditions (C1)–(C3) hold and assume that* $(\alpha_n)_{n\in\mathbb{N}}$ *is monotone decreasing and* $(\alpha_n)_{n\in\mathbb{N}}$ *and* $(\beta_n)_{n\in\mathbb{N}}$ *satisfy* (21). *Then, Algorithm 1 satisfies that, for all* $i \in \mathcal{I}$, (22) *holds and*

$$\limsup_{n\to+\infty} \mathbb{E}\left[\left\langle \exp_{x_n}^{-1}(x), \operatorname{grad} f(x_n)\right\rangle_{x_n}\right] \geq 0.$$

Suppose that Assumptions (A1)–(A5) and Conditions (C1)–(C3) hold and assume that $(\alpha_n(1-\beta_n))_{n\in\mathbb{N}}$ *and* $(\beta_n)_{n\in\mathbb{N}}$ *are monotone decreasing and satisfy* (23). *Then, Algorithm 1 satisfies that* (24) *holds and*

$$\liminf_{n\to+\infty} \mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n}\left\langle \exp_{x_k}^{-1}(x), \operatorname{grad} f(x_k)\right\rangle_{x_k}\right] \geq 0$$

with

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n}\left\langle \exp_{x_k}^{-1}(x), \operatorname{grad} f(x_k)\right\rangle_{x_k}\right]\\ &\geq -\frac{\sum_{i\in\mathcal{I}}\hat{B}^i {D^i}^2}{2(1-\beta_1)}\frac{1}{n\alpha_n} - \frac{\sum_{i\in\mathcal{I}}\zeta^i {\tilde{B}^i}^2 (\mathsf{h}_0^i)^{-1}}{2(1-\hat{\beta})(1-\beta_1)}\frac{1}{n}\sum_{k=1}^{n}\alpha_k - \frac{\sum_{i\in\mathcal{I}}\tilde{B}^i D^i}{1-\beta_1}\frac{1}{n}\sum_{k=1}^{n}\beta_k\end{aligned}$$

and the same convergence rate of $\mathrm{d}^i(T^i(y_k^i), y_k^i)$ *and* $\mathrm{d}^i(T^i(x_k^i), x_k^i)$ *(under Assumption (A1)') as in Theorem 2.*

A discussion similar to the one for obtaining Corollary 1 implies that Algorithm 1 with $\alpha_n := 1/n^\eta$ ($\eta \in [1/2, 1)$) and $(\beta_n)_{n\in\mathbb{N}}$ such that $\sum_{n=1}^{+\infty}\beta_n < +\infty$ satisfies that, under Assumption (A1)',

$$\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n}\sum_{i\in\mathcal{I}}\mathrm{d}^i\left(T^i(x_k^i), x_k^i\right)^2\right] = \mathcal{O}\left(\frac{1}{n^\eta}\right)$$

and

$$\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n}\left\langle \exp_{x_k}^{-1}(x), \operatorname{grad} f(x_k)\right\rangle_{x_k}\right] \geq -\mathcal{O}\left(\frac{1}{n^{1-\eta}}\right).$$

6 Numerical Comparisons

6.1 Preliminaries

The m-dimensional Poincaré disk model of hyperbolic space is defined by

$$\mathcal{D}^m := \{x \in \mathbb{R}^m \colon \|x\| < 1\},$$

where $\|\cdot\|$ denotes the Euclidean norm of $\mathbb{R}^m$. Let us also define $M := \underbrace{\mathcal{D}^m \times \mathcal{D}^m \times \cdots \times \mathcal{D}^m}_{I}$. Let $j \in \mathcal{J}^i := \{1, 2, \ldots, J^i\}$ ($i \in \mathcal{I} := \{1, 2, \ldots, I\}$). We define a closed ball with center $c_j^i \in \mathcal{D}^m$ and radius $r_j^i > 0$ in $\mathcal{D}^m$ by

$$\mathrm{B}_j^i := \left\{x \in \mathcal{D}^m \colon \mathrm{d}\left(c_j^i, x\right) \leq r_j^i\right\}, \tag{30}$$

where $\mathrm{d} \colon \mathcal{D}^m \times \mathcal{D}^m \to \mathbb{R}$ denotes the distance function of $\mathcal{D}^m$. Then, the metric projection onto the closed convex set B_j^i can be expressed as follows:

$$P_j^i(x) := \begin{cases} \exp_{c_j^i}^i \left(\dfrac{r_j^i \left(\exp_{c_j^i}^i\right)^{-1}(x)}{\left\| \left(\exp_{c_j^i}^i\right)^{-1}(x) \right\|_{c_j^i}} \right) & \text{if } x \notin \mathrm{B}_j^i, \\ x & \text{if } x \in \mathrm{B}_j^i. \end{cases}$$

We used the nonexpansive mapping $T^i \colon D^m \to D^m$ ($i \in \mathcal{I}$) defined by

$$T^i := P_1^i P_2^i \cdots P_{J^i}^i \tag{31}$$

and the smooth, nonconvex function $f \colon M \to \mathbb{R}$ defined for all $x \in M$ by

$$f(x) = \frac{1}{I} \sum_{i=1}^{I} \underbrace{\left\{ e^{(x^i)^\top x^j} + (x^i)^\top x^j \right\}}_{=F(x,i)}, \text{ where } j := (i \bmod I) + 1.$$

We implemented the following algorithms, all with $\alpha^i := 0.5$:

- Algorithm 1 with constant step-sizes
  - CSD: Algorithm 1 with h_n^i defined by Stochastic Gradient Descent [7] (i.e., $\mathsf{h}_n^i := 1$), $\alpha_n := 10^{-2}$, and $\beta_n = \hat{\beta} := 0$
  - CAG: Algorithm 1 with h_n^i defined by AdaGrad [10], $\alpha_n := 10^{-2}$, and $\beta_n = \hat{\beta} := 0$
  - CAM1: Algorithm 1 with h_n^i defined by AMSGrad (13), $\alpha_n := 10^{-2}$, $\beta_n = 0.9$, $\hat{\beta} := 0$, and $\bar{\beta} := 0.999$
  - CAM2: Algorithm 1 with h_n^i defined by AMSGrad (13), $\alpha_n := 10^{-2}$, $\beta_n = 10^{-3}$, $\hat{\beta} := 0$, and $\bar{\beta} := 0.999$
  - CAD1: Algorithm 1 with h_n^i defined by Adam (12), $\alpha_n := 10^{-2}$, $\beta_n = 0.9$, $\hat{\beta} := 0.9$, and $\bar{\beta} := 0.999$ ($\hat{\beta} := 0.9$ and $\bar{\beta} := 0.999$ were used in [18,27])
  - CAD2: Algorithm 1 with h_n^i defined by Adam (12), $\alpha_n := 10^{-2}$, $\beta_n = 10^{-3}$, $\hat{\beta} := 0.9$, and $\bar{\beta} := 0.999$

– Algorithm 1 with diminishing step-sizes
  DSD: Algorithm 1 with h_n^i defined by Stochastic Gradient Descent [7] (i.e., $\mathsf{h}_n^i := 1$), $\alpha_n := 10^{-1}/\sqrt{n}$, and $\beta_n = \hat{\beta} := 0$
  DAG: Algorithm 1 with h_n^i defined by AdaGrad [10], $\alpha_n := 10^{-1}/\sqrt{n}$, and $\beta_n = \hat{\beta} := 0$
  DAM1: Algorithm 1 with h_n^i defined by AMSGrad (13), $\alpha_n := 10^{-1}/\sqrt{n}$, $\beta_n = 0.5^n$, $\hat{\beta} := 0$, and $\bar{\beta} := 0.999$
  DAM2: Algorithm 1 with h_n^i defined by AMSGrad (13), $\alpha_n := 10^{-1}/\sqrt{n}$, $\beta_n = 0.9^n$, $\hat{\beta} := 0$, and $\bar{\beta} := 0.999$
  DAD1: Algorithm 1 with h_n^i defined by Adam (12), $\alpha_n := 10^{-1}/\sqrt{n}$, $\beta_n = 0.5^n$, $\hat{\beta} := 0.9$, and $\bar{\beta} := 0.999$
  DAD2: Algorithm 1 with h_n^i defined by Adam (12), $\alpha_n := 10^{-1}/\sqrt{n}$, $\beta_n = 0.9^n$, $\hat{\beta} := 0.9$, and $\bar{\beta} := 0.999$

The difference between CAM1 (resp. CAD1) and CAM2 (resp. CAD2) is the setting of β_n. The step-size $\beta_n = 0.9$ in CAM1 (resp. CAD1) is based on previously reported results (see, e.g., [5, Section 5]), while the step-size $\beta_n = 10^{-3}$ is based on Theorem 3 indicating that a small step-size approximates a solution to Problem 1. The algorithms with diminishing step-sizes all used $\alpha_n := 10^{-1}/\sqrt{n}$, which is based on previously reported results (see, e.g., [5, Theorems 1 and 2]).

Ten samplings, each starting from a different randomly chosen initial point $x_0(s) \in M$ $(s = 1, 2, \ldots, 10)$, were performed, and the results were averaged. The following two performance measures were used: for each $n \in \mathbb{N}$,

$$D_n := \frac{1}{10}\sum_{s=1}^{10}\sqrt{\sum_{i\in\mathcal{I}} \mathrm{d}\left(x_n^i(s), T^i(x_n^i(s))\right)^2} \text{ and } F_n := \frac{I}{10}\sum_{s=1}^{10} f(x_n(s)),$$

where $(x_n^i(s))_{n\in\mathbb{N}}$ denotes the sequence generated by Algorithm 1 with an initial point $x_0(s)$. If $(D_n)_{n\in\mathbb{N}}$ converges to 0, then Algorithm 1 converges to a fixed point of T^i.

The experiments were conducted on a MacBook Air (2017) with a 1.8 GHz Intel Core i5 CPU, 8 GB 1600 MHz DDR3 memory, and the macOS Mojave version 10.14.5 operating system. The algorithms were written in Python 3.7.6 with the NumPy 1.19.2 package and the Matplotlib 3.1.2 package.

6.2 Consistent case

We first consider the consistent case such that $\bigcap_{j\in\mathcal{J}^i} \mathrm{B}_j^i \neq \emptyset$ $(m = 2, 10, 100; I = 5; J_i = 5)$, where $c_j^i \in M$ and $r_j^i > 0$ in B_j^i defined by (30) were randomly chosen. A nonexpansive mapping $T^i \colon D^m \to D^m$ $(i \in \mathcal{I})$ defined by (31) satisfies $\mathrm{Fix}(T^i) = \bigcap_{j\in\mathcal{J}^i} \mathrm{B}_j^i$ (see also Proposition 2.3(ii) and Example 1).

Tables 1 and 2 show the average elapsed time (s) for the algorithms used in the experiment for $n = 500$ when $m = 2$, $n = 1000$ when $m = 10$, and $n = 1500$ when $m = 100$. The results in these tables indicate that the elapsed

times of the algorithms with constant step-sizes varied little from the elapsed times of the algorithms with diminishing step-sizes and that, for a fixed m, all of the algorithms ran at about the same speed.

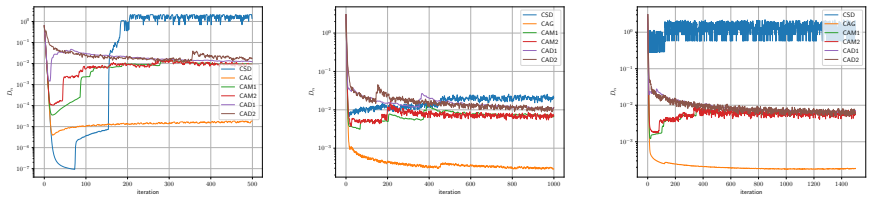
Table 1 Average time for the algorithms with constant step-sizes applied to consistent case

	CSD	CAG	CAM1	CAM2	CAD1	CAD2
$m = 2$	7.728	7.782	8.128	8.115	7.892	7.878
$m = 10$	16.219	16.381	16.974	16.534	16.546	17.077
$m = 100$	23.683	23.788	23.907	24.347	24.536	24.187

Table 2 Average time for the algorithms with diminishing step-sizes applied to consistent case

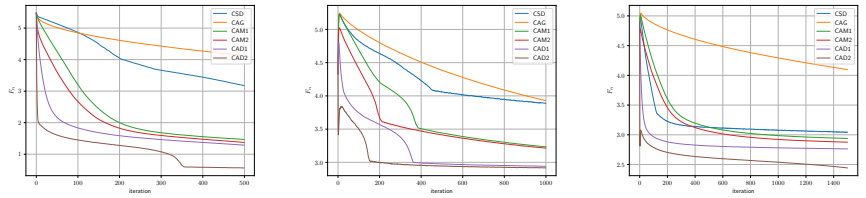
	DSD	DAG	DAM1	DAM2	DAD1	DAD2
$m = 2$	7.862	7.906	8.382	8.155	7.989	8.364
$m = 10$	16.216	16.635	16.706	16.440	16.373	16.962
$m = 100$	23.085	23.799	23.545	23.781	23.761	23.435

Figures 1 and 2 show the behaviors of D_n and F_n for the algorithms with constant step-sizes, and Figures 3 and 4 show the behaviors of D_n and F_n for the algorithms with diminishing step-sizes. The results in these figures indicate that all algorithms except for CSD, DSD, CAG, and DAG performed well. Although CAG and DAG converged to fixed points of T^i faster than the other algorithms, CAG and DAG did not minimize f. This is because CAG and DAG used $\beta_n = 0$ (i.e., $m_n = \mathsf{G}(x_n, \xi_n)$), which means that CAG and DAG attached more weight to converging to a point in $X = \mathrm{Fix}(T^1) \times \mathrm{Fix}(T^2) \times \cdots \times \mathrm{Fix}(T^I)$ than minimizing f. To verify why CSD and DSD did not converge to a fixed point of T^i, we checked the behaviors of CSD and DSD for ten samplings. CSD and DSD were sometimes good and sometimes not within ten samplings. As a result, the mean value D_n of $\sqrt{\sum_{i\in\mathcal{I}} \mathrm{d}(x_n^i(s), T^i(x_n^i(s)))^2}$ for CSD and DSD was not minimized.



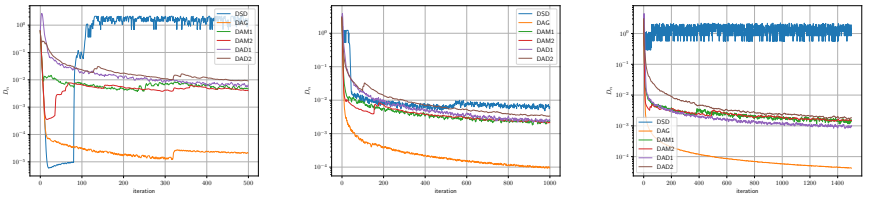
(a) $m = 2$, $I = 5$, $J_i = 5$ (b) $m = 10$, $I = 5$, $J_i = 5$ (c) $m = 100$, $I = 5$, $J_i = 5$

Fig. 1 D_n vs. iteration for Algorithm 1 with constant step-sizes (consistent case)



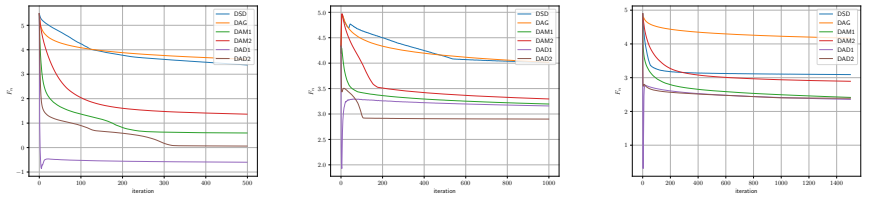
(a) $m = 2$, $I = 5$, $J_i = 5$ (b) $m = 10$, $I = 5$, $J_i = 5$ (c) $m = 100$, $I = 5$, $J_i = 5$

Fig. 2 F_n vs. iteration for Algorithm 1 with constant step-sizes (consistent case)



(a) $m = 2$, $I = 5$, $J_i = 5$ (b) $m = 10$, $I = 5$, $J_i = 5$ (c) $m = 100$, $I = 5$, $J_i = 5$

Fig. 3 D_n vs. iteration for Algorithm 1 with diminishing step-sizes (consistent case)



(a) $m = 2$, $I = 5$, $J_i = 5$ (b) $m = 10$, $I = 5$, $J_i = 5$ (c) $m = 100$, $I = 5$, $J_i = 5$

Fig. 4 F_n vs. iteration for Algorithm 1 with diminishing step-sizes (consistent case)

6.3 Inconsistent case

We next consider the inconsistent case such that $\bigcap_{j\in\mathcal{J}^i} \mathrm{B}_j^i = \emptyset$, where $c_j^i \in M$ and $r_j^i > 0$ in B_j^i ($m = 2, 10, 100; I = 5; \mathcal{J}^i = \{1, 2\}$) defined by (30) were randomly chosen so that $\bigcap_{j\in\mathcal{J}^i} \mathrm{B}_j^i = \emptyset$ was satisfied. Here, we define a *generalized convex feasible set* (see [9, Section I, Framework 2] and [37, Definition 4.1] for the definition under the Hilbert space setting) as follows:

$$C_{\mathrm{d}}^i := \left\{ x \in \mathrm{B}_1^i \colon \mathrm{d}\left(x, \mathrm{B}_2^i\right)^2 = \inf_{y\in\mathrm{B}_1^i} \mathrm{d}\left(y, \mathrm{B}_2^i\right)^2 \right\}. \tag{32}$$

The generalized convex feasible set plays an important role when the constraint set composed of the absolute set and the subsidiary set is not feasible. Let B_1^i be the absolute constrained set and B_2^i be the subsidiary constrained set. Then,

C^i_{d} is feasible (i.e., $C^i_{\mathrm{d}} \neq \emptyset$) even when $\mathrm{B}^i_1 \cap \mathrm{B}^i_2 = \emptyset$. Moreover, C^i_{d} is a subset of the absolute constrained set B^i_1 with the elements closest to the subsidiary constrained set B^i_2 in terms of the distance function. Accordingly, it would be reasonable to replace an inconsistent set with the generalized convex feasible set. The set C^i_{d} defined by (32) can be expressed as follows:

$$C^i_{\mathrm{d}} = \mathrm{Fix}\left(P^i_1\left(\exp\left[-\mathrm{grad}\,\frac{1}{2}\mathrm{d}\left(\cdot,\mathrm{B}^i_2\right)^2\right]\right)\right) = \mathrm{Fix}\left(P^i_1 P^i_2\right) = \mathrm{Fix}\left(T^i\right),$$

where the first equation comes from [23, Proposition 3.1, Corollaries 3.1 and 3.2, Theorem 3.3] (see also Proposition 3.1), the second equation comes from $\mathrm{grad}(1/2)\mathrm{d}(x,y)^2 = -\exp_x^{-1}(y)$ [11, Proposition 3.3], and the third equation comes from (31).

Tables 3 and 4 show that the elapsed times of the algorithms with constant step-sizes differed little from the elapsed times of the algorithms with diminishing step-sizes and that, for a fixed m, all of the algorithms ran at about the same speed.

Table 3 Average time for the algorithms with constant step-sizes applied to inconsistent case

	CSD	CAG	CAM1	CAM2	CAD1	CAD2
$m = 2$	3.906	3.878	4.010	4.014	3.940	3.933
$m = 10$	8.649	8.675	8.822	9.066	8.727	8.691
$m = 100$	13.727	13.831	14.303	14.021	14.003	14.061

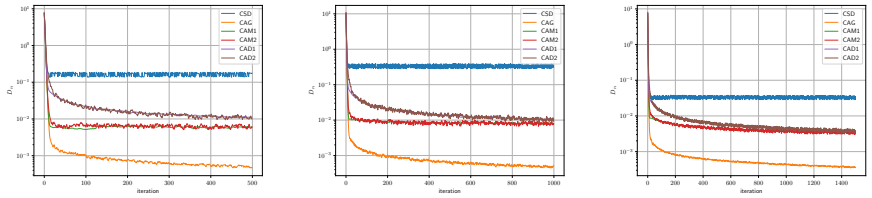
Table 4 Average time for the algorithms with diminishing step-sizes applied to inconsistent case

	DSD	DAG	DAM1	DAM2	DAD1	DAD2
$m = 2$	3.883	3.880	4.013	4.075	4.010	3.946
$m = 10$	8.783	8.739	9.043	9.105	8.912	8.910
$m = 100$	13.571	13.908	13.971	13.983	13.956	13.970

Figures 5 and 6 show the behaviors of D_n and F_n for the algorithms with constant step-sizes, and Figures 7 and 8 show the behaviors of D_n and F_n for the algorithms with diminishing step-sizes. The results shown in these figures indicate that all algorithms except for CSD, DSD, CAG, and DAG performed well, the same as in the consistent case (previous subsection).We checked the behaviors of CSD, DSD, CAG, and DAG and found that the reason they did not perform well was the same as in that case.

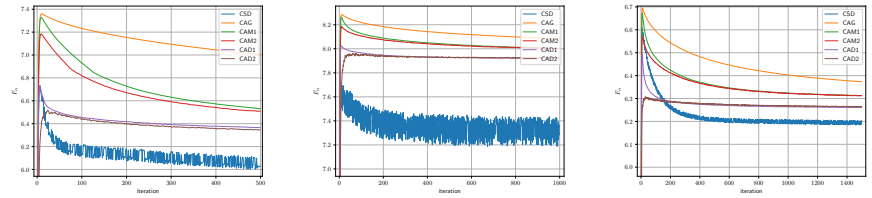
7 Conclusion

This paper proposed the Riemannian stochastic fixed point optimization algorithm for stochastic optimization with fixed point constraints of quasinonexpansive mappings defined on Riemannian manifolds. It also gave convergence



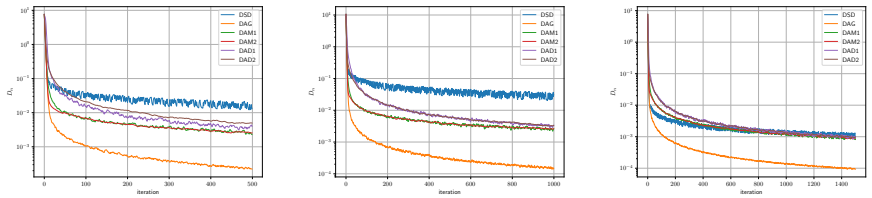
(a) $m = 2$, $I = 5$, $J_i = 2$ (b) $m = 10$, $I = 5$, $J_i = 2$ (c) $m = 100$, $I = 5$, $J_i = 2$

Fig. 5 D_n vs. iteration for Algorithm 1 with constant step-sizes (inconsistent case)



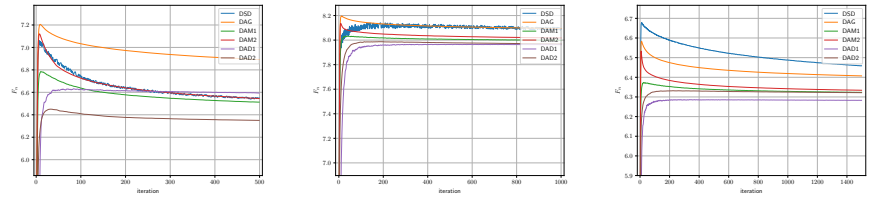
(a) $m = 2$, $I = 5$, $J_i = 2$ (b) $m = 10$, $I = 5$, $J_i = 2$ (c) $m = 100$, $I = 5$, $J_i = 2$

Fig. 6 F_n vs. iteration for Algorithm 1 with constant step-sizes (inconsistent case)



(a) $m = 2$, $I = 5$, $J_i = 2$ (b) $m = 10$, $I = 5$, $J_i = 2$ (c) $m = 100$, $I = 5$, $J_i = 2$

Fig. 7 D_n vs. iteration for Algorithm 1 with diminishing step-sizes (inconsistent case)



(a) $m = 2$, $I = 5$, $J_i = 2$ (b) $m = 10$, $I = 5$, $J_i = 2$ (c) $m = 100$, $I = 5$, $J_i = 2$

Fig. 8 F_n vs. iteration for Algorithm 1 with diminishing step-sizes (inconsistent case)

analyses of the algorithm for both constant and diminishing step-sizes and both nonsmooth convex and smooth nonconvex optimization. For small constant step-sizes, the analyses showed that the algorithm can approximate a solution to the problem. For diminishing step-sizes, the analyses suggested the general rate of convergence of the algorithm. Finally, the optimality and convergence of the algorithm with each of the formulas based on the adaptive learning rate optimization algorithms were demonstrated through numerical comparisons. In the process, the algorithms with formulas based on Adam and AMSGrad were found to be superior for performing stochastic Riemannian optimization with fixed point constraints.

References

1. Bauschke, H.H., Borwein, J.M.: On projection algorithms for solving convex feasibility problems. SIAM Review **38**, 367–426 (1996)
2. Bauschke, H.H., Chen, J.: A projection method for approximating fixed points of quasi nonexpansive mappings without the usual demiclosedness condition. Journal of Nonlinear and Convex Analysis **15**, 129–135 (2014)
3. Bauschke, H.H., Combettes, P.L.: A weak-to-strong convergence principle for Fejér-monotone methods in Hilbert space. Mathematics of Operations Research **26**, 248–264 (2001)
4. Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces, 2nd edn. Springer, New York (2017)
5. Bécigneul, G., Ganea, O.E.: Riemannian adaptive optimization methods. Proceedings of The International Conference on Learning Representations pp. 1–16 (2019)
6. Bento, G.C., Melo, J.G.: Subgradient method for convex feasibility on Riemannian manifolds. Journal of Optimization Theory and Applications **152**, 773–785 (2012)
7. Bonnabel, S.: Stochastic gradient descent on Riemannian manifolds. IEEE Transactions on Automatic Control **58**, 2217–2229 (2013)
8. Chaoha, P., Phon-on, A.: A note on fixed point sets in CAT(0) spaces. Journal of Mathematical Analysis and Applications **320**, 983–987 (2006)
9. Combettes, P.L., Bondon, P.: Hard-constrained inconsistent signal feasibility problems. IEEE Transactions on Signal Processing **47**, 2460–2468 (1999)
10. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research **12**, 2121–2159 (2011)
11. Ferreira, O., Oliveira, P.R.: Proximal point algorithm on Riemannian manifolds. Optimization **51**, 257–270 (2002)
12. Goebel, K., Reich, S.: Uniform Convexity, Hyperbolic Geometry, and Nonexpansive Mappings. Dekker, New York and Basel (1984)
13. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge (2016)
14. Grohs, P., Hosseini, S.: ε-subgradient algorithms for locally lipschitz functions on Riemannian manifolds. Advances in Computational Mathematics **42**, 333–360 (2016)
15. Hawe, S., Kleinsteuber, M., Diepold, K.: Analysis operator learning and its application to image reconstruction. IEEE Transactions on Image Processing **22**, 2138–2150 (2013)
16. Iiduka, H.: Stochastic fixed point optimization algorithm for classifier ensemble. IEEE Transactions on Cybernetics **50**, 4370–4380 (2020)
17. Kasai, H., Jawanpuria, P., Mishra, B.: Riemannian adaptive stochastic gradient algorithms on matrix manifolds. International Conference on Machine Learning pp. 3262–3271 (2019)
18. Kingma, D.P., Ba, J.L.: Adam: A method for stochastic optimization. Proceedings of The International Conference on Learning Representations pp. 1–15 (2015)
19. Kirk, W.A.: Geodesic geometry and fixed point theory II. pp. 113–142. in: Proceedings of the International Conference in Fixed Point Theory and Applications, Valencia, Spain (2003)

20. Li, C., López, G., Martín-Márquez, V.: Monotone vector fields and the proximal point algorithm on Hadamard manifolds. Journal of the London Mathematical Society **79**, 663–683 (2009)
21. Li, C., López, G., Martín-Márquez, V.: Iterative algorithms for nonexpansive mappings on Hadamard manifolds. Taiwanese Journal of Mathematics **14**, 541–559 (2010)
22. Li, C., López, G., Martín-Márquez, V., Wang, J.H.: Resolvents of set-valued monotone vector fields in Hadamard manifolds. Set-Valued Analysis **19**, 361–383 (2011)
23. Li, S.L., Li, C., Liou, Y.C., Yao, J.C.: Existence of solutions for variational inequalities on Riemannian manifolds. Nonlinear Analysis: Theory, Methods and Applications **71**, 5695–5706 (2009)
24. Liu, Y., Shang, F., Cheng, J., Cheng, H., Jiao, L.: Accelerated first-order methods for geodesically convex optimization on Riemannian manifolds. Advances in Neural Information Processing Systems **30**, 4868–4877 (2017)
25. Nemirovski, A., Juditsky, A., Lan, G., Shapiro, A.: Robust stochastic approximation approach to stochastic programming. SIAM Journal on Optimization **19**, 1574–1609 (2009)
26. Nickel, M., Kiela, D.: Poincaré embeddings for learning hierarchical representations. Advances in Neural Information Processing Systems **30**, 6338–6347 (2017)
27. Reddi, S.J., Kale, S., Kumar, S.: On the convergence of Adam and beyond. Proceedings of The International Conference on Learning Representations pp. 1–23 (2018)
28. Ring, W., Wirth, B.: Optimization methods on Riemannian manifolds and their application to shape space. SIAM Journal on Optimization **22**, 596–627 (2012)
29. Sakai, H., Iiduka, H.: Riemannian adaptive optimization algorithm and its application to natural language processing. IEEE Transactions on Cybernetics (2021)
30. Sakai, T.: Riemannian Geometry. Translations of Mathematical Monographs. American Marhmarical Society, Providence (1996)
31. Sato, H.: A Dai-Yuan-type Riemannian conjugate gradient method with the weak Wolfe conditions. Computational Optimization and Applications **64**, 101–118 (2016)
32. Sato, H., Iwai, T.: A new, globally convergent Riemannian conjugate gradient method. Optimization **64**, 1011–1031 (2015)
33. Sato, H., Kasai, H., Mishra, B.: Riemannian stochastic variance reduced gradient algorithm with retraction and vector transport. SIAM Journal on Optimization **29**, 1444–1472 (2019)
34. Tieleman, T., Hinton, G.: Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural networks for machine learning COURSERA: Neural networks for machine learning **4**, 26–31 (2012)
35. Vasin, V.V., Ageev, A.L.: Ill-posed problems with a priori information. V.S.P. Intl Science, Utrecht (1995)
36. Wang, X., Li, C., Wang, J., Yao, J.H.: Linear convergence of subgradient algorithm for convex feasibility on Riemannian manifolds. SIAM Journal on Optimization **25**, 2334–2358 (2015)
37. Yamada, I.: The hybrid steepest descent method for the variational inequality problem over the intersection of fixed point sets of nonexpansive mappings. In: D. Butnariu, Y. Censor, S. Reich (eds.) Inherently Parallel Algorithms for Feasibility and Optimization and Their Applications, pp. 473–504. Elsevier, New York (2001)
38. Zhang, H., Sra, S.: First-order methods for geodesically convex optimization. Journal of Machine Learning Research **49**, 1–22 (2016)

Declarations

Funding: This work was supported by JSPS KAKENHI Grant Number JP18K11184. **Conflicts of interest/Competing interests:** The authors declare that they have no conflict of interest. **Availability of data and material:** Not applicable. **Code availability:** Not applicable. However, after this paper is accepted for publication, Python implementations of the algorithms used in the numerical experiments will be available at `https://github.com/iiduka-researches/202011-fixed-ropt`.

Supplementary Material

Proofs of Propositions 2.2 and 2.3

Proof (Proof of Proposition 2.2)
(i) This follows from the definitions of firmly nonexpansive, nonexpansive, and quasinonexpansive mappings.

(ii) We prove that (1) implies (5). The comparison theorem for triangles (see, e.g., [21, Proposition 2.2]), together with [22, Proposition 5], ensures that, for all $x \in C$ and all $y \in \mathrm{Fix}(T)$,

$$\mathrm{d}(x,T(x))^2 + \mathrm{d}(T(x),y)^2 - 2\left\langle \exp^{-1}_{T(x)}(x), \exp^{-1}_{T(x)}(y)\right\rangle_{T(x)} \leq \mathrm{d}(x,y)^2,$$
$$\left\langle \exp^{-1}_{T(x)}(x), \exp^{-1}_{T(x)}(y)\right\rangle_{T(x)} \leq 0,$$

which implies (5). From (3), (4), and (5), we have that (5) implies (4), and (4) implies (3). □

Proof (Proof of Proposition 2.3)
(i) This follows from [22, Corollary 1].

(ii) Proposition 2.2(i) and Proposition 2.3(i) imply that P_j is nonexpansive. Accordingly, $T := P_1P_2\cdots P_J$ is nonexpansive. Proposition 2.2(ii) also ensures that P_j is strictly quasinonexpansive. Hence, the proofs of [4, Proposition 4.9, Corollary 4.50] lead to Proposition 2.3(ii).

(iii) This follows from [22, Theorem 4(i)].

(iv) The resolvent of ∂g coincides with the Moreau-Yosida regularization of g. Accordingly, Proposition 2.3(iii) implies Proposition 2.3(iv).

(v) From the definition of $P_{g,\lambda}$, we have that $\mathrm{lev}_{\leq 0}(g) \subset \mathrm{Fix}(P_{g,\lambda})$. To show that $\mathrm{lev}_{\leq 0}(g) \supset \mathrm{Fix}(P_{g,\lambda})$, we assume that $x \in \mathrm{Fix}(P_{g,\lambda})$ and $x \notin \mathrm{lev}_{\leq 0}(g)$. Then, the definition of $u_x \in \partial g(x)$ and the condition $x \notin \mathrm{lev}_{\leq 0}(g)$ guarantee that, for all $y \in \mathrm{lev}_{\leq 0}(g)$,

$$\left\langle u_x, \exp_x^{-1}(y)\right\rangle_x \leq g(y) - g(x) \leq -g(x) < 0,$$

which implies that u_x is not equal to the zero element 0_x of T_xM. Accordingly, the definition of $P_{g,\lambda}$ and the condition $x \in \mathrm{Fix}(P_{g,\lambda})$ mean that

$$\exp_x\left(-\lambda \frac{g(x)}{\|u_x\|_x} u_x\right) = P_{g,\lambda}(x) = x,$$

which implies that

$$0 = \mathrm{d}\left(\exp_x\left(-\lambda \frac{g(x)}{\|u_x\|_x} u_x\right), x\right) = \lambda \frac{g(x)}{\|u_x\|_x}.$$

From $\lambda > 0$ and $u_x \neq 0_x$, we have that $g(x) = 0$, which is a contradiction from $x \notin \mathrm{lev}_{\leq 0}(g)$. Hence, $\mathrm{lev}_{\leq 0}(g) \supset \mathrm{Fix}(P_{g,\lambda})$.

(vi) Lemma 5 in [38] and the definition of $P_{h_j,\lambda}$ ensure that there exists $\zeta = \zeta(\kappa, D) = \sqrt{|\kappa|}D/\tanh(\sqrt{|\kappa|}D) \in \mathbb{R}_+$ such that, for all $x \in C \backslash \mathrm{lev}_{\leq 0}(h_j)$ and all $y \in \mathrm{lev}_{\leq 0}(h_j) = \mathrm{Fix}(P_{h_j,\lambda})$ (by Proposition 2.3(v)),

$$\begin{aligned}\mathrm{d}(P_{h_j,\lambda}(x), y)^2 &\leq \zeta \mathrm{d}(P_{h_j,\lambda}(x), x)^2 + \mathrm{d}(x, y)^2 + 2\left\langle \exp_x^{-1}\left(P_{h_j,\lambda}(x)\right), \exp_x^{-1}(y)\right\rangle_x \\ &= \zeta \mathrm{d}(P_{h_j,\lambda}(x), x)^2 + \mathrm{d}(x, y)^2 + 2\lambda \frac{h_j(x)}{\|u_{j,x}\|_x^2}\left\langle u_{j,x}, \exp_x^{-1}(y)\right\rangle_x,\end{aligned}$$

where ($0_x \neq$) $u_{j,x} \in \partial h_j(x)$, which, together with the definitions of $P_{h_j,\lambda}$ and $u_{j,x} \in \partial h_j(x)$, implies that, for $\lambda \in (0, 2/\zeta)$,

$$\begin{aligned}\mathrm{d}(P_{h_j,\lambda}(x), y)^2 &\leq \mathrm{d}(x, y)^2 + \zeta\lambda^2 \frac{h_j(x)^2}{\|u_{j,x}\|_x^2} - 2\lambda \frac{h_j(x)^2}{\|u_{j,x}\|_x^2} \\ &= \mathrm{d}(x, y)^2 + \lambda(\zeta\lambda - 2)\frac{h_j(x)^2}{\|u_{j,x}\|_x^2}.\end{aligned}$$

From $h_j(x) > 0$, $\mathrm{d}(P_{h_j,\lambda}(x), y) < \mathrm{d}(x, y)$, i.e., $P_{h_j,\lambda}$ is strictly quasinonexpansive.

(vii) Proposition 2.3(vi) and the proofs of [4, Proposition 4.9, Corollary 4.50] lead to Proposition 2.3(vii). □

Proofs of Theorems 1, 2, 3, and 4

The history of the process $\xi_0, \xi_1, \ldots$ up to time n is denoted by $\xi_{[n]} = (\xi_0, \xi_1, \ldots, \xi_n)$. Let $\mathbb{E}[X|\xi_{[n]}]$ denote the conditional expectation of X given $\xi_{[n]} = (\xi_0, \xi_1, \ldots, \xi_n)$. Unless stated otherwise, all relations between random variables are supported to hold almost surely.

We prove the following lemma.

Lemma 1 *Suppose that Assumptions (A1)–(A3) and Conditions (C1)–(C2) hold and consider the sequences $(x_n)_{n\in\mathbb{N}}$, $(m_n)_{n\in\mathbb{N}}$, and $(\mathsf{d}_n)_{n\in\mathbb{N}}$ defined by Algorithm 1. Define y_n^i for all $i \in \mathcal{I}$ and all $n \in \mathbb{N}$ by*

$$y_n^i := \exp_{x_n^i}^i\left(\alpha_n \mathsf{d}_n^i\right) = \exp_{x_n^i}^i\left(-\alpha_n \frac{\hat{m}_n^i}{\mathsf{h}_n^i}\right).$$

Then, for all $i \in \mathcal{I}$, there exists a positive number ζ^i such that, for all $x^i \in X^i$ and all $n \in \mathbb{N}$, almost surely

$$\begin{aligned}\mathrm{d}^i(x_{n+1}^i, x^i)^2 \leq \mathrm{d}^i(x_n^i, x^i)^2 &+ \frac{2\alpha_n}{(1-\hat{\beta}^{n+1})\mathsf{h}_n^i}\left\langle m_n^i, \left(\exp_{x_n^i}^i\right)^{-1}(x^i)\right\rangle_{x_n^i} \\ &+ \frac{\zeta^i \alpha_n^2}{(1-\hat{\beta}^{n+1})^2}\frac{\left\|m_n^i\right\|_{x_n^i}^2}{(\mathsf{h}_n^i)^2} - \alpha^i(1-\alpha^i)\mathrm{d}^i\left(T^i(y_n^i), y_n^i\right)^2.\end{aligned} \tag{33}$$

Moreover, under (C3), for all $i \in \mathcal{I}$, there exists a positive number $\tilde{B^i}^2 := \max\{\|\tau_{-1}\|_{x_0^i}^2, B^{i\,2}\}$ such that, for all $n \in \mathbb{N}$, $\mathbb{E}[\|m_n^i\|_{x_n^i}^2] \leq \tilde{B^i}^2$.

Proof Lemma 5 in [38], together with Assumption (A3) (see also (11)), guarantees that, for all $i \in \mathcal{I}$, there exists $\zeta^i = \zeta(\kappa^i, D^i) \in \mathbb{R}_+$ such that, for all $x^i \in X^i$ and all $n \in \mathbb{N}$,

$$\mathrm{d}^i(y_n^i, x^i)^2 \leq \zeta^i \mathrm{d}^i\left(y_n^i, x_n^i\right)^2 + \mathrm{d}^i(x_n^i, x^i)^2 + 2\alpha_n \left\langle \frac{\hat{m}_n^i}{\mathsf{h}_n^i}, \left(\exp_{x_n^i}^i\right)^{-1}(x^i) \right\rangle_{x_n^i},$$

where κ^i denotes the lower bound of curvature of M^i and $\zeta(\kappa^i, c) := \sqrt{|\kappa^i|}c/\tanh(\sqrt{|\kappa^i|}c)$ for $c \in \mathbb{R}_+$. From the definitions of y_n^i and $\hat{m}_n^i$, we have that

$$\mathrm{d}^i\left(y_n^i, x_n^i\right)^2 = \left\| \left(\exp_{x_n^i}^i\right)^{-1}(y_n^i) \right\|_{x_n^i}^2 = \alpha_n^2 \frac{\left\|\hat{m}_n^i\right\|_{x_n^i}^2}{(\mathsf{h}_n^i)^2} = \frac{\alpha_n^2}{(1-\hat{\beta}^{n+1})^2} \frac{\left\|m_n^i\right\|_{x_n^i}^2}{(\mathsf{h}_n^i)^2}. \tag{34}$$

Accordingly, for all $i \in \mathcal{I}$ and all $x^i \in X^i$,

$$\begin{aligned}
&\mathrm{d}^i(y_n^i, x^i)^2 \\
&\leq \mathrm{d}^i(x_n^i, x^i)^2 + \frac{\zeta^i \alpha_n^2}{(1-\hat{\beta}^{n+1})^2} \frac{\left\|m_n^i\right\|_{x_n^i}^2}{(\mathsf{h}_n^i)^2} + \frac{2\alpha_n}{(1-\hat{\beta}^{n+1})\mathsf{h}_n^i} \left\langle m_n^i, \left(\exp_{x_n^i}^i\right)^{-1}(x^i) \right\rangle_{x_n^i}.
\end{aligned}$$

Meanwhile, from $Q_{\alpha^i}^i := P^i S_{\alpha^i}^i$ (see (9)) and $x_{n+1}^i = Q_{\alpha^i}^i(y_n^i)$, Proposition 2.3(i) ensures that

$$\mathrm{d}^i(x_{n+1}^i, x^i)^2 \leq \mathrm{d}^i(S_{\alpha^i}^i(y_n^i), x^i)^2,$$

which, together with (6) and (8), implies that

$$\mathrm{d}^i(x_{n+1}^i, x^i)^2 \leq \mathrm{d}^i(y_n^i, x^i)^2 - \alpha^i(1-\alpha^i)\mathrm{d}^i\left(T^i(y_n^i), y_n^i\right)^2.$$

Therefore, (33) holds.

The definitions of m_n and τ_n, together with the convexity of $\|\cdot\|_{x_n^i}^2$, guarantee that, for all $i \in \mathcal{I}$ and all $n \in \mathbb{N}$,

$$\begin{aligned}
\mathbb{E}\left[\left\|m_n^i\right\|_{x_n^i}^2\right] &\leq \beta_n \mathbb{E}\left[\left\|\varphi_{x_{n-1}^i \to x_n^i}^i(m_{n-1}^i)\right\|_{x_n^i}^2\right] + (1-\beta_n)\mathbb{E}\left[\left\|\mathsf{G}^i(x_n, \xi_n)\right\|_{x_n^i}^2\right] \\
&\leq \beta_n \mathbb{E}\left[\left\|m_{n-1}^i\right\|_{x_{n-1}^i}^2\right] + (1-\beta_n){B^i}^2.
\end{aligned}$$

Induction thus ensures that, for all $i \in \mathcal{I}$ and all $n \in \mathbb{N}$,

$$\mathbb{E}\left[\left\|m_n^i\right\|_{x_n^i}^2\right] \leq \tilde{B^i}^2 := \max\left\{\left\|\tau_{-1}\right\|_{x_0^i}^2, {B^i}^2\right\} < +\infty. \tag{35}$$

This completes the proof. □

Lemma 1 also leads to the following lemma, which is used to show the main theorems.

Lemma 2 *Suppose that Assumptions (A1)–(A5) and Conditions (C1)–(C3) hold and $X_n(x)$ is defined for all $x \in X$ and all $n \in \mathbb{N}$ by*

$$X_n(x) := \mathbb{E}\left[\sum_{i\in\mathcal{I}} \mathsf{h}_n^i \mathrm{d}^i(x_n^i, x^i)\right].$$

Then, for all $x \in X$ and all $n \in \mathbb{N}$,

$$X_{n+1}(x) \leq X_n(x) + \frac{2\alpha_n(1-\beta_n)}{1-\hat{\beta}^{n+1}}\mathbb{E}\left[\left\langle \exp_{x_n}^{-1}(x), \mathsf{g}(x_n)\right\rangle_{x_n}\right] + \mathbb{E}\left[\sum_{i\in\mathcal{I}} D^i\left(\mathsf{h}_{n+1}^i - \mathsf{h}_n^i\right)\right] + \frac{2\alpha_n\beta_n}{1-\hat{\beta}}\sum_{i\in\mathcal{I}}\tilde{B}^i D^i + \frac{\alpha_n^2}{(1-\hat{\beta})^2}\sum_{i\in\mathcal{I}}\frac{\zeta^i \tilde{B^i}^2}{\mathsf{h}_0^i},$$

where ζ^i and $\tilde{B}^i$ are defined as in Lemma 1.

Proof Condition (C1) and $x_n = x_n(\xi_{[n-1]})$ mean that, for all $i \in \mathcal{I}$, all $x^i \in X^i$, and all $n \in \mathbb{N}$,

$$\begin{aligned}
&\mathbb{E}\left[\left\langle \mathsf{G}^i(x_n, \xi_n), \left(\exp_{x_n^i}^i\right)^{-1}(x^i)\right\rangle_{x_n^i}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left\langle \mathsf{G}^i(x_n, \xi_n), \left(\exp_{x_n^i}^i\right)^{-1}(x^i)\right\rangle_{x_n^i} \middle| \xi_{[n-1]}\right]\right] \\
&= \mathbb{E}\left[\left\langle \mathbb{E}\left[\mathsf{G}^i(x_n, \xi_n)\middle|\xi_{[n-1]}\right], \left(\exp_{x_n^i}^i\right)^{-1}(x^i)\right\rangle_{x_n^i}\right] \\
&= \mathbb{E}\left[\left\langle \mathsf{g}^i(x_n), \left(\exp_{x_n^i}^i\right)^{-1}(x^i)\right\rangle_{x_n^i}\right].
\end{aligned}$$

Since Condition (C2) implies that, for all $x \in X$ and all $n \in \mathbb{N}$,

$$\left\langle \mathsf{g}(x_n), \exp_{x_n}^{-1}(x)\right\rangle_{x_n} = \sum_{i\in\mathcal{I}}\left\langle \mathsf{g}^i(x_n), \left(\exp_{x_n^i}^i\right)^{-1}(x^i)\right\rangle_{x_n^i},$$

we have that, for all $x \in X$ and all $n \in \mathbb{N}$,

$$\mathbb{E}\left[\left\langle \mathsf{g}(x_n), \exp_{x_n}^{-1}(x)\right\rangle_{x_n}\right] = \mathbb{E}\left[\sum_{i\in\mathcal{I}}\left\langle \mathsf{G}^i(x_n, \xi_n), \left(\exp_{x_n^i}^i\right)^{-1}(x^i)\right\rangle_{x_n^i}\right]. \quad (36)$$

The Cauchy-Schwarz inequality ensures that, for all $x \in X$ and all $n \in \mathbb{N}$,

$$\mathbb{E}\left[\sum_{i\in\mathcal{I}}\left\langle \tau_{n-1}^i, (\exp_{x_n^i}^i)^{-1}(x^i)\right\rangle_{x_n^i}\right] \leq \mathbb{E}\left[\sum_{i\in\mathcal{I}}\left\|\tau_{n-1}^i\right\|_{x_n^i}\left\|\left(\exp_{x_n^i}^i\right)^{-1}(x^i)\right\|_{x_n^i}\right],$$

which, together with (11) and Lemma 1, implies that

$$\mathbb{E}\left[\sum_{i\in\mathcal{I}}\left\langle \tau_{n-1}^i, (\exp_{x_n^i}^i)^{-1}(x^i)\right\rangle_{x_n^i}\right] \leq \sum_{i\in\mathcal{I}} \tilde{B}^i D^i. \tag{37}$$

Moreover, from Lemma 1, $\hat{\beta}\in[0,1)$, and $1/\mathsf{h}_n^i \leq 1/\mathsf{h}_0^i$ (by Assumption (A4)),

$$\mathbb{E}\left[\sum_{i\in\mathcal{I}} \frac{\zeta^i \alpha_n^2}{(1-\hat{\beta}^{n+1})^2} \frac{\left\|m_n^i\right\|_{x_n^i}^2}{\mathsf{h}_n^i}\right] \leq \frac{\alpha_n^2}{(1-\hat{\beta})^2} \sum_{i\in\mathcal{I}} \frac{\zeta^i \tilde{B^i}^2}{\mathsf{h}_0^i}. \tag{38}$$

Accordingly, Lemma 1, together with (36), (37), and (38), leads to the assertion in Lemma 2. □

The following is a convergence analysis of Algorithm 1.

Theorem 5 *Suppose that Assumptions (A1)–(A5) and Conditions (C1)–(C3) hold. Then, Algorithm 1 satisfies that, for all $i\in\mathcal{I}$ and all $n\geq 1$,*

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^n \mathrm{d}^i\left(T^i(y_k^i), y_k^i\right)^2\right] \leq \frac{1}{\hat{\alpha}^i}\left\{\frac{D^i}{n} + \frac{2\tilde{B}^i D^i}{\hat{\mathsf{h}}_0^i}\frac{1}{n}\sum_{k=1}^n \alpha_k + \frac{\zeta^i \tilde{B^i}^2}{(\hat{\mathsf{h}}_0^i)^2}\frac{1}{n}\sum_{k=1}^n \alpha_k^2\right\},\\
&\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^n \mathrm{d}^i\left(y_k^i, x_k^i\right)^2\right] \leq \frac{\tilde{B^i}^2}{(1-\hat{\beta})^2(\hat{\mathsf{h}}_0^i)^2}\frac{1}{n}\sum_{k=1}^n \alpha_k^2,
\end{aligned} \tag{39}$$

where $\hat{\alpha}^i := \alpha^i(1-\alpha^i)$ and $\hat{\mathsf{h}}_0^i := (1-\hat{\beta})\mathsf{h}_0^i$. Moreover, if $(\alpha_n(1-\beta_n))_{n\in\mathbb{N}}$ and $(\beta_n)_{n\in\mathbb{N}}$ are monotone decreasing, then, for all $n\geq 1$,

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^n \left\langle \exp_{x_k}^{-1}(x), \mathsf{g}(x_k)\right\rangle_{x_k}\right]\\
&\quad\leq \frac{\sum_{i\in\mathcal{I}} \hat{B}^i D^{i2}}{2(1-\beta_1)}\frac{1}{n\alpha_n} + \frac{\sum_{i\in\mathcal{I}} \zeta^i \tilde{B^i}^2 (\mathsf{h}_0^i)^{-1}}{2(1-\hat{\beta})(1-\beta_1)}\frac{1}{n}\sum_{k=1}^n \alpha_k + \frac{\sum_{i\in\mathcal{I}} \tilde{B}^i D^i}{1-\beta_1}\frac{1}{n}\sum_{k=1}^n \beta_k.
\end{aligned} \tag{40}$$

If (A1)' $T^i\colon M^i \to M^i$ $(i\in\mathcal{I})$ is nonexpansive with $\mathrm{Fix}(T^i)\neq\emptyset$, then, for all $n\geq 1$,

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^n \mathrm{d}^i\left(T^i(x_k^i), x_k^i\right)^2\right]\\
&\quad\leq \frac{2}{\hat{\alpha}^i}\frac{D^i}{n} + \frac{4\tilde{B}^i D^i}{\hat{\alpha}^i \hat{\mathsf{h}}_0^i}\frac{1}{n}\sum_{k=1}^n \alpha_k + \frac{2\tilde{B^i}^2}{(\hat{\mathsf{h}}_0^i)^2}\left\{\frac{\zeta^i}{\hat{\alpha}^i} + \frac{4}{(1-\hat{\beta})^2}\right\}\frac{1}{n}\sum_{k=1}^n \alpha_k^2.
\end{aligned} \tag{41}$$

Proof The Cauchy-Schwarz inequality, together with Lemma 1 and Assumption (A3) (see (11)), ensures that, for all $i \in \mathcal{I}$, all $x^i \in X^i$, and all $n \in \mathbb{N}$,

$$\mathbb{E}\left[\left\langle m_n^i, (\exp_{x_n^i}^i)^{-1}(x^i)\right\rangle_{x_n^i}\right] \leq \mathbb{E}\left[\left\|m_n^i\right\|_{x_n^i}\left\|(\exp_{x_n^i}^i)^{-1}(x^i)\right\|_{x_n^i}\right] \leq \tilde{B}^i D^i.$$

Lemma 1, together with $\hat{\beta} \in [0,1)$ and Assumption (A4), guarantees that, for all $i \in \mathcal{I}$, all $x^i \in X^i$, and all $k \geq 1$,

$$\begin{aligned}&\alpha^i(1-\alpha^i)\mathbb{E}\left[\mathrm{d}^i\left(T^i(y_k^i), y_k^i\right)^2\right]\\ &\quad\leq \mathbb{E}\left[\mathrm{d}^i(x_k^i, x^i)^2\right] - \mathbb{E}\left[\mathrm{d}^i(x_{k+1}^i, x^i)^2\right] + \frac{\zeta^i \tilde{B^i}^2}{(1-\hat{\beta})^2(\mathsf{h}_0^i)^2}\alpha_k^2 + \frac{2\tilde{B}^i D^i}{(1-\hat{\beta})\mathsf{h}_0^i}\alpha_k.\end{aligned} \tag{42}$$

Accordingly, we have that, for all $i \in \mathcal{I}$ and all $n \geq 1$,

$$\hat{\alpha}^i\mathbb{E}\left[\sum_{k=1}^n \mathrm{d}^i\left(T^i(y_k^i), y_k^i\right)^2\right] \leq D^i + \frac{\zeta^i \tilde{B^i}^2}{(1-\hat{\beta})^2(\mathsf{h}_0^i)^2}\sum_{k=1}^n \alpha_k^2 + \frac{2\tilde{B}^i D^i}{(1-\hat{\beta})\mathsf{h}_0^i}\sum_{k=1}^n \alpha_k,$$

where $\hat{\alpha}^i := \alpha^i(1-\alpha^i)$ and (11) implies that $\mathbb{E}\left[\mathrm{d}^i(x_1^i, x^i)^2\right] \leq D^i$. From (34), for all $i \in \mathcal{I}$ and all $n \geq 1$,

$$\mathbb{E}\left[\sum_{k=1}^n \mathrm{d}^i\left(y_k^i, x_k^i\right)^2\right] \leq \frac{\tilde{B^i}^2}{(1-\hat{\beta})^2(\mathsf{h}_0^i)^2}\sum_{k=1}^n \alpha_k^2,$$

which implies that (39) holds. Lemma 1, together with the definition of m_n^i, implies that, for all $i \in \mathcal{I}$, all $x^i \in X^i$, and all $n \in \mathbb{N}$,

$$\begin{aligned}&\left\langle -\mathsf{G}^i(x_n, \xi_n), \left(\exp_{x_n^i}^i\right)^{-1}(x^i)\right\rangle_{x_n^i}\\ &\leq \underbrace{\frac{(1-\hat{\beta}^{n+1})\mathsf{h}_n^i}{2\alpha_n(1-\beta_n)}\left\{\mathrm{d}^i(x_n^i, x^i)^2 - \mathrm{d}^i(x_{n+1}^i, x^i)^2\right\}}_{H_n^i(x^i)}\\ &\quad + \underbrace{\frac{\beta_n}{1-\beta_n}\left\langle \tau_{n-1}^i, \left(\exp_{x_n^i}^i\right)^{-1}(x^i)\right\rangle_{x_n^i}}_{B_n^i(x^i)} + \underbrace{\frac{\zeta^i\alpha_n}{2(1-\hat{\beta}^{n+1})(1-\beta_n)}\frac{\left\|m_n^i\right\|_{x_n^i}^2}{\mathsf{h}_n^i}}_{A_n^i(x^i)},\end{aligned}$$

which, together with (36), implies that, for all $x_\star \in X_\star$ and all $n \geq 1$,

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^n \left\langle \exp_{x_k}^{-1}(x), \mathsf{g}(x_k)\right\rangle_{x_k}\right]\\ &\leq \frac{1}{n}\mathbb{E}\left[\sum_{k=1}^n\sum_{i\in\mathcal{I}} H_k^i(x_\star^i)\right] + \frac{1}{n}\mathbb{E}\left[\sum_{k=1}^n\sum_{i\in\mathcal{I}} B_k^i(x_\star^i)\right] + \frac{1}{n}\mathbb{E}\left[\sum_{k=1}^n\sum_{i\in\mathcal{I}} A_k^i(x_\star^i)\right].\end{aligned} \tag{43}$$

The definition of $H_n^i(x^i)$ ($i \in \mathcal{I}, n \in \mathbb{N}$) and (11) guarantee that, for all $i \in \mathcal{I}$, all $x_\star \in X_\star$, and all $n \geq 1$,

$$\begin{aligned}
&\sum_{k=1}^{n} H_k^i(x_\star^i) \\
&\leq \frac{(1-\hat{\beta}^2)\mathsf{h}_1^i}{2\alpha_1(1-\beta_1)} {D^i}^2 + \sum_{k=2}^{n} \left\{ \frac{(1-\hat{\beta}^{k+1})\mathsf{h}_k^i}{2\alpha_k(1-\beta_k)} - \frac{(1-\hat{\beta}^k)\mathsf{h}_{k-1}^i}{2\alpha_{k-1}(1-\beta_{k-1})} \right\} \mathrm{d}^i(x_k^i, x_\star^i)^2.
\end{aligned}$$

Since $\hat{\beta} \in [0,1)$ and Assumption (A4) hold and $(\alpha_n(1-\beta_n))_{n\in\mathbb{N}}$ is monotone decreasing, we have that, for all $k \geq 2$,

$$\frac{(1-\hat{\beta}^{k+1})\mathsf{h}_k^i}{2\alpha_k(1-\beta_k)} - \frac{(1-\hat{\beta}^k)\mathsf{h}_{k-1}^i}{2\alpha_{k-1}(1-\beta_{k-1})} \geq 0.$$

Accordingly, for all $i \in \mathcal{I}$ and all $x_\star \in X_\star$,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{n} H_k^i(x_\star^i)\right] \\
&\leq \mathbb{E}\left[\frac{(1-\hat{\beta}^2)\mathsf{h}_1^i}{2\alpha_1(1-\beta_1)} {D^i}^2 + \sum_{k=2}^{n} \left\{ \frac{(1-\hat{\beta}^{k+1})\mathsf{h}_k^i}{2\alpha_k(1-\beta_k)} - \frac{(1-\hat{\beta}^k)\mathsf{h}_{k-1}^i}{2\alpha_{k-1}(1-\beta_{k-1})} \right\} {D^i}^2\right] \\
&= \mathbb{E}\left[\frac{(1-\hat{\beta}^{n+1})\mathsf{h}_n^i}{2\alpha_n(1-\beta_n)} {D^i}^2\right] \\
&\leq \frac{\hat{B}^i {D^i}^2}{2(1-\beta_1)\alpha_n},
\end{aligned} \tag{44}$$

where the second inequality comes from $\hat{\beta} \in [0,1)$, Assumption (A5), and $\beta_n \leq \beta_1$ ($n \geq 1$). The Cauchy-Schwarz inequality ensures that, for all $x_\star \in X_\star$ and all $n \geq 1$,

$$\mathbb{E}\left[\sum_{k=1}^{n}\sum_{i\in\mathcal{I}} B_k^i(x_\star^i)\right] \leq \mathbb{E}\left[\sum_{i\in\mathcal{I}}\sum_{k=1}^{n} \frac{\beta_k}{1-\beta_k} \left\|\tau_{k-1}^i\right\|_{x_k^i} \left\|(\exp_{x_k^i}^i)^{-1}(x_\star^i)\right\|_{x_k^i}\right],$$

which, together with (11), Lemma 1, and $\beta_n \leq \beta_1$ ($n \geq 1$), implies that

$$\mathbb{E}\left[\sum_{k=1}^{n}\sum_{i\in\mathcal{I}} B_k^i(x_\star^i)\right] \leq \frac{\sum_{i\in\mathcal{I}} \tilde{B}^i D^i}{1-\beta_1} \sum_{k=1}^{n} \beta_k. \tag{45}$$

Moreover, from Lemma 1, $\hat{\beta} \in [0,1)$, Assumption (A4), and $\beta_n \leq \beta_1$ ($n \geq 1$),

$$\begin{aligned}
\mathbb{E}\left[\sum_{k=1}^{n}\sum_{i\in\mathcal{I}} A_k^i(x_\star^i)\right] &= \mathbb{E}\left[\sum_{k=1}^{n}\sum_{i\in\mathcal{I}} \frac{\zeta^i \alpha_k}{2(1-\hat{\beta}^{k+1})(1-\beta_k)} \frac{\left\|m_k^i\right\|_{x_k^i}^2}{\mathsf{h}_k^i}\right] \\
&\leq \frac{1}{2(1-\hat{\beta})(1-\beta_1)} \sum_{i\in\mathcal{I}} \frac{\zeta^i {\tilde{B}^i}^2}{\mathsf{h}_0^i} \sum_{k=1}^{n} \alpha_k.
\end{aligned} \tag{46}$$

Hence, (43), (44), (45), and (46) lead to (40).

Suppose that Assumption (A1)' holds. Since the triangle inequality implies that, for all $i \in \mathcal{I}$ and all $n \in \mathbb{N}$,

$$\mathrm{d}^i\left(T^i(x_n^i), x_n^i\right) \leq \mathrm{d}^i\left(T^i(x_n^i), T^i(y_n^i)\right) + \mathrm{d}^i\left(T^i(y_n^i), y_n^i\right) + \mathrm{d}^i\left(y_n^i, x_n^i\right),$$

Assumption (A1)' ensures that

$$\mathrm{d}^i\left(T^i(x_n^i), x_n^i\right) \leq \mathrm{d}^i\left(T^i(y_n^i), y_n^i\right) + 2\mathrm{d}^i\left(y_n^i, x_n^i\right),$$

which implies that, for all $i \in \mathcal{I}$ and all $n \in \mathbb{N}$,

$$\mathrm{d}^i\left(T^i(x_n^i), x_n^i\right)^2 \leq 2\mathrm{d}^i\left(T^i(y_n^i), y_n^i\right)^2 + 8\mathrm{d}^i\left(y_n^i, x_n^i\right)^2. \tag{47}$$

(39) and (47) thus lead to (41), which completes the proof. □

Proof (Proof of Theorem 1) Let $i \in \mathcal{I}$ be fixed arbitrarily. From (34) and Lemma 1, together with Assumption (A4) and $\alpha_n := \alpha$ ($n \in \mathbb{N}$), (14) holds. If (15) does not hold, then there exists $\delta > 0$ such that

$$\alpha^i(1-\alpha^i)\liminf_{n\to+\infty}\mathbb{E}\left[\mathrm{d}^i\left(T^i(y_n^i), y_n^i\right)^2\right] > \frac{2\tilde{B}^i D^i}{(1-\hat{\beta})\mathsf{h}_0^i}\alpha + \frac{\zeta^i \tilde{B}^{i\,2}}{(1-\hat{\beta})^2(\mathsf{h}_0^i)^2}\alpha^2 + \delta.$$

The definition of the limit inferior of $(\mathbb{E}[\mathrm{d}^i(T^i(y_n^i), y_n^i)^2])_{n\in\mathbb{N}}$ ensures that there exists $n_0 \in \mathbb{N}$ such that, for all $n \geq n_0$,

$$\alpha^i(1-\alpha^i)\liminf_{n\to+\infty}\mathbb{E}\left[\mathrm{d}^i\left(T^i(y_n^i), y_n^i\right)^2\right] - \frac{1}{2}\delta \leq \alpha^i(1-\alpha^i)\mathbb{E}\left[\mathrm{d}^i\left(T^i(y_n^i), y_n^i\right)^2\right],$$

which implies that, for all $n \geq n_0$,

$$\alpha^i(1-\alpha^i)\mathbb{E}\left[\mathrm{d}^i\left(T^i(y_n^i), y_n^i\right)^2\right] > \frac{2\tilde{B}^i D^i}{(1-\hat{\beta})\mathsf{h}_0^i}\alpha + \frac{\zeta^i \tilde{B}^{i\,2}}{(1-\hat{\beta})^2(\mathsf{h}_0^i)^2}\alpha^2 + \frac{1}{2}\delta.$$

From (42) with $\alpha_n := \alpha$ and $\beta_n := \beta$ ($n \in \mathbb{N}$),

$$\begin{aligned}\mathbb{E}\left[\mathrm{d}^i(x_{n+1}^i, x^i)^2\right] \leq{}& \mathbb{E}\left[\mathrm{d}^i(x_n^i, x^i)^2\right] - \alpha^i(1-\alpha^i)\mathbb{E}\left[\mathrm{d}^i\left(T^i(y_n^i), y_n^i\right)^2\right] \\ &+ \frac{2\tilde{B}^i D^i}{(1-\hat{\beta})\mathsf{h}_0^i}\alpha + \frac{\zeta^i \tilde{B}^{i\,2}}{(1-\hat{\beta})^2(\mathsf{h}_0^i)^2}\alpha^2,\end{aligned}$$

which implies that, for all $n \geq n_0$,

$$\begin{aligned}\mathbb{E}\left[\mathrm{d}^i(x_{n+1}^i, x^i)^2\right] &< \mathbb{E}\left[\mathrm{d}^i(x_n^i, x^i)^2\right] - \left\{\frac{2\tilde{B}^i D^i}{(1-\hat{\beta})\mathsf{h}_0^i}\alpha + \frac{\zeta^i \tilde{B}^{i\,2}}{(1-\hat{\beta})^2(\mathsf{h}_0^i)^2}\alpha^2 + \frac{1}{2}\delta\right\} \\ &\quad + \frac{2\tilde{B}^i D^i}{(1-\hat{\beta})\mathsf{h}_0^i}\alpha + \frac{\zeta^i \tilde{B}^{i\,2}}{(1-\hat{\beta})^2(\mathsf{h}_0^i)^2}\alpha^2 \\ &= \mathbb{E}\left[\mathrm{d}^i(x_n^i, x^i)^2\right] - \frac{1}{2}\delta \\ &< \mathbb{E}\left[\mathrm{d}^i(x_{n_0}^i, x^i)^2\right] - \frac{1}{2}\delta(n+1-n_0).\end{aligned}$$

Since the right-hand side of the above inequality approaches minus infinity when n diverges, we have a contradiction. Hence, (15) holds.

Assumptions (A4) and (A5) and the conditions, $\lim_{n\to+\infty}\hat{\beta}^{n+1}=0$ and $X_n^\star := X_n(x_\star) \leq \sum_{i\in\mathcal{I}}\hat{B}^i D^i < +\infty$ ($x_\star \in X_\star$) (by Assumptions (A3) and (A5)), guarantee that, for all $\epsilon>0$, there exists $n_1\in\mathbb{N}$ such that, for all $n\in\mathbb{N}$, $n\geq n_1$ implies that

$$\mathbb{E}\left[\sum_{i\in\mathcal{I}} D^i\left(\mathsf{h}_{n+1}^i-\mathsf{h}_n^i\right)\right]+\hat{\beta}^{n+1}\left(X_{n+1}^\star-X_n^\star\right)\leq\alpha(1-\beta)\epsilon. \tag{48}$$

Let us show that, for all $\epsilon>0$,

$$\liminf_{n\to+\infty}\mathbb{E}\left[f(x_n)-f_\star\right]\leq\frac{\sum_{i\in\mathcal{I}}\zeta^i\tilde{B^i}^2(\mathsf{h}_0^i)^{-1}}{2(1-\beta)(1-\hat{\beta})}\alpha+\frac{\sum_{i\in\mathcal{I}}\tilde{B}^i D^i}{(1-\beta)(1-\hat{\beta})}\beta+\frac{3}{2}\epsilon. \tag{49}$$

If (49) does not hold, then there exists $\epsilon_0>0$ such that

$$\liminf_{n\to+\infty}\mathbb{E}\left[f(x_n)-f_\star\right]>\frac{\sum_{i\in\mathcal{I}}\zeta^i\tilde{B^i}^2(\mathsf{h}_0^i)^{-1}}{2(1-\beta)(1-\hat{\beta})}\alpha+\frac{\sum_{i\in\mathcal{I}}\tilde{B}^i D^i}{(1-\beta)(1-\hat{\beta})}\beta+\frac{3}{2}\epsilon_0.$$

From the definition of the limit inferior of $(\mathbb{E}[f(x_n)-f_\star])_{n\in\mathbb{N}}$, there exists $n_2\in\mathbb{N}$ such that, for all $n\geq n_2$,

$$\liminf_{n\to+\infty}\mathbb{E}\left[f(x_n)-f_\star\right]-\frac{1}{2}\epsilon_0\leq\mathbb{E}\left[f(x_n)-f_\star\right].$$

Hence, we have that, for all $n\geq n_2$,

$$\mathbb{E}\left[f(x_n)-f_\star\right]>\frac{\sum_{i\in\mathcal{I}}\zeta^i\tilde{B^i}^2(\mathsf{h}_0^i)^{-1}}{2(1-\beta)(1-\hat{\beta})}\alpha+\frac{\sum_{i\in\mathcal{I}}\tilde{B}^i D^i}{(1-\beta)(1-\hat{\beta})}\beta+\epsilon_0.$$

The convexity of f implies that, for all $n\in\mathbb{N}$,

$$\mathbb{E}\left[\left\langle\exp_{x_n}^{-1}(x_\star),\mathsf{g}(x_n)\right\rangle_{x_n}\right]\leq f_\star-f(x_n). \tag{50}$$

Since Lemma 2, together with $\alpha_n:=\alpha$, $\beta_n:=\beta$ ($n\in\mathbb{N}$), (48), and (50), ensures that, for all $n\geq n_1$,

$$\begin{aligned}X_{n+1}^\star\leq X_n^\star+\alpha(1-\beta)\epsilon-2\alpha(1-\beta)\mathbb{E}\left[f(x_n)-f_\star\right]+\frac{2\alpha\beta}{1-\hat{\beta}}\sum_{i\in\mathcal{I}}\tilde{B}^i D^i\\+\frac{\alpha^2}{1-\hat{\beta}}\sum_{i\in\mathcal{I}}\frac{\zeta^i\tilde{B^i}^2}{\mathsf{h}_0^i},\end{aligned}$$

we find that, for all $n \geq n_3 := \max\{n_1, n_2\}$,

$$\begin{aligned} X^\star_{n+1} &< X^\star_n + \alpha(1-\beta)\epsilon_0 + \frac{2\alpha\beta}{1-\hat{\beta}} \sum_{i\in\mathcal{I}} \tilde{B}^i D^i + \frac{\alpha^2}{1-\hat{\beta}} \sum_{i\in\mathcal{I}} \frac{\zeta^i {\tilde{B}^i}^2}{\mathsf{h}^i_0} \\ &\quad - 2\alpha(1-\beta) \left\{ \frac{\sum_{i\in\mathcal{I}} \zeta^i {\tilde{B}^i}^2 (\mathsf{h}^i_0)^{-1}}{2(1-\beta)(1-\hat{\beta})}\alpha + \frac{\sum_{i\in\mathcal{I}} \tilde{B}^i D^i}{(1-\beta)(1-\hat{\beta})}\beta + \epsilon_0 \right\} \\ &= X^\star_n - \alpha(1-\beta)\epsilon_0 \\ &< X^\star_{n_3} - \alpha(1-\beta)\epsilon_0 \left(n+1-n_3\right), \end{aligned}$$

which is a contradiction. Hence, (49) holds for all $\epsilon > 0$. This implies that (16) holds. Obviously, (17) holds from (39) with $\alpha_n := \alpha$ and $\beta_n := \beta$ $(n \in \mathbb{N})$.

The conditions $\alpha_n := \alpha$ and $\beta_n := \beta$ $(n \in \mathbb{N})$ satisfy that $(\alpha_n(1-\beta_n))_{n\in\mathbb{N}}$ and $(\beta_n)_{n\in\mathbb{N}}$ are monotone decreasing. Since f is convex, induction shows that

$$f(\bar{x}_n) \leq \frac{1}{n} \sum_{k=1}^{n} f(x_k). \tag{51}$$

Therefore, (40) in Theorem 5 leads to (19). If Assumption (A1)' holds, then Theorem 5 leads to (20). □

Proof (Proof of Theorem 2) From (34) and Lemma 1, together with Assumption (A4) and $\lim_{n\to+\infty} \alpha_n = 0$ (by $\sum_{n=0}^{+\infty} \alpha_n^2 < +\infty$), we have that $\lim_{n\to+\infty} \mathbb{E}[\mathrm{d}^i(y^i_n, x^i_n)^2] = 0$. Define Y^i_n for all $x \in X$, all $i \in \mathcal{I}$, and all $n \in \mathbb{N}$ by

$$Y^i_n(x) := \alpha_n \mathbb{E}\left[\mathrm{d}^i(x^i_n, x^i)^2\right].$$

Inequality (42) then ensures that, for all $x \in X$, all $i \in \mathcal{I}$, and all $k \in \mathbb{N}$,

$$\begin{aligned} &\alpha^i(1-\alpha^i)\alpha_k \mathbb{E}\left[\mathrm{d}^i\left(T^i(y^i_k), y^i_k\right)^2\right] \\ &\quad \leq Y^i_k(x) - \alpha_k \mathbb{E}\left[\mathrm{d}^i(x^i_{k+1}, x^i)^2\right] + \frac{\zeta^i {\tilde{B}^i}^2}{(1-\hat{\beta})^2(\mathsf{h}^i_0)^2}\alpha_k^3 + \frac{2\tilde{B}^i D^i}{(1-\hat{\beta})\mathsf{h}^i_0}\alpha_k^2, \end{aligned}$$

which, together with $\alpha_{n+1} \leq \alpha_n$ $(n \in \mathbb{N})$, implies that

$$\begin{aligned} &\alpha^i(1-\alpha^i)\alpha_k \mathbb{E}\left[\mathrm{d}^i\left(T^i(y^i_k), y^i_k\right)^2\right] \\ &\quad \leq Y^i_k(x) - Y^i_{k+1}(x) + \frac{\zeta^i {\tilde{B}^i}^2}{(1-\hat{\beta})^2(\mathsf{h}^i_0)^2}\alpha_k^3 + \frac{2\tilde{B}^i D^i}{(1-\hat{\beta})\mathsf{h}^i_0}\alpha_k^2. \end{aligned}$$

Summing the above inequality from $k = 0$ to $k = n$ means that, for all $x \in X$, all $i \in \mathcal{I}$, and all $n \in \mathbb{N}$,

$$\hat{\alpha}^i \sum_{k=0}^{n} \alpha_k \mathbb{E}\left[\mathrm{d}^i\left(T^i(y^i_k), y^i_k\right)^2\right] \leq Y^i_0(x) + \frac{\zeta^i {\tilde{B}^i}^2}{(1-\hat{\beta})^2(\mathsf{h}^i_0)^2} \sum_{k=0}^{n} \alpha_k^3 + \frac{2\tilde{B}^i D^i}{(1-\hat{\beta})\mathsf{h}^i_0} \sum_{k=0}^{n} \alpha_k^2,$$

where $\hat{\alpha}^i := \alpha^i(1-\alpha^i)$. Since $(\alpha_n)_{n\in\mathbb{N}} \subset (0,1)$ satisfies $\sum_{n=0}^{+\infty} \alpha_n^2 < +\infty$, we have that, for all $x \in X$ and all $i \in \mathcal{I}$,

$$\sum_{n=0}^{+\infty} \alpha_n \mathbb{E}\left[\mathrm{d}^i\left(T^i(y_n^i), y_n^i\right)^2\right] < +\infty. \tag{52}$$

We prove that, for all $i \in \mathcal{I}$,

$$\liminf_{n\to+\infty} \mathbb{E}\left[\mathrm{d}^i\left(T^i(y_n^i), y_n^i\right)^2\right] \leq 0. \tag{53}$$

Assume that (53) does not hold. Then, there exist $i \in \mathcal{I}$, $\gamma > 0$, and $m_0 \in \mathbb{N}$ such that, for all $n \geq m_0$,

$$\mathbb{E}\left[\mathrm{d}^i\left(T^i(y_n^i), y_n^i\right)^2\right] \geq \gamma,$$

which, together with $\sum_{n=0}^{+\infty} \alpha_n = +\infty$ and (52), implies that

$$+\infty = \gamma \sum_{n=m_0}^{+\infty} \alpha_n \leq \sum_{n=m_0}^{+\infty} \alpha_n \mathbb{E}\left[\mathrm{d}^i\left(T^i(y_n^i), y_n^i\right)^2\right] < +\infty.$$

This is a contradiction. Hence, we have (53), which implies that, for all $i \in \mathcal{I}$, $\liminf_{n\to+\infty} \mathbb{E}[\mathrm{d}^i(T^i(y_n^i), y_n^i)^2] = 0$. Lemma 2 with (50) guarantees that, for all $k \in \mathbb{N}$,

$$\begin{aligned}\frac{2\alpha_k}{1-\hat{\beta}^{k+1}} \mathbb{E}\left[f(x_k) - f_\star\right] \leq X_k^\star - X_{k+1}^\star + \mathbb{E}\left[\sum_{i\in\mathcal{I}} D^i\left(\mathsf{h}_{k+1}^i - \mathsf{h}_k^i\right)\right] \\ + \frac{2\alpha_k\beta_k}{1-\hat{\beta}}\left(\sum_{i\in\mathcal{I}} \tilde{B}^i D^i + F\right) + \frac{\alpha_k^2}{(1-\hat{\beta})^2} \sum_{i\in\mathcal{I}} \frac{\zeta^i \tilde{B^i}^2}{\mathsf{h}_0^i},\end{aligned}$$

where $X_n^\star := X_n(x^\star)$ and $F := \sup\{|\mathbb{E}[f(x_n) - f_\star]| \colon n \in \mathbb{N}\}$ is finite from Assumptions (A2) and (A3). Summing the above inequality from $k = 0$ to $k = n$ implies that

$$\begin{aligned}2\sum_{k=0}^{n} \frac{\alpha_k}{1-\hat{\beta}^{k+1}} \mathbb{E}\left[f(x_k) - f_\star\right] \leq X_0^\star + \sum_{i\in\mathcal{I}} D^i B^i + \frac{1}{(1-\hat{\beta})^2} \sum_{i\in\mathcal{I}} \frac{\zeta^i \tilde{B^i}^2}{\mathsf{h}_0^i} \sum_{k=0}^{n} \alpha_k^2 \\ + \frac{2}{1-\hat{\beta}}\left(\sum_{i\in\mathcal{I}} \tilde{B}^i D^i + F\right) \sum_{k=0}^{n} \alpha_k\beta_k,\end{aligned}$$

where $\sup\{\mathbb{E}[\mathsf{h}_n^i] \colon n \in \mathbb{N}\} \leq \hat{B}^i$ holds from Assumption (A5). From $\sum_{n=0}^{+\infty} \alpha_n\beta_n < +\infty$ and $\sum_{n=0}^{+\infty} \alpha_n^2 < +\infty$, we have that

$$\sum_{k=0}^{+\infty} \frac{\alpha_k}{1-\hat{\beta}^{k+1}} \mathbb{E}\left[f(x_k) - f_\star\right] < +\infty.$$

We also have that $\sum_{n=0}^{+\infty} \alpha_k/(1-\hat{\beta}^{n+1}) \geq \sum_{n=0}^{+\infty} \alpha_k = +\infty$. Accordingly, a discussion similar to the one for proving (53) leads to the finding that

$$\liminf_{n\to+\infty} \mathbb{E}\left[f(x_k) - f_\star\right] \leq 0.$$

Theorem 5, together with (51) and (23), leads to (24) and (25) with rate of convergence (39), (40), and (41). This completes the proof. □

Proof (Proofs of Theorems 3 and 4) A discussion similar to the one for showing Theorem 1 (resp. Theorem 2) with $\mathtt{g} := \mathrm{grad} f$ leads to Theorem 3 (resp. Theorem 4). □

Proof of Corollary 1

Proof The step-sizes $\alpha_n := 1/n^\eta$ ($\eta \in (1/2, 1], n \geq 1$) and $(\beta_n)_{n\in\mathbb{N}}$ such that $\sum_{n=1}^{+\infty} \alpha_n\beta_n < +\infty$ satisfy (21). Hence, Theorem 2 leads to (26).

Let $\alpha_n := 1/n^\eta$ ($\eta \in [1/2, 1), n \geq 1$) and $(\beta_n)_{n\in\mathbb{N}}$ such that $\sum_{n=1}^{+\infty} \beta_n < +\infty$. We have that

$$\lim_{n\to+\infty} \frac{1}{n\alpha_n} = \lim_{n\to+\infty} \frac{1}{n^{1-\eta}} = 0. \tag{54}$$

Moreover,

$$\begin{aligned} \frac{1}{n}\sum_{k=1}^{n} \alpha_k^2 \leq \frac{1}{n}\sum_{k=1}^{n} \alpha_k &\leq \frac{1}{n}\left\{1 + \int_1^n \frac{\mathrm{d}t}{t^\eta}\right\} = \frac{1}{n}\left\{\frac{n^{1-\eta}}{1-\eta} - \frac{\eta}{1-\eta}\right\} \\ &\leq \frac{1}{1-\eta}\frac{1}{n^\eta}. \end{aligned} \tag{55}$$

Hence, $\lim_{n\to+\infty}(1/n)\sum_{k=1}^{n} \alpha_k = \lim_{n\to+\infty}(1/n)\sum_{k=1}^{n} \alpha_k^2 = 0$. From $\sum_{n=1}^{+\infty} \beta_n < +\infty$, $\lim_{n\to+\infty}(1/n)\sum_{k=1}^{n} \beta_k = 0$. Hence, $\alpha_n := 1/n^\eta$ and $(\beta_n)_{n\in\mathbb{N}}$ such that $\sum_{n=1}^{+\infty} \beta_n < +\infty$ satisfy (23). Accordingly, from Theorem 2 with (51), (54), and (55), we have the assertions in Corollary 1. □