Optimization without Retraction on the Random Generalized Stiefel Manifold

Simon Vary [1 2] **Pierre Ablin** [3] **Bin Gao** [4] **P.-A. Absil** [1]

Abstract

Optimization over the set of matrices X that satisfy $X^\top BX = I_p$, referred to as the generalized Stiefel manifold, appears in many applications involving sampled covariance matrices such as the canonical correlation analysis (CCA), independent component analysis (ICA), and the generalized eigenvalue problem (GEVP). Solving these problems is typically done by iterative methods that require a fully formed B. We propose a cheap stochastic iterative method that solves the optimization problem while having access only to a random estimates of B. Our method does not enforce the constraint in every iteration; instead, it produces iterations that converge to critical points on the generalized Stiefel manifold defined in expectation. The method has lower per-iteration cost, requires only matrix multiplications, and has the same convergence rates as its Riemannian optimization counterparts that require the full matrix B. Experiments demonstrate its effectiveness in various machine learning applications involving generalized orthogonality constraints, including CCA, ICA, and the GEVP.

1. Introduction

Many problems in machine learning and engineering, including canonical correlation analysis (CCA) (Hotelling, 1936), independent component analysis (ICA) (Comon, 1994), linear discriminant analysis (McLachlan, 1992), and the generalized eigenvalue problem (GEVP) (Saad, 2011), can be formulated as the following optimization problem:

$$\begin{aligned} &\min_{X \in \mathrm{St}_B(p,n)} f(X) := \mathbb{E}[f_\xi(X)], \quad \text{s.t.} \quad B = \mathbb{E}[B_\zeta], \\ &\mathrm{St}_B(p,n) := \left\{ X \in \mathbb{R}^{n\times p} \mid X^\top BX = I_p \right\} \end{aligned} \tag{1}$$

where the objective function f is the expectation of L-smooth functions f_ξ, $B \in \mathbb{R}^{n\times n}$ is a positive-definite matrix, and ξ, ζ are independent random variables. The individual random matrices B_ζ are only assumed to be positive semidefinite. The feasible set $\mathrm{St}_B(p,n) \subset \mathbb{R}^{n\times p}$ defines a smooth manifold referred to as the *generalized Stiefel manifold*.

In the deterministic case, when we have access to the matrix B, the optimization problem can be solved by Riemannian techniques (Absil et al., 2008; Boumal, 2023). Riemannian methods produce a sequence of iterates belonging to the set $\mathrm{St}_B(p,n)$, often by repeatedly applying a *retraction* that maps tangent vectors to points on the manifold. In the case of $\mathrm{St}_B(p,n)$, retractions require non-trivial linear algebra operations such as eigenvalue or Cholesky decomposition. On the other hand, optimization on $\mathrm{St}_B(p,n)$ also lends itself to infeasible optimization methods, such as the augmented Lagrangian method. Such methods are typically employed in deterministic setting when the feasible set does not have a convenient projection, e.g., it lacks a closed-form expression or it requires solving an expensive optimization problem (Bertsekas, 1982). Infeasible approaches produce iterates that do not belong to the feasible set but converge to it by solving a sequence of unconstrained optimization problems. However, solving the optimization subproblems in each iteration might be computationally expensive and the methods are sensitive to the choice of hyper-parameters, both in theory and in practice.

In this paper, unlike in the aforementioned areas of study, we consider the setting (1) where the feasible set itself is *stochastic*, i.e., the matrix B is unknown and is an expectation of random estimates B_ζ, for which neither Riemannian methods nor infeasible optimization techniques are well-suited. In particular, we are interested in the case where we only have access to i.i.d. samples from ξ and ζ, and not to the full function f and matrix B.

We design an iterative *landing* method requiring only random estimates B_ζ that provably converges to critical points of (1) while performing only matrix multiplications. The

[1]ICTEAM Institute, UCLouvain, Louvain-la-neuve, Belgium [2]Department of Statistics, University of Oxford, Oxford, United Kingdom [3]Apple Machine Learning Group, Paris, France [4]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China. Correspondence to: Simon Vary <simon.vary@stats.ox.ac.uk>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

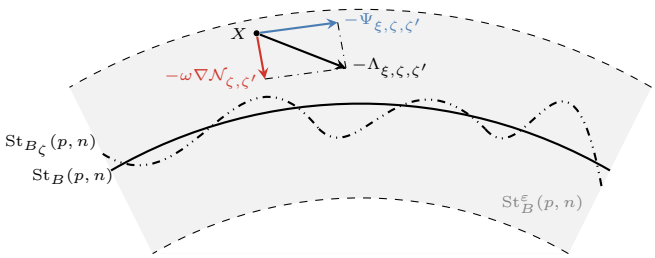


Figure 1. Illustration of the landing field and the random feasible set.

main principle of the method is depicted in the diagram in Figure 1. It is inspired by a recent line of work that first considered the orthogonal group $\mathrm{St}_{I_n}(n,n)$ (Ablin & Peyré, 2022) and was later extended to the Stiefel manifold $\mathrm{St}_{I_n}(p,n)$ (Gao et al., 2022b; Ablin et al., 2023; Schechtman et al., 2023). Instead of performing retractions after each iteration, the proposed algorithm performs an update along the sum of two orthogonal vectors—one is an unbiased estimator of a *relative ascent direction* (a concept defined in Section 2) and the other is an unbiased estimator of a direction towards $\mathrm{St}_B(p,n)$. The algorithm does not enforce the constraint in every iteration; instead it produces iterations that remain within an initially prescribed ε-safe region, and finally "lands" on, i.e., converges to, the manifold.

Specifically, the proposed stochastic landing iteration for solving (1) is the simple, cheap, and stochastic update rule

$$\begin{aligned} X^{k+1} &= X^k - \eta_k \Lambda_{\xi^k,\zeta^k,\zeta'^k}(X^k) \\ \text{with} \quad \Lambda_{\xi,\zeta,\zeta'}(X) &= \Psi_{\xi,\zeta,\zeta'}(X) + \omega\nabla\mathcal{N}_{\zeta,\zeta'}(X), \end{aligned} \tag{2}$$

in $\mathbb{R}^{n\times p}$ whose two components are

$$\begin{aligned} \Psi_{\xi,\zeta,\zeta'}(X) &= 2\,\mathrm{skew}\left(\nabla f_\xi(X)X^\top B_\zeta\right)B_{\zeta'}X, \\ \nabla\mathcal{N}_{\zeta,\zeta'}(X) &= 2B_{\zeta'}X\left(X^\top B_\zeta X - I_p\right), \end{aligned}$$

where $\omega > 0$, $\nabla\mathcal{N}_{\zeta,\zeta'}(X)$ is an unbiased stochastic estimator of the gradient of $\mathcal{N}(X) = \frac{1}{2}\|X^\top BX - I_p\|_{\mathrm{F}}^2$, and $\mathrm{skew}(A) = (A - A^\top)/2$. The above landing field formula (2) applies in the general case when both the function f and the constraint matrix B are stochastic; the deterministic case is recovered by substituting $\nabla f_\xi = \nabla f$ and $B_\zeta = B_{\zeta'} = B$. Note that in many applications of interest, $B_\zeta = \sum_{i=1}^r x_i x_i^\top/r$ is a subsampled covariance matrix with batch-size r, that is of rank at most r when $r \le n$. Unlike retractions, the landing method benefits in this setting since the cost of multiplication by B_ζ, which is the dominant cost of (2), becomes $\mathcal{O}(npr)$ instead of $\mathcal{O}(n^2p)$ where r is the batch size. The landing method never requires to form the matrix B, thus having space complexity defined by only saving the iterates and the samples: $\mathcal{O}\left(n(p+r)\right)$ instead of $\mathcal{O}(n^2)$.

We prove that the landing iteration converges to e-critical points, i.e., points X such that $\|\mathrm{grad} f(X)\| \le e$ (where $\mathrm{grad} f$ denotes the Riemannian gradient defined in (10)) and $\|\mathcal{N}(X)\| \le e$, with a fixed step-size in the deterministic case (Theorem 2.8) and with a decaying step-size in the stochastic case (Theorem 2.9), with a rate that matches that of deterministic (Boumal et al., 2019) and stochastic (Bonnabel, 2013) Riemannian gradient descent on $\mathrm{St}_B(p,n)$. The advantages of the landing field in (2) are that i) its computation involves only parallelizable matrix multiplications, which is cheaper than the computations of the Riemannian gradient and retraction and ii) it handles gracefully the stochastic constraint, while Riemannian approaches need form the full estimate of B.

While the presented theory holds for a general smooth, possibly non-convex objective f, a particular problem that can be either phrased as (1) or framed as an optimization over the product manifold of two $\mathrm{St}_B(p,n)$ is CCA, which is a widely used technique for measuring similarity between datasets (Raghu et al., 2017). CCA aims to find the top-p most correlated principal components $X, Y \in \mathbb{R}^{n\times p}$, for two zero-centered datasets $D_1 = (d_1^1,\dots,d_1^N), D_2 = (d_2^1,\dots,d_2^N) \in \mathbb{R}^{n\times N}$ of N i.i.d. samples from two different distributions and is formulated as

$$\begin{aligned} &\min_{X,Y\in\mathbb{R}^{n\times p}} \mathbb{E}_i\left[-\mathrm{Tr}(X^\top d_1^i (d_2^i)^\top Y)\right] \\ &X^\top\mathbb{E}_i[d_1^i(d_1^i)^\top]X = I_p \text{ and } Y^\top\mathbb{E}_i[d_2^i(d_2^i)^\top]Y = I_p, \end{aligned} \tag{3}$$

where the expectations are w.r.t. the uniform distribution over $\{1,\dots,N\}$. Here, the constraint matrices B_ζ correspond to individual or mini-batch sample covariances, and the constraints are that the two matrices $X, Y \in \mathbb{R}^{n\times p}$ are in the generalized Stiefel manifold. The proposed landing method is able to solve (3) while only having a stochastic estimate of the covariance matrices.

The rest of the introduction gives a brief overview of the current optimization techniques for solving (1) and its forthcoming generalization (4) when the feasible set is deterministic, since we are not aware of existing techniques for (1) with stochastic feasible set. Afterwards, the paper is organized as follows:

- In Section 2 we give a form to a generalized landing algorithm for solving a smooth optimization problem $\min_{x\in\mathcal{M}} f(x)$ on a smooth manifold $\mathcal{M}$ defined below in (4). Under suitable conditions, the algorithm converges to critical points with the same sublinear rate, $\mathcal{O}(1/K)$, as its Riemannian counterpart (Boumal et al., 2019), see Theorem 2.8. Unlike in Schechtman et al. (2023), our analysis is based on a smooth merit function allowing us to obtain a convergence result for the stochastic variant of the algorithm, when having an unbiased estimator for the landing field, see Theorem 2.9.

- In Section 3 we build on the general theory developed in the previous section and prove that the update rule for the generalized Stiefel manifold in (2) converges to critical points of (1), both in the deterministic case with the rate $\mathcal{O}(1/K)$, and in expectation with the rate $\mathcal{O}(1/\sqrt{K})$ in the case when both the gradient of the objective function *and* the feasible set are stochastic estimates.

- In Section 4 we numerically demonstrate the efficiency of the proposed method on a deterministic example of solving a generalized eigenvalue problem, stochastic CCA and ICA.

Notation and terminology. We denote vectors by lowercase letters $x, y, z, \ldots$, matrices with uppercase letters $X, Y, Z, \ldots$, and I_n denotes the $n \times n$ identity matrix. We let β_i denote the ith eigenvalue of B and $\kappa_B = \beta_1/\beta_n$ the condition number of B. Let $\mathrm{D}f(x)[v] = \lim_{t\to 0}(f(x + tv) - f(x))/t$ denote the derivative of f at x along v. We let $\|\cdot\|$ denote the ℓ_2-norm also termed Frobenius norm for matrices, whereas $\|\cdot\|_2$ denotes the operator norm induced by ℓ_2-norm. We denote the Frobenius inner product as $\langle\cdot\,,\,\cdot\rangle$, with respect to which we define the adjoint of a linear operator $\mathrm{A}[v]$ denoted by $\mathrm{A}^*[w]$. We say that a function $f : \mathbb{R}^n \to \mathbb{R}$ is L_f-smooth if it is continuously differentiable and its gradient is Lipschitz continous with Lipschitz constant L_f, i.e., $\|\nabla f(x) - \nabla f(y)\|_2 \leq L_f\|x - y\|_2$, for all $x, y \in \mathbb{R}^n$.

1.1. Prior work related to optimization on the generalized Stiefel manifold

Riemannian optimization. A widely used approach to solving optimization problems constrained to manifolds as in (4) are the techniques from Riemannian optimization. These methods are based on the observation that smooth sets can be locally approximated by a linear subspace, which allows to extend classical Euclidean optimization methods, such as gradient descent and the stochastic gradient descent, to the Riemannian setting. For example, Riemannian gradient descent iterates $x^{k+1} = \mathrm{Retr}_{\mathcal{M}}(x^k, -\eta_k \mathrm{grad} f(x^k))$, where $\eta_k > 0$ is the step-size at iteration k, $\mathrm{grad} f(x^k)$ is the Riemannian gradient that is computed as a projection of $\nabla f(x^k)$ on the tangent space of $\mathcal{M}$ at x^k, and Retr is the *retraction* operation, which maps the updated iterate along the direction $-\eta_k \mathrm{grad} f(x^k)$ onto the manifold and is accurate up to the first-order, i.e., $\mathrm{Retr}_{\mathcal{M}}(x, d) = x + d + o(\|d\|)$. Retractions allow the implementation of Riemannian counterparts to classical Euclidean methods on the generalized Stiefel manifold, such as Riemannian (stochastic) gradient descent (Bonnabel, 2013; Zhang & Sra, 2016), trust-region methods (Absil et al., 2007), and accelerated methods (Ahn & Sra, 2020); for an overview, see Absil et al. (2008); Boumal (2023).

There are several ways to compute a retraction to the generalized Stiefel manifold, which we summarize in Table 1 and we give a more detailed explanation in Appendix A. Overall, we see that the landing field (2) is much cheaper to compute than all these retractions in two cases: i) when $n \simeq p$, then the bottleneck in the retractions becomes the matrix factorizations, which, although they are of the same complexity as matrix multiplications, are much more expensive and hard to parallelize, ii) when $n \gg p$, the dominant cost of all retractions lies in matrix multiplications that require in practice $\mathcal{O}(n^2p)$, whereas the use of the batches of size r mentioned above allows computing the landing field in $\mathcal{O}(npr)$. We demonstrate numerically the practical cost of computing retractions in Figure 7b in the appendices.

Infeasible optimization methods. A popular approach for solving constrained optimization is to employ the squared ℓ_2-penalty method by adding the $\omega\mathcal{N}(X)$ regularizer to the objective. However, unlike the landing method, the iterates of the squared penalty method do not converge to the feasible set for any fixed choice of ω and converge only when ω goes to ∞ (Nocedal & Wright, 2006). In contrast, the landing method provably converges to the feasible set for any fixed $\omega > 0$, which is enabled by the structure of the landing field (2) as the sum of two orthogonal components, the second one being the gradient of the infeasibility measure $\mathcal{N}$.

Augmented Lagrangian methods seek to solve a deterministic minimization problem with an augmented Lagrangian function $\mathcal{L}(x, \lambda)$, such as the one introduced later in (6), by updating the solution vector x and the vector of Lagrange multipliers λ respectively (Bertsekas, 1982). This is typically done by solving a sequence of optimization problems of $\mathcal{L}(\cdot, \lambda_k)$ followed by a first-order update of the multipliers $\lambda_{k+1} = \lambda_k - 2\beta h(x_k)$ depending on the penalty parameter β. The iterates are gradually pushed towards the feasible set by increasing the penalty parameter β. However, each optimization subproblem may be expensive, and the methods are sensitive to the choice of the penalty parameter β.

	Matrix factorizations	Complexity
Polar	matrix inverse square root	$\mathcal{O}(n^2p)$
SVD-based	SVD	$\mathcal{O}(n^2p)$
Cholesky-QR	Cholesky, matrix inverse	$\mathcal{O}(n^2p)$
$\Lambda(X)$ formula in (2)	None	$\min\{\mathcal{O}(n^2p), \mathcal{O}(npr)\}$

Table 1. Cost comparison of retractions and the landing formula on the generalized Stiefel manifold. We assume naive flop count for the matrix-matrix multiplication and no additional structure on matrix B_ζ apart from being rank-r in the stochastic setting. The matrices are of size $n \times p$ with $p \leq n$, and r is the rank of the stochastic matrices B_ζ. Matrix factorizations are hard to parallelize. The retractions do not allow for reduced complexities when B_ζ is low-rank and are not suited for stochastic B_ζ. For the numerical timings, see Figure 7b in the appendices.

Recently, a number of works explored the possibility of infeasible methods for optimization on Riemannian manifolds, when the feasible set is deterministic, in order to eliminate the cost of retractions, which can be limiting in some situations, e.g., when the evaluation of stochastic gradients of the objective is cheap. The works of Gao et al. (2019a; 2022a) proposed a modified augmented Lagrangian method which allows for fast computation and better bounds on the penalty parameter β. Ablin & Peyré (2022) designed a simple iterative method called *landing*, consisting of two orthogonal components, to be used on the orthogonal group, which was later expanded to the Stiefel manifold (Gao et al., 2022b; Ablin et al., 2023). Schechtman et al. (2023) expanded the *landing* approach to be used on a general smooth constraint using a non-smooth merit function. More recently, Goyens et al. (2024) analysed the classical Fletcher's augmented Lagrangian for solving smoothly constrained problems through the Riemannian perspective and proposed an algorithm that provably finds second-order critical points of the minimization problem. As the differentiability of the infeasible models relies on the second-order information of the objective, Xiao et al. (2024) proposed a constraint-dissolving model where the exact gradient and Hessian are convenient to compute.

1.2. Existing methods for the GEVP and CCA

Deterministic methods. A lot of effort has been spent in recent years on finding fast and memory-efficient solvers for CCA and the GEVP. The top-p GEVP, that seeks to find the eigenspace corresponding to the p largest eigenvalues of the pair (A, B), can be formulated as (1); this can be deduced from Absil et al. (2008, Proposition 2.2.1). As for CCA, it can be framed as (1) (Ge et al., 2016; Bhatia & Pacchiano, 2018) or as a minimization over a Cartesian product of two generalized Stiefel manifolds as in (3). The majority of the existing methods specialized for CCA and the GEVP that compute the top-p vector solution aim to circumvent the need to compute $B^{-\frac{1}{2}}$ or B^{-1}, e.g., by using an approximate solver to compute the action of multiplying with B^{-1}. The classic Lanczos algorithm for computation of eigenvalues can be adapted to the GEVP by noting that we can look for standard eigenvectors of $B^{-1}A$, see (Saad, 2011, Algorithm 9.1). Ma et al. (2015) propose `AppGrad` which performs a projected gradient descent with ℓ_2-regularization and proves its convergence when initialized sufficiently close to the minimum. The work of Ge et al. (2016) proposes `GenELinK` algorithm based on the block power method, using inexact linear solvers, that has provable convergence with a rate depending on $1/\delta$, where $\delta = \beta_p - \beta_{p+1}$ is the eigenvalue gap. Allen-Zhu & Li (2017) improve upon this in terms of the eigenvalue gap and proposes the doubly accelerated method `LazyCCA`, which is based on the shift-and-invert strategy with iteration complexity that depends on $1/\sqrt{\delta}$. Xu & Li (2020) present a first-order Riemannian algorithm that computes gradients using fast linear solvers to approximate the action of B^{-1} and performs polar retraction.

Stochastic methods. While the stochastic CCA problem is of high practical interest, fewer works consider it. Although several of the aforementioned deterministic solvers can be implemented for streaming data using sampled information (Ma et al., 2015; Wang et al., 2016; Meng et al., 2021), they do not analyse stochastic convergence. The main challenge comes from designing an unbiased estimator for the whitening part of the method that ensures the constraint $X^\top BX = I$ in expectation. Arora et al. (2017) propose a stochastic approximation algorithm, MSG, that keeps a running weighted average of covariance matrices used for projection, requiring computing $B^{-1/2}$ at each iteration. Additionally, the work of Gao et al. (2019b) proves stochastic convergence of an algorithm based on the shift-and-invert scheme and SVRG to solve linear subproblems, but only for the top-1 setting.

Comparison with the landing. Constrained optimization methods such as the augmented Lagrangian methods and Riemannian optimization techniques can be applied on stochastic problems when the gradient of the objective function is random but not on problems when the *feasible set is random*. The landing method has provable global convergence guarantees with the same asymptotic rate as its Riemannian counterpart, while also allowing for stochasticity in the constraint. Our work is conceptually related to the recently developed infeasible methods (Ablin & Peyré, 2022; Ablin et al., 2023; Schechtman et al., 2023), with the key difference of constructing a smooth merit function for a general constraint $h(x) = 0$, which is necessary for the convergence analysis of stochastic iterative updates that can have error in the normal space of $\mathcal{M}$. In Table 2 we show the overview of relevant GEVP/CCA methods by comparing their per-iteration complexity, memory requirements, and the type of proved convergence. Despite the landing iteration (2) being designed for a general non-convex smooth problem (1) and not being tailored specifically to the GEVP/CCA, we achieve theoretically interesting rate of convergence. Additionally, we provide an improved space complexity $\mathcal{O}(n(p+r))$ by not having to form the full matrix B and only to save the iterates and the streaming samples.

2. Landing on General Stochastic Constraints

This section is devoted to analyzing the landing method in the general case where the feasible set is given by the zero set of a smooth function. We will later use these results in Section 3 devoted to extending and analyzing the landing

	Stochastic	Matrix factorizations	Convergence	Per-iteration complexity	Memory
AppGrad (Ma et al., 2015, Theorem 2.1)	-	SVD	local linear	$\mathcal{O}(n^2p + p^3)$	n^2
`CCALin` (Ge et al., 2016, Theorem 7)	-	inexact linear solver	global linear	$\mathcal{O}(n^2p + p^3)$	n^2
`rgCCALin` (Xu & Li, 2020, Theorem 6.1)	-	inexact linear solver	global linear	$\mathcal{O}(n^2p + p^3)$	n^2
LazyCCA (Allen-Zhu & Li, 2017, Theorem 4.2)	-	inexact linear solver	global linear	$\mathcal{O}(n^2p + p^2n)$	n^2
MSG (Arora et al., 2017, Theorem 2.3)	✓	inverse square root	global sublinear	$\mathcal{O}(n^3)$	n^2
$\Lambda(X)$ formula in (2)	✓	None	global sublinear	$\mathcal{O}(npr)$	$n(p+r)$

Table 2. Summary of CCA and GEVP solvers for finding top-p vectors simultaneously. For CCA based on covariance matrices we assume that the number of samples is much greater than the dimension, *i.e.*, $N \gg n$. "Stochastic" marks methods with convergence analysis in expectation for the stochastic case. We assume that deterministic methods require forming the matrix B at the start with additional cost $\mathcal{O}(Nn^2)$ and store it in iterations to remove dependence of the complexity on N.

method (2) on $\mathrm{St}_B(p,n)$. The theory presented here improves on that of Schechtman et al. (2023) in two important directions. First, we introduce the notion of relative ascent direction, which allows us to consider a richer class than that of *geometry-aware orthogonal directions* (Schechtman et al., 2023, Eq. 18). Second, we do not require any structure on the noise term $\tilde{E}$ defined later in (7), for the stochastic case, while A2(iii) in Schechtman et al. (2023) requires the noise to be in the tangent space. This enhancement is due to the smoothness of our merit function $\mathcal{L}$, while Schechtman et al. (2023) consider a non-smooth merit function. Importantly, for the case of $\mathrm{St}_B(p,n)$ with the formula given in (2), there is indeed noise in the normal space, rendering Schechtman et al. (2023)'s theory inapplicable, while we show in the next section that Theorem 2.9 applies in that case.

Given a continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, we address the optimization problem

$$\min_{x\in\mathbb{R}^d} f(x) \quad \text{s.t.} \quad x \in \mathcal{M} = \left\{x \in \mathbb{R}^d \,:\, h(x) = 0\right\}, \quad (4)$$

where $h : \mathbb{R}^d \to \mathbb{R}^q$ is continuously differentiable, $q \in \mathbb{N}$ represents the number of constraints, and $\mathcal{M}$ defines a smooth manifold set. We will consider algorithms that stay within an initially prescribed ε-*safe region*

$$\mathcal{M}^\varepsilon = \left\{x \in \mathbb{R}^d \,:\, \|h(x)\| \le \varepsilon\right\},$$

which can be split into a collection of layered manifolds (Goyens et al., 2024)

$$\mathcal{M}_c = \left\{x \in \mathbb{R}^d \,:\, h(x) = c\right\},$$

with $\|c\| \le \varepsilon$.

The first assumption we make is that the gradient of f is Lipschitz continuous. The second one requires that the differential $\mathrm{D}h(x)$ inside the ε-safe region has bounded singular values.

Assumption 2.1 (Smoothness of the objective)**.** *The objective function $f : \mathbb{R}^d \to \mathbb{R}$ is L_f-smooth.*

Assumption 2.2 (Smoothness of the constraint)**.** *The differential of the constraint function has bounded singular values for x in the ε-safe region, i.e.,*

$$\forall x \in \mathcal{M}^\varepsilon : \quad \bar{C}_h \le \sigma\left(\mathrm{D}h(x)\right) \le C_h.$$

Additionally, the gradient $\nabla\mathcal{N}(x)$ of the penalty term $\mathcal{N}(x) = \frac{1}{2}\|h(x)\|^2$ is $L_\mathcal{N}$-smooth over $\mathcal{M}^\varepsilon$.

Assumption 2.1 is standard in optimization. Assumption 2.2 is necessary for the analysis of smooth constrained optimization (Goyens et al., 2024) and holds, e.g., when $\mathcal{M}^\varepsilon$ is a compact set, and $\mathrm{D}h(x)$ has full rank for all $x \in \mathcal{M}^\varepsilon$. This ensures that every layered manifold $\mathcal{M}_c$ is an embedded submanifold of $\mathbb{R}^d$. The tangent space to $\mathcal{M}_c$ at x is the null space of $\mathrm{D}h(x)$, the normal space at x (in the sense of the Frobenius inner product) is the range (i.e., image) of $\mathrm{D}h(x)^*$, and a *critical point* is then a point x in $\mathcal{M}$ where $\nabla f(x)$ belongs to the normal space.

Next we define the notion of relative ascent direction, used to guarantee that (2) produces a descent when the second term of the landing field vanishes.

Definition 2.1 (Relative ascent direction)**.** *A relative ascent direction $\Psi(x) : \mathbb{R}^d \to \mathbb{R}^d$, with a parameter $\rho > 0$ that may depend on ε, satisfies:*

(i) $\forall x \in \mathcal{M}^\varepsilon, \quad \forall v \in \mathrm{range}(\mathrm{D}h(x)^*) : \langle \Psi(x), v\rangle = 0$;

(ii) $\forall x \in \mathcal{M}^\varepsilon, \quad \langle \Psi(x), \nabla f(x)\rangle \ge \rho\|\Psi(x)\|^2$;

(iii) $\forall x \in \mathcal{M}, \quad \langle \Psi(x), \nabla f(x)\rangle = 0$ *if and only if x is a critical point of f subject to $\mathcal{M}$.*

In short, the relative ascent direction must be in the tangent space to every layered manifold $\mathcal{M}_{h(x)}$ while remaining positively aligned with the Euclidean gradient $\nabla f(x)$. Note that the above definition is not scale invariant to ρ, i.e., taking $c\Psi(x)$ for $c > 0$ will result in $c\rho$, and this is in line with the forthcoming convergence guarantees deriving an upper bound on $\|\Psi(x)\|$. While there may be many examples of relative ascent directions, a particular example is given next.

Definition 2.2 (Riemannian gradient on the layered manifold $\mathcal{M}_c$)**.** *Let $f : \mathcal{M}_c \to \mathbb{R}$ be a smooth function on*

$\mathcal{M}_c$. *The* Riemannian gradient *of* f*, denoted by* $\operatorname{grad} f$*, is uniquely defined by*

$$\forall x \in \mathcal{M}_c, v \in \mathrm{T}_x\mathcal{M}_c, \quad \mathrm{D}f(x)[v] = \langle v, \operatorname{grad} f(x)\rangle,$$

where $\mathrm{T}_x\mathcal{M}_c$ *denotes the tangent space of* $\mathcal{M}_c$ *at* x.

Proposition 2.3 (Riemannian gradient is a relative ascent direction)**.** *The Riemannian gradient defined in Definition 2.2 is a relative ascent direction on* $\mathcal{M}^\varepsilon$ *with* $\rho = 1$.

The proof can be found in the appendices in subsection C.1. Such extension of the Riemannian gradient to the collection of layered manifolds was already considered by Gao et al. (2022b) in the particular case of the Stiefel manifold and by Schechtman et al. (2023).

2.1. Deterministic case

We now define the general form of the *deterministic* landing iteration as

$$x^{k+1} = x^k - \eta_k \Lambda(x^k) \text{ with } \Lambda(x) = \Psi(x) + \omega \nabla \mathcal{N}(x), \quad (5)$$

where $\Psi(x)$ is a relative ascent direction described in Definition 2.1, $\nabla\mathcal{N}(x) = \mathrm{D}h(x)^* h(x)$ is the gradient of the penalty $\mathcal{N}(x) = \frac{1}{2}\|h(x)\|^2$, $\omega > 0$ is a parameter, and $\|\cdot\|$ is the ℓ_2-norm. Condition (i) in Definition 2.1 guarantees that $\langle \nabla\mathcal{N}(x), \Psi(x)\rangle = 0$, so that the two terms in Λ are orthogonal.

Note that we can use *any* relative ascent direction as $\Psi(x)$. The Riemannian gradient in (10) is just one special case, which has some shortcomings. Firstly, it requires a potentially expensive projection $\mathrm{D}h(x)^* \left(\mathrm{D}h(x)^*\right)^\dagger$. Secondly, if the constraint involves a random noise on h, formula (10) does not give an unbiased formula in expectation. An important contribution of the present work is the derivation of a computationally convenient relative ascent direction in the specific case of the generalized Stiefel manifold in Section 3.

We now turn to the analysis of the convergence of this method. The main object allowing for the convergence analysis is Fletcher's augmented Lagrangian

$$\mathcal{L}(x) = f(x) - \langle h(x), \lambda(x)\rangle + \beta\|h(x)\|^2, \quad (6)$$

with the Lagrange multiplier $\lambda(x) \in \mathbb{R}^p$ defined as $\lambda(x) = (\mathrm{D}h(x)^*)^\dagger[\nabla f(x)]$ (Goyens et al., 2024). The next assumption that the differential of $\lambda(x)$ is bounded is met when $\mathcal{M}^\varepsilon$ is a compact set.

Assumption 2.3 (Multipliers of Fletcher's augmented Lagrangian)**.** *The norm of the differential of the multipliers of Fletcher's augmented Lagrangian is bounded:* $\sup_{x\in\mathcal{M}^\varepsilon} \|\mathrm{D}\lambda(x)\| \le C_\lambda$.

Proposition 2.4 (Lipschitz constant of Fletcher's augmented Lagrangian)**.** *Fletcher's augmented Lagrangian* $\mathcal{L}$ *in* (6) *is* $L_\mathcal{L}$*-smooth on* $\mathcal{M}^\varepsilon$*, with* $L_\mathcal{L} = L_{f+\lambda} + 2\beta L_\mathcal{N}$*, where* $L_{f+\lambda}$ *is the smoothness constant of* $f(x) - \langle h(x), \lambda(x)\rangle$ *and* $L_\mathcal{N}$ *is that of* $\mathcal{N}(x)$.

Proof. By the smoothness of $f(x) - \langle h(x), \lambda(x)\rangle$ and $\mathcal{N}(x)$ combined with the triangle inequality for $\|\cdot\|$. □

The following two lemmas show that there exists a positive step-size η that guarantees that the next landing iteration stays within $\mathcal{M}^\varepsilon$ provided that the current iterate is inside $\mathcal{M}^\varepsilon$.

Lemma 2.5 (A step-size safeguard)**.** *Let* $x \in \mathcal{M}^\varepsilon$ *and consider the iterative update* $\tilde{x} = x - \eta\Lambda(x)$*, where* $\eta > 0$ *is a step-size and* $\Lambda(x)$ *is the landing field in* (5)*. If the step-size satisfies* $\eta \le \eta(x)$ *with*

$$\eta(x) := \frac{1}{L_\mathcal{N}\|\Lambda(x)\|^2}\Big(\omega\|\nabla\mathcal{N}(x)\|^2 + \sqrt{\omega^2\|\nabla\mathcal{N}(x)\|^4 + L_\mathcal{N}\|\Lambda(x)\|^2(\varepsilon^2 - \|h(x)\|^2)}\Big)$$

where $L_\mathcal{N}$ *is from Assumption 2.2, then the line segment from the current to the next iterate remains in the safe region.*

The proof can be found in the appendices in subsection C.2. Next, we require that the norm of the relative ascent direction must remain bounded in the safe region.

Assumption 2.4 (Bounded relative ascent direction)**.** *We require that* $\sup_{x\in\mathcal{M}^\varepsilon} \|\Psi(x)\| \le C_\Psi$.

This holds, for instance, if ∇f is bounded in $\mathcal{M}^\varepsilon$, using Definition 2.1 (ii) and Cauchy-Schwarz inequality. Under this assumption, we can lower bound the step-size safeguard in Lemma 2.5 for all $x \in \mathcal{M}^\varepsilon$, implying that there is always a positive step-size that keeps the next iterate in the safe region.

Lemma 2.6 (A lower-bound on the step-size safeguard)**.** *The step-size safeguard* $\eta(x)$ *in Lemma 2.5 is lower bounded away from zero by*

$$\underline{\eta} := \min\left\{\frac{\omega\bar{C}_h^2\alpha^2\varepsilon^2}{L_\mathcal{N}\left(C_\Psi^2 + \omega^2 C_h^2\varepsilon^2\right)}, \frac{(1-\alpha)\varepsilon}{\sqrt{2L_\mathcal{N}}}, \frac{(1-\alpha)\varepsilon}{\sqrt{2L_\mathcal{N}}\left(C_\Psi^2 + \omega^2 C_h^2\varepsilon^2\right)}, \frac{1}{\omega L_\mathcal{N}}\left(\frac{\bar{C}_h}{C_h}\right)^2\right\}$$

for any choice of $0 < \alpha < 1$ *where* $C_h, \bar{C}_h, C_\Psi > 0$ *are constants from Assumption 2.2 and 2.4.*

The proof can be found in subsection C.3.

Lemma 2.7. *Let* $\mathcal{L}(x)$ *be Fletcher's augmented Lagrangian in* (6) *with* $\beta \ge (\frac{\rho}{4C_h^2} + \frac{\omega C_\lambda}{2C_h} + \frac{C_\lambda^2}{4\rho C_h^2})/\omega$*, where* ρ *is defined in Definition 2.1. We have that* $\langle \nabla\mathcal{L}(x), \Lambda(x)\rangle \ge \frac{\rho}{2}\left(\|\Psi(x)\|^2 + \|h(x)\|^2\right)$.

The proof can be found in the appendices in subsection C.4. This critical lemma shows that $\mathcal{L}$ is a valid merit function for the landing iterations and allows the study of convergence of the method with ease.

The following statement combines Lemma 2.7 with the bound on the step-size safeguard in Lemma 2.6 to prove sublinear convergence to critical points of f subject to $\mathcal{M}$.

Theorem 2.8 (Convergence of the deterministic landing)**.** *Under the above assumptions, the landing iteration in* (5) *starting from $x_0 \in \mathcal{M}^\varepsilon$ satisfies*

$$\frac{1}{K}\sum_{k=0}^{K}\|\Psi(x^k)\|^2 \leq 4\frac{\mathcal{L}(x^0)-\mathcal{L}^*}{\eta\rho K},$$
$$\frac{1}{K}\sum_{k=0}^{K}\|h(x_k)\|^2 \leq 4\frac{\mathcal{L}(x^0)-\mathcal{L}^*}{\eta\rho\omega^2 K}.$$

for $\eta \leq \min\left\{\frac{\rho}{2L_{\mathcal{L}}}, \frac{\rho}{2L_{\mathcal{L}}C_h^2}, \underline{\eta}\right\}$ and $\mathcal{L}^* = \min_{x\in\mathcal{M}^\varepsilon}\mathcal{L}(x)$, where $\underline{\eta}$ comes from Lemma 2.6.

The proof is given in subsection C.5 and implies that the iterates x^k converge to critical points with the sublinear rate $\mathcal{O}(1/K)$.

2.2. Stochastic case

Due to the smoothness of Fletcher's augmented Lagrangian in the $\mathcal{M}^\varepsilon$ region, we can extend the convergence result to the stochastic setting. The iteration is

$$x^{k+1} = x^k - \eta_k\left[\Lambda(x^k) + \tilde{E}(x^k, \Xi^k)\right], \tag{7}$$

where the Ξ^k are i.i.d. random variables, $\tilde{E}(x^k, \Xi^k)$ is the random error term at iteration x^k, and $\Lambda(x^k)$ is the landing field in (5). We require the landing update in (7) to be an unbiased estimator with bounded variance.

Assumption 2.5 (An unbiased estimator of $\Lambda(x^k)$ with bounded variance)**.** *There exists $\gamma > 0$ such that for all $x \in \mathcal{M}^\varepsilon$, we have $\mathbb{E}_\Xi[\tilde{E}(x,\Xi)] = 0$ and $\mathbb{E}_\Xi[\|\tilde{E}(x,\Xi)\|^2] \leq \gamma^2$.*

We obtain the following result with decaying step-sizes.

Theorem 2.9 (Convergence of the stochastic landing)**.** *Under the above assumptions, the stochastic landing iteration in* (7) *with a diminishing step-size $\eta_k = \eta_0 \times (1+k)^{-1/2}$, and assuming the line segments between the iterates remain within $\mathcal{M}^\varepsilon$ with probability one, produces iterates for which*

$$\inf_{k\leq K}\mathbb{E}\left[\|\Psi(x_k)\|^2\right] \leq 4\frac{\mathcal{L}(x^0)-\mathcal{L}^*}{\rho\eta_0\sqrt{K}} + \frac{2\eta_0\gamma^2 L_{\mathcal{L}}(1+\log(K+1))}{\rho\sqrt{K}},$$
$$\inf_{k\leq K}\mathbb{E}\left[\|h(x)\|^2\right] \leq 4\frac{\mathcal{L}(x^0)-\mathcal{L}^*}{\omega^2\rho\eta_0\sqrt{K}} + \frac{2\eta_0\gamma^2 L_{\mathcal{L}}(1+\log(K+1))}{\omega^2\rho\sqrt{K}},$$

for $\eta_0 \leq \frac{\rho}{2L_{\mathcal{L}}}\min\left\{1, C_h^{-2}\right\}$ and $\mathcal{L}^* = \min_{x\in\mathcal{M}^\varepsilon}\mathcal{L}(x)$.

The theorem is proved in subsection C.6. Unlike in the deterministic case in Lemma 2.5, without further assumption on the distribution of Ξ^k, it cannot be ensured that the line segments connecting the successive iterates are within $\mathcal{M}^\varepsilon$ with probability one. Under that assumption, we recover the same convergence rate as Riemannian SGD in the non-convex setting for a deterministic feasible set (Bonnabel, 2013), but in our case, we require only an online estimate of the random manifold feasible set.

3. Landing on the Generalized Stiefel Manifold

This section builds on the results of the previous Section 2 and proves that the simple landing update rule $X^{k+1} = X^k - \eta_k\Lambda(X^k)$, defined in (2), converges to the critical points of (1). The generalized Stiefel manifold $\mathrm{St}_B(p,n)$ is defined by the constraint function $h(X) = X^\top BX - I_p$, and we have $\nabla\mathcal{N}(X) = 2BX(X^TBX - I_p)$. We now derive the quantities required for Assumption 2.2. Recall that β_i denotes the ith eigenvalue of B and $\kappa_B = \beta_1/\beta_n$ is the condition number of B.

Proposition 3.1 (Smoothness constants for the generalized Stiefel manifold)**.** *Smoothness constants in Assumption 2.2 for the generalized Stiefel manifold are*

$$C_h = 2\sqrt{(1+\varepsilon)\beta_1\kappa_B} \quad \textit{and} \quad \bar{C}_h = 2\sqrt{(1-\varepsilon)\beta_n\kappa_B^{-1}}.$$

The proof is presented in subsection D.2.

We show two candidates for the relative ascent direction:

Proposition 3.2 (Relative ascent directions for the generalized Stiefel manifold)**.** *The following two formulas are relative ascent directions on the generalized Stiefel manifold:*

$$\Psi_B(X) = 2\mathrm{skew}(\nabla f(X)X^\top B)BX$$
$$\Psi_B^{\mathrm{R}}(X) = 2\mathrm{skew}(B^{-1}\nabla f(X)X^\top)BX$$

with $\Psi_B(X)$ having $\rho_B = 1/(\beta_1\kappa_B(1+\varepsilon))$ and $\Psi_B^{\mathrm{R}}(X)$ having $\rho_B^{\mathrm{R}} = \beta_n/(1+\varepsilon)$.

The proof is given in subsection D.3. The formula for the relative ascent $\Psi_B^{\mathrm{R}}(X)$ can be derived as a Riemannian gradient for $\mathrm{St}_B(p,n)$ in a metric derived from a canonical metric on the standard Stiefel manifold via a specific isometry; see Appendix E.

3.1. Deterministic generalized Stiefel case

The fact that $\Psi_B(X)$ above meets the conditions of Definition 2.1 allows us to define the deterministic landing iterations as $X^{k+1} = X^k - \eta^k \Lambda(X^k)$ with

$$\Lambda(X) = 2\,\mathrm{skew}(\nabla f(X) X^\top B) BX + 2\omega BX(X^T BX - I_p), \quad (8)$$

and Theorem 2.8 applies to these iterations, showing that they converge to critical points.

3.2. Stochastic generalized Stiefel case

One of the main features of the formulation in (8) is that it seamlessly extends to the stochastic case when both the objective f and the constraint matrix B are expectations. Indeed, using the stochastic estimate $\Lambda_{\xi,\zeta,\zeta'}$ defined in (2), we have $\mathbb{E}_{\xi,\zeta,\zeta'}[\Lambda_{\xi,\zeta,\zeta'}(X)] = \Lambda(X)$. The stochastic landing iterations are, therefore, of the same form as in (7). To apply Theorem 2.9 we need to bound the variance of $\tilde{E}(X,\Xi) = \Lambda_{\xi,\zeta,\zeta'}(X) - \Lambda(X)$ where the random variable Ξ is the triplet (ξ,ζ,ζ') using standard U-statistics techniques (Van der Vaart, 2000).

Proposition 3.3 (Variance estimation of the generalized Stiefel landing iteration)**.** *Let σ_B^2 be the variance of B_ζ and σ_G^2 the variance of $\nabla f_\xi(X)$. We have that*

$$\mathbb{E}_\Xi[\|\tilde{E}(X,\Xi)\|^2] \le \sigma_G^2 \alpha_G + \sigma_B^2(\alpha_B + \omega^2 \gamma_B) \quad (9)$$

where the constants $\alpha_G, \alpha_B, \gamma_B$ are given explicitly in subsection D.5, and depend only on ε, the distribution of B_ζ, and the function f.

The proof is found in subsection D.5. Note that, as expected, the variance in (9) is zero in the deterministic setting where both variances σ_B and σ_G are zero. A consequence of Proposition 3.3 is that Theorem 2.9 applies in the case of the stochastic landing method on the generalized Stiefel manifold, and more specifically, also for solving the stochastic GEVP.

4. Numerical Experiments

Generalized eigenvalue problem. We compare the methods on the deterministic top-p GEVP that consists of solving $\min_{X\in\mathbb{R}^{n\times p}} -\frac{1}{2}\mathrm{Tr}(X^\top AX)$ for $X \in \mathrm{St}_B(p,n)$. The two matrices are randomly generated with a condition number $\kappa_A = \kappa_B = 100$ and with the size $n = 1000$ and $p = 500$; see further specifics in Appendix B.[1]

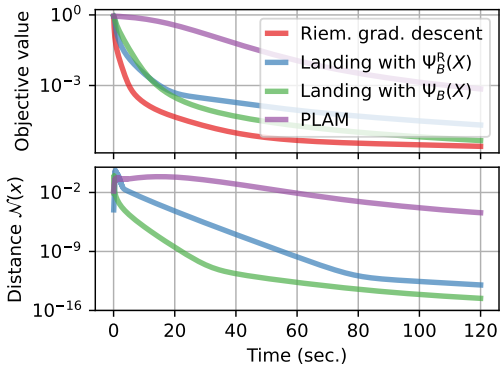


Figure 2. Generalized eigenvalue problem ($n = 1000, p = 500$).

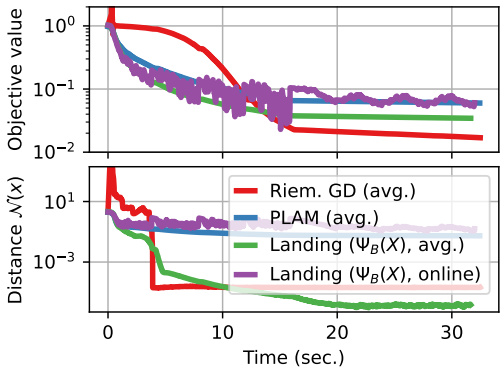


Figure 3. Stochastic CCA on the split MNIST dataset for $p = 5$. An epoch takes roughly 2.5 sec.

Figure 2 shows the timings of four methods with fixed step-size: Riemannian steepest descent with QR-based Cholesky retraction (Sato & Aihara, 2019), the PLAM method (Gao et al., 2022a), and the two landing methods with either $\Psi_B^{\mathrm{R}}(X)$ or $\Psi_B(X)$ in Proposition 3.2. The landing method with $\Psi_B(X)$ converges the fastest in terms of time, due to its cheap per-iteration computation, which is also demonstrated in Figure 5 and Figure 7 in the appendices. It can be also observed that the landing method with $\Psi_B(X)$ is more robust to the choice of parameters η and ω compared to PLAM, which we show in Figure 8 and Figure 10 in the appendices, and is in line with the equivalent observations previously made for the orthogonal manifold (Ablin & Peyré, 2022, Figure 9). In Figure 9 in the appendices we track numerically the value of the step-size safeguard $\eta(X)$ in Lemma 2.5.

Stochastic CCA and ICA. For stochastic CCA, we use the benchmark problem used by Ma et al. (2015); Wang et al. (2016), in which the MNIST dataset is split in half by

[1] The code is available at: https://github.com/simonvary/landing-generalized-stiefel.

taking left and right halves of each image, and compute the top-p canonical correlation components by solving (3). In our experiments, we have $N = 60\,000$, $n = 392$, $p = 5$, and $r = 512$.

The stochastic ICA is performed by solving (Hyvarinen, 1999; Ablin et al., 2018)

$$\min_{X \in \mathbb{R}^{n\times n}} \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{n} \sigma([AX]_{i,j}), \text{ s.t. } X \in \mathrm{St}_{\frac{1}{N}A^\top A}(n,n)$$

where $\sigma(x) = \log(\cosh(x))$ is performed elementwise and $\sigma'(x) = \tanh(x)$. We generate the data matrix A as $A = SW^\top$, where S is a $N \times n$ matrix of random i.i.d. data sampled from a Laplace distribution and W is a $n \times n$ random orthogonal matrix. We take $N = 100\,000$ and $n = 10$. By solving the above optimization problem, the goal of ICA is to recover the mixing matrix W, up to scaling and permutations invariances; to monitor this we track the Amari distance (Amari et al., 1995) between X and W^{-1}.

Figure 3 and Figure 4 show the timings for the Riemannian gradient descent with rolling averaged covariance matrix and the landing algorithm with $\Psi_B(X)$ in its online and averaged form for the CCA and the ICA experiment respectively. The averaged methods keep track of the covariance matrices during the first pass through the dataset, which is around 3 sec. and 0.6 sec. respectively, after which they have the exact fully sampled covariance matrices. The online methods have always only the sampled estimate with the batch size of $r = 512$. All methods use the fixed stepsize $\eta = 0.1$, and the landing methods have $\omega = 1$. In practice, the hyperparameters can be picked by grid-search as is common for stochastic optimization methods.

The online landing method outperforms the averaged Riemannian gradient descent in the online setting in terms of the objective value after only a few passes over the data, e.g., at the 3 sec. mark and the 0.6 sec. mark respectively in Figure 3 and Figure 4, which corresponds to the first epoch, at which point each sample appeared just once. After the first epoch, the rolling avg. methods get the advantage of the exact fully sampled covariance matrix and, consequently, have better distance $\mathcal{N}(X)$, but at the cost of requiring $\mathcal{O}(n^2)$ memory for the full covariance matrix. The online method does not form B and requires only $\mathcal{O}(n(p+r))$ memory. The behavior is also consistent when $p = 10$ as shown in Figure 6 in the appendices.

5. Conclusion

We have extended the theory of the landing method from the Stiefel manifold to the general case of a feasible set defined by smooth equations $h(x) = 0$. We have improved the existing analysis by using a smooth merit function, which allows us to also consider situations where we have only random estimates of the manifold. We have showed that the random generalized Stiefel manifold, which is central to problems such as stochastic CCA, ICA, and the GEVP, falls into the category of random manifold feasible set and derived specific bounds for it.

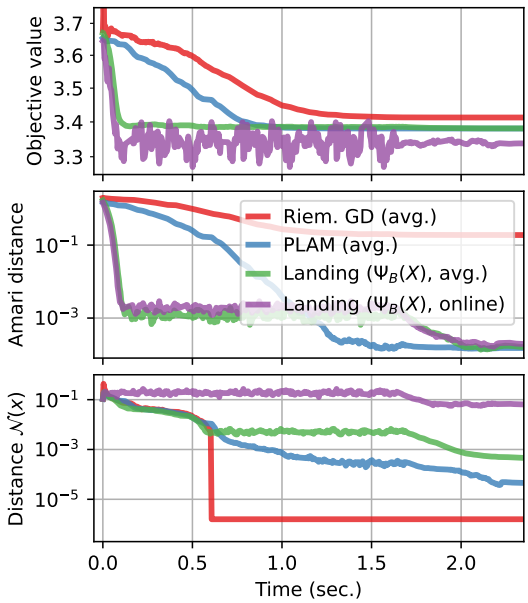


Figure 4. Stochastic ICA on the synthetic dataset for $n = 10$.

Acknowledgments

This work was supported by the Fonds de la Recherche Scientifique-FNRS under Grant no T.0001.23. Simon Vary was supported by the FSR Incoming Post-doctoral Fellowship and the Incentive Grant for Scientific Research (MIS) "Learning from Pairwise Data" of the F.R.S.-FNRS. Bin Gao was supported by the Young Elite Scientist Sponsorship Program by CAST and the National Natural Science Foundation of China (grant No. 12288201).

Impact Statement

This paper presents theoretical work which aims to advance the field of machine learning. There is no broad impact other than the consequences discussed in the paper.

References

Ablin, P. and Peyré, G. Fast and accurate optimization on the orthogonal manifold without retraction. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 51, Valencia, Spain, 2022. PMLR.

Ablin, P., Cardoso, J.-F., and Gramfort, A. Faster ICA under orthogonal constraint. In *2018 IEEE International*

Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4464–4468. IEEE, 2018.

Ablin, P., Vary, S., Gao, B., and Absil, P.-A. Infeasible Deterministic, Stochastic, and Variance-Reduction Algorithms for Optimization under Orthogonality Constraints. *arXiv preprint arXiv:2303.16510*, 2023.

Absil, P.-A., Baker, C., and Gallivan, K. Trust-Region Methods on Riemannian Manifolds. *Foundations of Computational Mathematics*, 7(3):303–330, July 2007. doi: 10.1007/s10208-005-0179-9.

Absil, P.-A., Mahony, R., and Sepulchre, R. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, January 2008. ISBN 978-1-4008-3024-4. doi: 10.1515/9781400830244.

Ahn, K. and Sra, S. From Nesterov's Estimate Sequence to Riemannian Acceleration. In *Proceedings of Machine Learning Research*, volume 125, pp. 1–35, 2020.

Allen-Zhu, Z. and Li, Y. Doubly Accelerated Methods for Faster CCA and Generalized Eigendecomposition. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, Sydney, Australia, 2017.

Amari, S.-i., Cichocki, A., and Yang, H. A new learning algorithm for blind signal separation. *Advances in neural information processing systems*, 8, 1995.

Arora, R., Marinov, T. V., Mianjy, P., and Srebro, N. Stochastic approximation for canonical correlation analysis. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Berger, G. O., Absil, P.-A., Jungers, R. M., and Nesterov, Y. On the quality of first-order approximation of functions with Hölder continuous gradient. *Journal of Optimization Theory and Applications*, 185:17–33, 2020.

Bertsekas, D. P. *Constrained Optimization and Lagrange Multiplier Methods*. Athena Scientific, 1982. ISBN 1-886529-04-3.

Bhatia, K. and Pacchiano, A. Gen-Oja: A Simple and Efficient Algorithm for Streaming Generalized Eigenvector Computation. In *Proceedings of the 32nd Conference on Neural Information Processing Systems*, 2018.

Bonnabel, S. Stochastic Gradient Descent on Riemannian Manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, September 2013. doi: 10.1109/TAC.2013.2254619.

Boumal, N. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023. doi: 10.1017/9781009166164. URL https://www.nicolasboumal.net/book.

Boumal, N., Absil, P. A., and Cartis, C. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 39(1):1–33, 2019. doi: 10.1093/imanum/drx080.

Comon, P. Independent component analysis, A new concept? *Signal Processing*, 36(3):287–314, April 1994. doi: 10.1016/0165-1684(94)90029-9.

Gao, B., Liu, X., and Yuan, Y.-x. Parallelizable Algorithms for Optimization Problems with Orthogonality Constraints. *SIAM Journal on Scientific Computing*, 41(3):A1949–A1983, January 2019a. doi: 10.1137/18M1221679.

Gao, B., Hu, G., Kuang, Y., and Liu, X. An orthogonalization-free parallelizable framework for all-electron calculations in density functional theory. *SIAM Journal on Scientific Computing*, 44(3):B723–B745, 2022a. doi: 10.1137/20M1355884.

Gao, B., Vary, S., Ablin, P., and Absil, P.-A. Optimization flows landing on the Stiefel manifold. *IFAC-PapersOnLine*, 55(30):25–30, 2022b. doi: 10.1016/j.ifacol.2022.11.023. 25th IFAC Symposium on Mathematical Theory of Networks and Systems MTNS 2022.

Gao, C., Garber, D., Srebro, N., Wang, J., and Wang, W. Stochastic Canonical Correlation Analysis. *Journal of Machine Learning Research*, 20:1–46, 2019b.

Ge, R., Jin, C., Kakade, S., Netrapalli, P., and Sidford, A. Efficient Algorithms for Large-scale Generalized Eigenvector Computation and Canonical Correlation Analysis. In *Proceedings of the 33th International Conference on Machine Learning*, volume 48, New York, NY, USA, 2016.

Goyens, F., Eftekhari, A., and Boumal, N. Computing Second-Order Points Under Equality Constraints: Revisiting Fletcher's Augmented Lagrangian. *Journal of Optimization Theory and Applications*, April 2024. doi: 10.1007/s10957-024-02421-6.

Hotelling, H. Relations between two sets of variates. *Biometrika*, 28(3-4):321–377, 1936. doi: 10.1093/biomet/28.3-4.321.

Hyvarinen, A. Fast and robust fixed-point algorithms for independent component analysis. *IEEE transactions on Neural Networks*, 10(3):626–634, 1999.

Ma, Z., Lu, Y., and Foster, D. Finding Linear Structure in Large Datasets with Scalable Canonical Correlation Analysis. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

McLachlan, G. J. *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons, 1992. ISBN 9780471615316.

Meng, Z., Chakraborty, R., and Singh, V. An Online Riemannian PCA for Stochastic Canonical Correlation Analysis. In *35th Conference on Neural Information Processing Systems (NeurIPS 2021)*, 2021.

Mishra, B. and Sepulchre, R. Riemannian Preconditioning. *SIAM Journal on Optimization*, 26(1):635–660, January 2016. doi: 10.1137/140970860.

Nocedal, J. and Wright, S. J. *Numerical Optimization*. Springer, New York, NY, USA, 2 edition, 2006.

Raghu, M., Gilmer, J., Yosinski, J., and Sohl-Dickstein, J. SVCCA: Singular Vector Canonical Correlation Analysis for Deep Learning Dynamics and Interpretability. In *31st Conference on Neural Information Processing Systems*, Long Beach, CA, USA, 2017.

Saad, Y. *Numerical Methods for Large Eigenvalue Problems*. Society for Industrial and Applied Mathematics, 2011. doi: 10.1137/1.9781611970739.

Sato, H. and Aihara, K. Cholesky QR-based retraction on the generalized Stiefel manifold. *Computational Optimization and Applications*, 72(2):293–308, March 2019. doi: 10.1007/s10589-018-0046-7.

Schechtman, S., Tiapkin, D., Muehlebach, M., and Moulines, E. Orthogonal Directions Constrained Gradient Method: From non-linear equality constraints to Stiefel manifold. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195, pp. 1228–1258. PMLR, 2023.

Van der Vaart, A. W. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Wang, W., Wang, J., Garber, D., and Srebro, N. Efficient Globally Convergent Stochastic Optimization for Canonical Correlation Analysis. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016.

Xiao, N., Liu, X., and Toh, K.-C. Dissolving constraints for Riemannian optimization. *Mathematics of Operations Research*, 49(1):366–397, 2024.

Xu, Z. and Li, P. A Practical Riemannian Algorithm for Computing Dominant Generalized Eigenspace. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, volume 124, pp. 819–828. PMLR, 2020.

Yger, F., Berar, M., Gasso, G., and Rakotomamonjy, A. Adaptive Canonical Correlation Analysis Based On Matrix Manifolds. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

Zhang, H. and Sra, S. First-order Methods for Geodesically Convex Optimization. In *Conference on Learning Theory (COLT 2016)*, volume 49, pp. 1–22, 2016.

A. Summary of Retractions on the Generalized Stiefel Manifold

For an update to a matrix $X \in \mathrm{St}_B(p,n)$ following a direction in the tangent space $Z \in \mathrm{T}_X\mathrm{St}_B(p,n)$ (see Appendix E for an expression of $\mathrm{T}_X\mathrm{St}_B(p,n)$), there are several ways to compute a retraction. The following asymptotic flop counts provide a simplified picture of computational cost: they do not reflect opportunities for parallelism and assume no structure on matrix B.

- The *Polar decomposition* (Yger et al., 2012) uses

$$\mathrm{Retr}_{\mathrm{St}_B}(X,Z) = (X+Z)\left(I_p + Z^\top BZ\right)^{-1/2},$$

  involving the multiplication of B by an $n \times p$ matrix and the computation of the inverse matrix square root of a $p \times p$ matrix, which in naive implementation amounts to $\mathcal{O}(n^2p)$ flops.

- Mishra & Sepulchre (2016) observed that the aforementioned polar decomposition can be expressed as $UV^\top$ in terms of an SVD-like decomposition of $X+Z = U\Sigma V^\top$, where U, V are orthogonal with respect to B-inner product, whose main cost is the eigendecomposition of $(X+Z)^\top B(X+Z)$.

- Recently, Sato & Aihara (2019) proposed the *Cholesky-QR based retraction*

$$\mathrm{Retr}_{\mathrm{St}_B}(X,Z) = (X+Z)R^{-1},$$

  where $R \in \mathbb{R}^{p\times p}$ comes from the Cholesky factorization of $R^\top R = (X+Z)^\top B(X+Z)$. The flops required for the computation, in naive implementation, amount to $\mathcal{O}(n^2p)$, which comes from the matrix multiplications. The Cholesky factorization of an $p \times p$ matrix and the inverse multiplication by a small triangular $p \times p$ matrix requires $\mathcal{O}(p^3)$ to form and $\mathcal{O}(np^2)$ to multiply with.

B. Additional Experiments and Figures

For the experiment showed in Fig. 2, we generate the matrix $A \in \mathbb{R}^{n\times n}$ to have equidistant eigenvalues $\lambda_i(A) \in [1/\kappa_A, 1]$ and $B \in \mathbb{R}^{n\times n}$ has exponentially decaying eigenvalues $\lambda_i(B) \in [1/\kappa_B, 1]$. We pick the step-size η parameter to be $\eta = 0.01$ for the Riemannian gradient descent, the landing with $\Psi_B^{\mathrm{R}}(X)$, and PLAM, and $\eta = 200$ for the landing with $\Psi_B(X)$ and we run a grid-search with step-sizes $c\eta$, where $c \in [1/4, 1/2, 1, 2, 4, 8]$. The normalizing parameter ω is chosen to be $\omega = 10^5$ for the landing with $\Psi_B^{\mathrm{R}}(X)$, $\omega = 0.1$ for the landing with $\Psi_B(X)$, and $\omega = 200$ for PLAM.

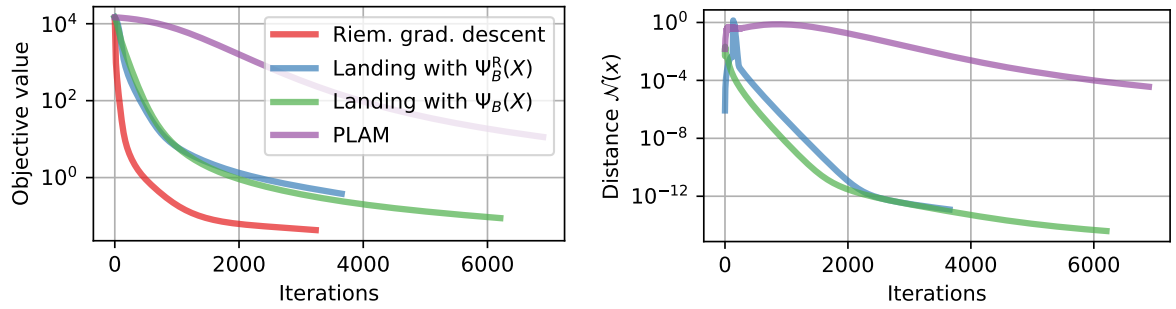


Figure 5. Deterministic computation of the generalized eigenvalue problem with $n = 1000, p = 500$, the condition number of the two matrices $\kappa_B = \kappa_A = 100$. Each algorithm is given a time limit of 120 seconds.

C. Proofs for Section 2

C.1. Proof of Proposition 2.3

Proof. The Riemannian gradient can be computed as

$$\mathrm{grad} f(x) = \nabla f(x) - \mathrm{D}h(x)^* \left(\mathrm{D}h(x)^*\right)^\dagger \nabla f(x), \tag{10}$$

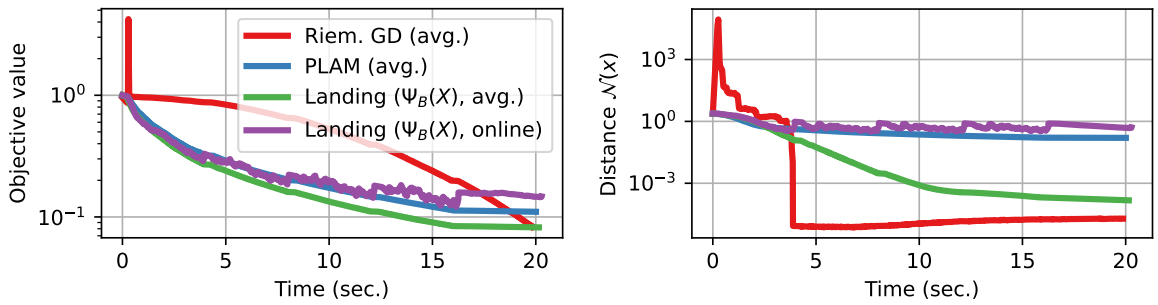
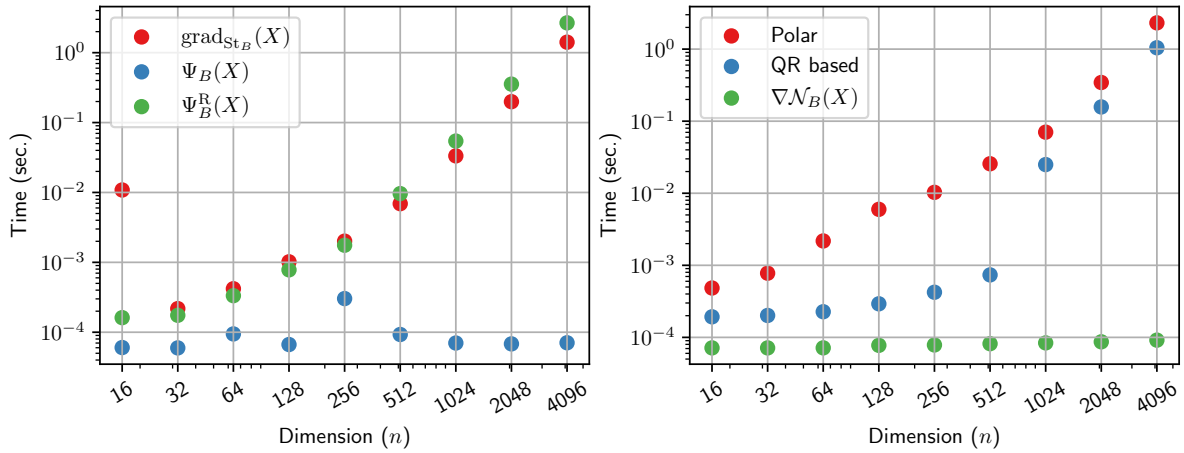


Figure 6. Stochastic canonical correlation analysis on the split MNIST dataset for $p = 10$ canonical components.



(a) Descent directions. (b) Retractions.

Figure 7. Comparison of per-iteration computational time for different problem sizes of the descent directions of algorithms in Fig. 2 and the cost of retractions compared to $\nabla\mathcal{N}(X)$, both in the deterministic setting when $n = p = r$, for which the matrix multiplication in $\Psi_B(X)$ and $\nabla_{\mathcal{N}}(X)$ are at the disadvantage. Computation time of randomly generated $B, X \in \mathbb{R}^{n\times n}$ averaged over 100 runs with `CUDA` implementation using `cupy`.

where $\mathrm{D}h(x)^* \left(\mathrm{D}h(x)^*\right)^\dagger$ is the orthogonal projection on the normal space of $\mathcal{M}_{h(x)}$. It follows from (10) and $\mathrm{D}h(x)\mathrm{D}h(x)^* \left(\mathrm{D}h(x)^*\right)^\dagger = \mathrm{D}h(x)$ that $\mathrm{D}h(x)[\mathrm{grad} f(x)] = 0$, which implies the first condition in Definition 2.1 holds, i.e., $\langle \mathrm{grad} f(x),\, v\rangle = 0$ for all $v \in \mathrm{range}(\mathrm{D}h(x)^*)$. Since $\mathrm{D}h(x)^* \left(\mathrm{D}h(x)^*\right)^\dagger \nabla f(x) \in \mathrm{range}(\mathrm{D}h(x)^*)$, we have

$$
\begin{aligned}
\|\mathrm{grad} f(x)\|^2 &= \langle \mathrm{grad} f(x),\, \mathrm{grad} f(x)\rangle \\
&= \left\langle \mathrm{grad} f(x),\, \nabla f(x) - \mathrm{D}h(x)^* \left(\mathrm{D}h(x)^*\right)^\dagger \nabla f(x) \right\rangle \\
&= \langle \mathrm{grad} f(x),\, \nabla f(x)\rangle\,,
\end{aligned}
$$

which verifies the second condition with $\rho = 1$. It also satisfies the third condition since the critical points are the points of $\mathcal{M}$ where $\mathrm{grad} f$ is zero. □

C.2. Proof of Lemma 2.5

Proof. It is assumed that $\Lambda(x) \neq 0$, otherwise the conclusion of Lemma 2.5 holds regardless of $\eta(x)$. Let $\tilde{\eta} = \inf\{\eta > 0 : \mathcal{N}(x - \eta\Lambda(x)) > \frac{\varepsilon^2}{2}\}$. If $\tilde{\eta} = \infty$, then the conclusion of Lemma 2.5 trivially holds; hence we now consider that $\tilde{\eta} < \infty$, i.e., $\tilde{\eta}$ is the first η beyond which $x - \eta\Lambda(x)$ is no longer in the safe region $\mathcal{M}^\varepsilon$. Let $\tilde{x} = x - \tilde{\eta}\Lambda(x)$, and observe that the line segment from x to $\tilde{x}$ is in $\mathcal{M}^\varepsilon$. Since $\mathcal{N}$ is $L_{\mathcal{N}}$-smooth in $\mathcal{M}^\varepsilon$ (Assumption 2.2), it follows from a standard bound (see,

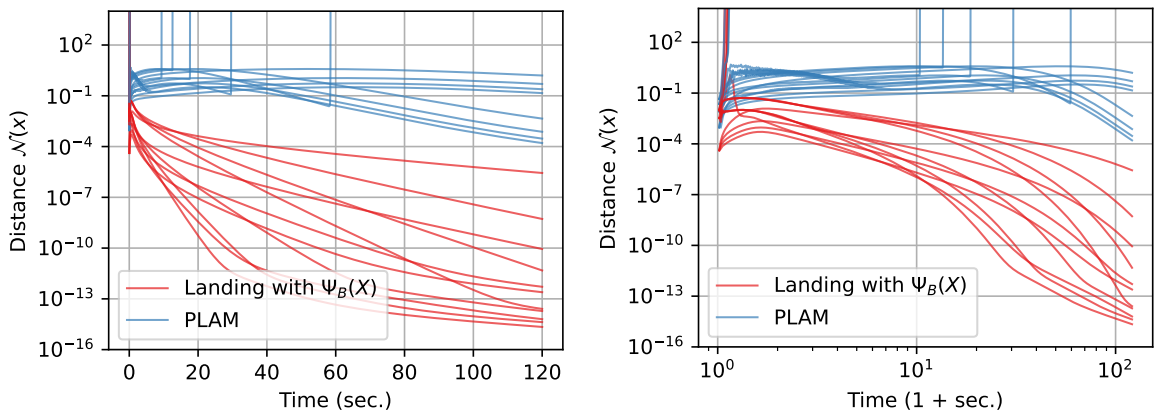


Figure 8. Comparison of the sensitivity to the choice of the step-size η and ω of the landing with $\Psi_B(X)$ and the PLAM method (Gao et al., 2022a) in the generalized eigenvalue problem experiment presented in Fig. 2 with $n = 1000, p = 500$, and the condition number of the two matrices $\kappa_B = \kappa_A = 100$. On the right we show log-log scale to better see the effect in earlier iterations. Both parameters are picked as in the experiment for Fig. 2 and multiplied by a scalar from the set $\{0.25, 0.75, 1.25, 1.75\}$ for all possible pair combinations.

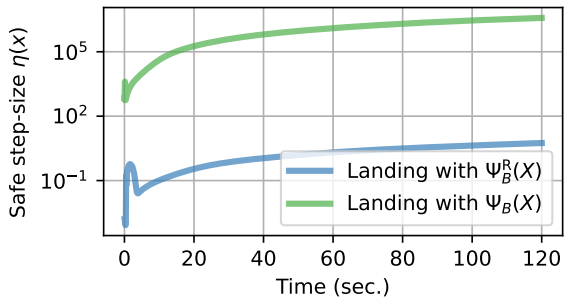
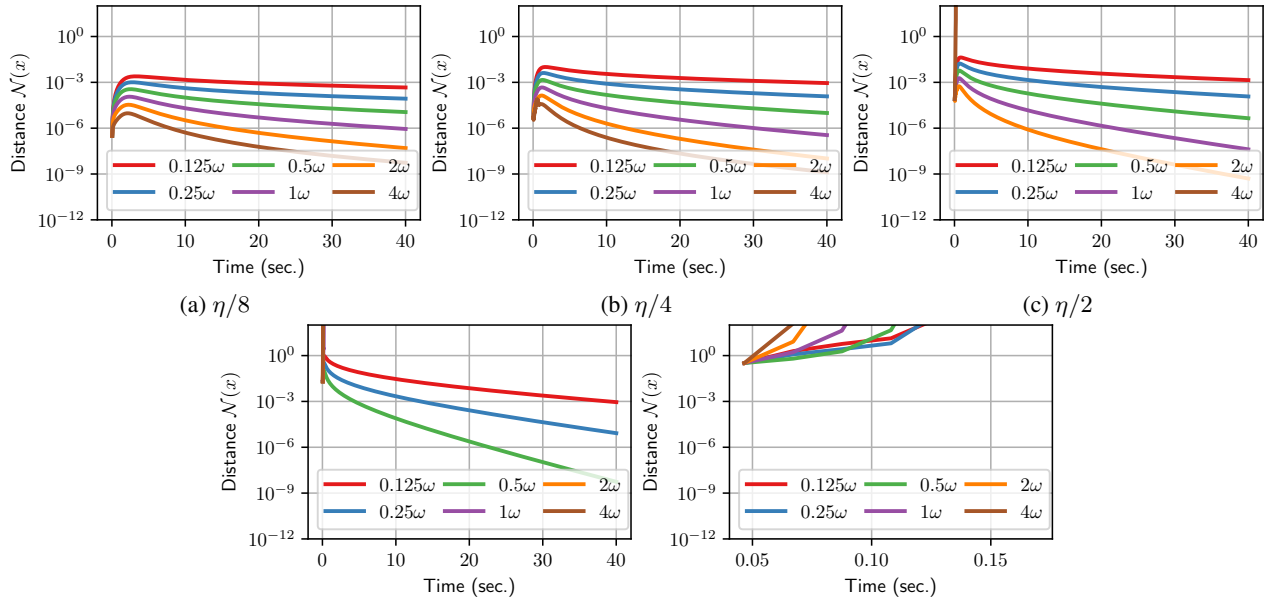


Figure 9. Numerical evaluation of the step-size safeguard $\eta(X)$ in Lemmma 2.5 per time, which ensures that the iterates stay in $\mathrm{St}_B^\varepsilon(p,n)$, for the two landing methods tested in Fig. 2 with the $L_\mathcal{N}$ bounded for the GEVP as in Lemma D.1.

e.g. Berger et al. (2020)) that

$$\begin{aligned}\frac{\varepsilon^2}{2} = \mathcal{N}(\tilde{x}) &\leq \mathcal{N}(x) + \langle \nabla\mathcal{N}(x),\, -\tilde{\eta}\Lambda(x)\rangle + \frac{\tilde{\eta}^2 L_\mathcal{N}}{2}\|\Lambda(x)\|^2 \\ &= \mathcal{N}(x) - \tilde{\eta}\omega\|\nabla\mathcal{N}(x)\|^2 + \frac{\tilde{\eta}^2 L_\mathcal{N}}{2}\|\Lambda(x)\|^2.\end{aligned}$$

The function $\bar{\mathcal{N}}(\eta) := \mathcal{N}(x) - \tilde{\eta}\omega\|\nabla\mathcal{N}(x)\|^2 + \frac{\tilde{\eta}^2 L_\mathcal{N}}{2}\|\Lambda(x)\|^2$ appearing on the right-hand side is a strictly convex quadratic function with $\bar{\mathcal{N}}(0) < \frac{\varepsilon^2}{2}$. Since $\tilde{\eta} \geq 0$, it follows that $\tilde{\eta} \geq \eta(x)$, where $\eta(x)$ is the positive solution of $\bar{\mathcal{N}}(\eta) = \frac{\varepsilon^2}{2}$, whose formula is the one given in the statement of Lemma 2.5. Hence $x - \eta(x)\Lambda(x)$ is in the line segment from x to $\tilde{x}$, which is included in $\mathcal{M}^\varepsilon$. □



(a) $\eta/8$ (b) $\eta/4$ (c) $\eta/2$

(d) 2η (e) 4η.

Figure 10. Robustness of the convergence towards the $\mathrm{St}_B(p,n)$ for the landing with $\Psi_B(X)$ in the experiment for Fig. 2 based on the multiplicative perturbations of η and ω parameters with the values from $\{1/8, 1/4, 1/2, 2, 4\}$.

C.3. Proof of Lemma 2.6

Proof. In view of Assumption 2.2, $\|\nabla\mathcal{N}(x)\| \geq \bar{C}_h\|h(x)\|$ holds in $\mathcal{M}^\varepsilon$. We proceed to lower bound the numerator of the step-size safeguard $\eta(x)$ in Lemma 2.5 as follows

$$\begin{aligned}
\omega\|\nabla\mathcal{N}(x)\|^2 &+ \sqrt{\omega^2\|\nabla\mathcal{N}(x)\|^4 + L_\mathcal{N}\|\Lambda(x)\|^2(\varepsilon^2 - \|h(x)\|^2)} \\
&\geq \omega\bar{C}_h^2\|h(x)\|^2 + \sqrt{\omega^2\bar{C}_h^4\|h(x)\|^4 + L_\mathcal{N}\|\Psi(x)\|^2\left(\varepsilon^2 - \|h(x)\|^2\right)} \\
&\geq \omega\bar{C}_h^2\|h(x)\|^2\left(1+\frac{1}{\sqrt{2}}\right) + \frac{1}{\sqrt{2}}\|\Psi(x)\|\sqrt{L_\mathcal{N}\left(\varepsilon^2 - \|h(x)\|^2\right)} \\
&\geq \sqrt{\frac{L_\mathcal{N}}{2}}\|\Psi(x)\|(\varepsilon - \|h(x)\|) + \left(1+\frac{1}{\sqrt{2}}\right)\omega\bar{C}_h^2\|h(x)\|^2
\end{aligned}$$

where the first inequality comes from using bounds from Assumption 2.2, the second inequality comes from $\sqrt{a+b} \geq (\sqrt{a}+\sqrt{b})/\sqrt{2}$ for $a, b \geq 0$, and the final inequality from the fact that $\sqrt{a-b} \geq \sqrt{a} - \sqrt{b}$ for $a, b \geq 0$ and $a \geq b$. As a result we have that $\eta(x)$ in Lemma 2.5 is lower-bounded by

$$\eta(x) \geq \frac{\sqrt{\frac{L_\mathcal{N}}{2}}\|\Psi(x)\|(\varepsilon - \|h(x)\|) + \left(1+\frac{1}{\sqrt{2}}\right)\omega\bar{C}_h^2\|h(x)\|^2}{L_\mathcal{N}\left(\|\Psi(x)\|^2 + \omega^2 C_h^2\|h(x)\|^2\right)}, \tag{11}$$

using the fact that $\|\Lambda(x)\|^2 = \|\Psi(x)\|^2 + \omega^2\|\nabla\mathcal{N}(x)\|^2$ and $\|\nabla\mathcal{N}(x)\|^2 \leq C_h^2\|h(x)\|^2$.

The right-hand side of (11) takes the form

$$\underline{\eta} := \frac{aP(\varepsilon - H) + bH^2}{cP^2 + dH^2},$$

where $a = \sqrt{\frac{L_\mathcal{N}}{2}}$, $b = \left(1+\frac{1}{\sqrt{2}}\right)\omega\bar{C}_h^2$, $c = L_\mathcal{N}$, and $d = L_\mathcal{N}\omega^2 C_h^2$ are constants and $P = \|\Psi(x)\|$, $H = \|h(x)\|$ are variables in bounded intervals: $0 \leq P \leq C_\Psi$ and $0 \leq H \leq \varepsilon$.

Pick $\alpha \in (0,1)$. There are two cases, either $H \geq \alpha\varepsilon$ or $H < \alpha\varepsilon$.

When $H \geq \alpha\varepsilon$, we have

$$\underline{\eta} \geq \frac{b\alpha^2\varepsilon^2}{cC_\Psi^2 + d\varepsilon^2}. \tag{12}$$

For the second case, when $H < \alpha\varepsilon$:

$$\underline{\eta} \geq \frac{aP(1-\alpha)\varepsilon + bH^2}{cP^2 + dH^2}. \tag{13}$$

We can again distinguish two cases to further lower bound (13). When $P \geq 1$, we have

$$(13) \geq \frac{a(1-\alpha)\varepsilon}{cC_\Psi^2 + d\varepsilon^2}.$$

When $P \leq 1$, we get

$$\begin{aligned}
(13) &\geq \frac{aP^2(1-\alpha)\varepsilon + bH^2}{cP^2 + dH^2} \\
&= \frac{\frac{a(1-\alpha)\varepsilon}{c}cP^2 + \frac{b}{d}dH^2}{cP^2 + dH^2} \\
&\geq \min\{\frac{a(1-\alpha)\varepsilon}{c}, \frac{b}{d}\}\frac{cP^2(1-\alpha)\varepsilon + dH^2}{cP^2 + dH^2} \\
&= \min\{\frac{a(1-\alpha)\varepsilon}{c}, \frac{b}{d}\}.
\end{aligned}$$

Putting all the cases together, we are left with the minimum of four terms

$$\begin{aligned}
\underline{\eta} &\geq \min\left\{\frac{b\alpha^2\varepsilon^2}{cC_\Psi^2 + d\varepsilon^2}, \frac{a(1-\alpha)\varepsilon}{cC_\Psi^2 + d\varepsilon^2}, \frac{a(1-\alpha)\varepsilon}{c}, \frac{b}{d}\right\} \\
&\geq \min\left\{\frac{\omega\bar{C}_h^2\alpha^2\varepsilon^2}{L_\mathcal{N}\left(C_\Psi^2 + \omega^2 C_h^2\varepsilon^2\right)}, \frac{(1-\alpha)\varepsilon}{\sqrt{2L_\mathcal{N}}\left(C_\Psi^2 + \omega^2 C_h^2\varepsilon^2\right)}, \frac{(1-\alpha)\varepsilon}{\sqrt{2L_\mathcal{N}}}, \frac{1}{\omega L_\mathcal{N}}\left(\frac{\bar{C}_h}{C_h}\right)^2\right\},
\end{aligned}$$

where in the second line we also further lower bounded the first and the last terms in the minimum by using that $1 \leq \left(1 + \frac{1}{\sqrt{2}}\right)$. □

C.4. Proof of Lemma 2.7

Proof. The inner product has two parts

$$\begin{aligned}
\langle \nabla\mathcal{L}(x),\, \Lambda(x)\rangle &= \mathrm{D}\mathcal{L}(x)[\Lambda(x)] \\
&= \mathrm{D}\mathcal{L}(x)[\Psi(x)] + \omega\mathrm{D}\mathcal{L}(x)[\nabla\mathcal{N}(x)].
\end{aligned} \tag{14}$$

We expand the first term of the right hand side of (14) as

$$\begin{aligned}
\mathrm{D}\mathcal{L}(x)[\Psi(x)] &= \langle \nabla f(x),\, \Psi(x)\rangle - \left\langle (\mathrm{D}h(x)^*)^\dagger \nabla f(x),\, \mathrm{D}h(x)\Psi(x)\right\rangle \\
&\quad - \langle \mathrm{D}\lambda(x)[\Psi(x)],\, h(x)\rangle + 2\beta\,\langle \nabla\mathcal{N}(x),\, \Psi(x)\rangle \\
&= \langle \nabla f(x),\, \Psi(x)\rangle - \langle \mathrm{D}\lambda(x)[\Psi(x)],\, h(x)\rangle
\end{aligned} \tag{15}$$

where we use that $\nabla\|h(x)\|^2 = 2\nabla\mathcal{N}(x)$ and that the second and the third term are zero due to the orthogonality of $\Psi(x)$ with the range of $\mathrm{D}h(x)^*$. We expand the second term of the right hand side of in (14) as

$$\begin{aligned}
\mathrm{D}\mathcal{L}(x)[\nabla\mathcal{N}(x)] &= \langle \nabla f(x),\, \nabla\mathcal{N}(x)\rangle - \left\langle (\mathrm{D}h(x)^*)^\dagger \nabla f(x),\, \mathrm{D}h(x)\nabla\mathcal{N}(x)\right\rangle \\
&\quad - \langle \mathrm{D}\lambda(x)[\nabla\mathcal{N}(x)],\, h(x)\rangle + 2\beta\|\nabla\mathcal{N}(x)\|^2 \\
&= \left\langle (I_n - \mathrm{D}h(x)^*(\mathrm{D}h(x)^*)^\dagger)\nabla f(x),\, \nabla\mathcal{N}(x)\right\rangle \\
&\quad - \langle \mathrm{D}\lambda(x)[\nabla\mathcal{N}(x)],\, h(x)\rangle + 2\beta\|\nabla\mathcal{N}(x)\|^2 \\
&= -\langle \mathrm{D}\lambda(x)[\nabla\mathcal{N}(x)],\, h(x)\rangle + 2\beta\|\nabla\mathcal{N}(x)\|^2,
\end{aligned} \tag{16}$$

where in the second equality we move the adjoint $\mathrm{D}h(x)^*$ in the second inner product to the left side and join it with the first inner product. The third equality comes from the fact that the projection of $\nabla f(x)$ on the null space of $\mathrm{D}h(x)$ and $\nabla\mathcal{N}(x) = \mathrm{D}h(x)^* h(x)$ are orthogonal.

Joining the two components (15) and (16) together we get

$$\begin{aligned}
\langle \nabla\mathcal{L}(x),\, \Lambda(x)\rangle &= \langle \nabla f(x),\, \Psi(x)\rangle - \langle \mathrm{D}\lambda(x)[\Lambda(x)],\, h(x)\rangle + 2\beta\omega\|\nabla\mathcal{N}(x)\|^2 \\
&\geq \rho\|\Psi(x)\|^2 - C_\lambda\left(\|\Psi(x)\| + \omega\|\nabla\mathcal{N}(x)\|\right)\|h(x)\| + 2\beta\omega\|\nabla\mathcal{N}(x)\|^2 \\
&\geq \rho\|\Psi(x)\|^2 + \omega(2\beta C_h - C_\lambda)C_h\|h(x)\|^2 - C_\lambda\|\Psi(x)\|\|h(x)\| \\
&\geq \rho\|\Psi(x)\|^2 + \omega(2\beta C_h - C_\lambda)C_h\|h(x)\|^2 - \frac{C_\lambda}{2}\left(\alpha\|\Psi(x)\|^2 + \alpha^{-1}\|h(x)\|^2\right) \\
&\geq \left(\rho - \frac{C_\lambda}{2}\alpha\right)\|\Psi(x)\|^2 + \left(2\omega\beta C_h^2 - \omega C_h C_\lambda - \alpha^{-1}\frac{C_\lambda}{2}\right)\|h(x)\|^2 \\
&\geq \frac{\rho}{2}\left(\|\Psi(x)\|^2 + \|h(x)\|^2\right)
\end{aligned}$$

where the first inequality comes from $\langle \nabla f(x),\, \Psi(x)\rangle \geq \rho\|\Psi(x)\|^2$ in Definition 2.1 combined with the bound $\sup_{x\in\mathcal{M}^\varepsilon}\|\mathrm{D}\lambda(x)\| \leq C_\lambda$ and the triangle inequality, the second inequality comes from bounding $\|\nabla\mathcal{N}(x)\| \leq C_h\|h(x)\|$ using Assumption 2.2 and rearranging terms, the third inequality comes from using the AG-inequality $\sqrt{ab} \leq (a+b)/2$ with $a = \alpha\|h(x)\|^2$ and $b = \alpha^{-1}\|\Psi(x)\|^2$ for an arbitrary $\alpha > 0$, in the fourth inequality we only rearrange terms, and finally, in the fifth inequality we choose $\alpha = \rho/C_\lambda$ and use that $\beta \geq (\frac{\rho}{4C_h^2} + \frac{\omega C_\lambda}{2C_h} + \frac{C_\lambda^2}{4\rho C_h^2})/\omega$. □

C.5. Proof of Theorem 2.8

Proof. Due to $x_0 \in \mathcal{M}^\varepsilon$ and the step-size η being smaller than the bound on the step-size safeguard in Lemma 2.6, we have that the segment $[x^k, x^{k+1}]$ is included in $\mathcal{M}^\varepsilon$ for all k. By $L_\mathcal{L}$-smoothness of Fletcher's augmented Lagrangian in $\mathcal{M}^\varepsilon$, we can expand

$$\mathcal{L}(x^{k+1}) \leq \mathcal{L}(x^k) - \eta\left\langle \Lambda(x^k),\, \nabla\mathcal{L}(x^k)\right\rangle + \frac{L_\mathcal{L}\eta^2}{2}\|\Lambda(x^k)\|^2 \tag{17}$$

$$\leq \mathcal{L}(x^k) - \frac{\eta\rho}{2}\left(\|\Psi(x^k)\|^2 + \omega^2\|h(x^k)\|^2\right) + \frac{L_\mathcal{L}\eta^2}{2}\|\Lambda(x^k)\|^2 \tag{18}$$

$$\leq \mathcal{L}(x^k) - \frac{\eta}{2}\Big(\left(\rho - L_\mathcal{L}\eta\right)\|\Psi(x^k)\|^2 + \omega^2\left(\rho - \eta L_\mathcal{L} C_h^2\right)\|h(x^k)\|^2\Big), \tag{19}$$

where in the second inequality we used the results of Lemma 2.7, and in the third inequality we use the bound on $\|\nabla\mathcal{N}(x)\| \leq C_h\|h(x)\|$ by Assumption 2.2. By the step-size $\eta < \min\left\{\frac{\rho}{2L_\mathcal{L}}, \frac{\rho}{2L_\mathcal{L}C_h^2}\right\}$ we have

$$\frac{\eta\rho}{4}\|\Psi(x^k)\|^2 + \frac{\eta\rho\omega^2}{4}\|h(x)\|^2 \leq \mathcal{L}(x^k) - \mathcal{L}(x^{k+1}). \tag{20}$$

Telescopically summing the first $K+1$ terms gives

$$\frac{\eta\rho}{4}\sum_{k=0}^{K}\|\Psi(x^k)\|^2 + \frac{\eta\rho\omega^2}{4}\sum_{k=0}^{K}\|h(x)\|^2 \leq \mathcal{L}(x^0) - \mathcal{L}(x^{K+1}) \leq \mathcal{L}(x^0) - \mathcal{L}^*,$$

which implies that the inequalities hold individually also

$$\frac{\eta\rho}{4}\sum_{k=0}^{K}\|\Psi(x^k)\|^2 \leq \mathcal{L}(x^0) - \mathcal{L}^* \qquad \text{and} \qquad \frac{\eta\rho\omega^2}{4}\sum_{k=0}^{K}\|h(x)\|^2 \leq \mathcal{L}(x^0) - \mathcal{L}^*.$$

□

C.6. Proof of Theorem 2.9

Proof. Let $x^{k+1} = x^k - \eta_k \tilde{\Lambda}(x^k)$, where we denote by $\tilde{\Lambda}(x^k) = \Lambda(x^k) + \tilde{E}(x^k, \Xi^k)$ the unbiased estimator of the landing update, and we assume that the line segment between the iterates remain within $\mathcal{M}^\varepsilon$. By $L_\mathcal{L}$-smoothness of Fletcher's augmented Lagrangian inside $\mathcal{M}^\varepsilon$, we have

$$\begin{aligned}
\mathbb{E}_{\Xi^k}\left[\mathcal{L}(x^{k+1})\right] &\leq \mathbb{E}_{\Xi^k}\left[\mathcal{L}(x^k) - \eta_k \left\langle \tilde{\Lambda}(x^k), \nabla\mathcal{L}(x^k)\right\rangle + \frac{L_\mathcal{L}\eta_k^2}{2}\|\tilde{\Lambda}(x^k)\|^2\right] \\
&\leq \mathcal{L}(x^k) - \eta_k \left\langle \Lambda(x^k), \nabla\mathcal{L}(x^k)\right\rangle + \frac{L_\mathcal{L}\eta_k^2}{2}\left(\|\Lambda(x^k)\|^2 + \gamma^2\right) \\
&\leq \mathcal{L}(x^k) - \frac{\eta_k\rho}{2}\left(\|\Psi(x^k)\|^2 + \omega^2\|h(x)\|^2\right) + \frac{L_\mathcal{L}\eta_k^2}{2}\left(\|\Lambda(x^k)\|^2 + \gamma^2\right) \\
&\leq \mathcal{L}(x^k) + \frac{L_\mathcal{L}\eta_k^2}{2}\gamma^2 - \frac{\eta_k}{2}\left((\rho - L_\mathcal{L}\eta_k)\|\Psi(x^k)\|^2 + \omega^2\left(\rho - \eta_k L_\mathcal{L} C_h^2\right)\|h(x^k)\|^2\right),
\end{aligned}$$

where the first inequality comes from taking an expectation of a bound akin the first bound of subsection C.5, in the second inequality we take the expectation inside the inner product using the fact that $\tilde{E}(x^k, \Xi^k)$ is zero-centered and has bounded variance, the third inequality comes as a consequence of Lemma 2.7. The last inequality comes as a consequence of $\Lambda(x^k)$ having two orthogonal components and rearranging terms in the same way as in (19). Note that by $\mathbb{E}_{\Xi^k}$ we denote expectation only with respect to the last random realization Ξ^k.

By the step-size being smaller than $\eta_k \leq \eta_0 < \min\left\{\frac{\rho}{2L_\mathcal{L}}, \frac{\rho}{2L_\mathcal{L}C_h^2}\right\}$ we have that

$$\frac{\eta_k\rho}{4}\|\Psi(x^k)\|^2 + \frac{\eta_k\rho\omega^2}{4}\|h(x^k)\|^2 \leq \mathcal{L}(x^k) - \mathbb{E}_{\Xi^k}\left[\mathcal{L}(x^{k+1})\right] + \frac{L_\mathcal{L}\eta_k^2}{2}\gamma^2. \tag{21}$$

Taking the expectation of (21) with respect to the whole past random realizations $\Xi^0, \ldots, \Xi^k$, denoted for short simply as $\mathbb{E}$, yields

$$\mathbb{E}\left[\frac{\eta_k\rho}{4}\|\Psi(x^k)\|^2 + \frac{\eta_k\rho\omega^2}{4}\|h(x^k)\|^2\right] \leq \mathbb{E}[\mathcal{L}(x^k)] - \mathbb{E}\left[\mathbb{E}_{\Xi^k}\left[\mathcal{L}(x^{k+1})\right]\right] + \frac{L_\mathcal{L}\eta_k^2}{2}\gamma^2. \tag{22}$$

Since $x^{k+1} = x^k - \eta_k\tilde{\Lambda}(x^k)$, we have that $\mathbb{E}\left[\mathbb{E}_{\Xi^k}[\cdot]\right] = \mathbb{E}[\cdot]$, and we can telescopically sum the first $K+1$ terms of (21) for $k = 0, 1, \ldots, K$:

$$\begin{aligned}
\frac{\rho}{4}\left(\sum_{k=0}^{K}\eta_k\mathbb{E}\left[\|\Psi(x^k)\|^2\right] + \omega^2\sum_{k=0}^{K}\eta_k\mathbb{E}\left[\|h(x^k)\|^2\right]\right) &\leq \mathcal{L}(x^0) - \mathbb{E}\left[\mathcal{L}(x^{K+1})\right] + \frac{L_\mathcal{L}\eta_0^2\gamma^2}{2}\sum_{k=0}^{K}(1+k)^{-1} \\
&\leq \mathcal{L}(x^0) - \mathcal{L}^* + \frac{L_\mathcal{L}\eta_0^2\gamma^2}{2}\left(1 + \log(K+1)\right)
\end{aligned} \tag{23}$$

which implies that the inequalities hold also individually

$$\begin{aligned}
\inf_{k\leq K}\mathbb{E}\left[\|\Psi(x^k)\|^2\right] &\leq 4\frac{\mathcal{L}(x^0) - \mathcal{L}^*}{\rho\eta_0\sqrt{K}} + 2\frac{\eta_0 L_\mathcal{L}\gamma^2}{\rho}\left(\frac{1+\log(K+1)}{\sqrt{K}}\right), \\
\inf_{k\leq K}\mathbb{E}\left[\|h(x^k)\|^2\right] &\leq 4\frac{\mathcal{L}(x^0) - \mathcal{L}^*}{\rho\omega^2\eta_0\sqrt{K}} + 2\frac{\eta_0 L_\mathcal{L}\gamma^2}{\rho\omega^2}\left(\frac{1+\log(K+1)}{\sqrt{K}}\right),
\end{aligned}$$

where we used that $\inf_{k\leq K}\mathbb{E}\|\Psi(x^k)\|^2 \leq \sum_{k=0}^{K}\eta_k\mathbb{E}\|\Psi(x^k)\|^2\left(\sum_{k=0}^{K}\eta_k\right)^{-1}$ and the fact that $\sum_{k\leq K}\eta_k \geq \eta_0\sqrt{K}$. □

D. Proofs for Section 3

D.1. Specific forms of $\mathrm{D}h(x), \lambda(X)$ for $\mathrm{St}_B(p,n)$

We begin by showing the specific form of the formulations derived in the previous section for the case of the generalized Stiefel manifold. Let $h : \mathbb{R}^{n\times p} \to \mathrm{sym}(p) : X \mapsto X^\top BX - I_p$, where letting $\mathrm{sym}(p)$ be the codomain is essential for

Assumption 2.2 to hold. Differentiating the generalized Stiefel constraint yields $\mathrm{D}h(X)[V] = X^\top BV + V^\top BX$ and the adjoint $\mathrm{D}h(X)^* : \mathrm{sym}(p) \to \mathbb{R}^{n\times p}$ is derived as

$$\langle \mathrm{D}h(X)^*[V],\, W\rangle = \langle V,\, \mathrm{D}h(X)[W]\rangle = \left\langle V,\, W^TBX + X^TBW\right\rangle = \langle 2BXV,\, W\rangle\,, \tag{24}$$

as such we have that $\mathrm{D}h(X)^*[V] = 2BXV$. Consequently

$$\mathrm{D}h(X)\mathrm{D}h(X)^*[V] = 2VX^\top B^2X + 2X^\top B^2XV, \tag{25}$$

and the Lagrange multiplier $\lambda(X)$ is defined in the case of the generalized Stiefel manifold as the solution to the following Lyapunov equation

$$2\lambda(X)X^\top B^2X + 2X^\top B^2X\lambda(X) = X^\top B\nabla f(X) + \nabla f(X)^\top BX. \tag{26}$$

Importantly, due to $\lambda(X)$ being the unique solution to the linear equation, which is ensured by BX having a full rank since X is in the ε-safe region, and by the the linear operator being smooth in X, since $\nabla f(X)$ is smooth, we have that $\lambda(X)$ is invertible and smooth with respect to X. Thus, as a smooth function defined over a compact set $\mathrm{St}_B^\varepsilon(p,n)$, its operator norm is bounded: $\sup_{X\in\mathrm{St}_B^\varepsilon(p,n)} \|\mathrm{D}\lambda(X)\|_F \le C_\lambda$ as required by Assumption 2.3.

D.2. Proof of Proposition 3.1

Proof. For $\|X^\top BX - I_p\|_F \le \varepsilon$, let $X = U\Sigma V^\top$ be a truncated singular value decomposition of X, and $QDQ^\top$ be an eigendecomposition of B. We then have

$$\varepsilon^2 \ge \|X^\top BX - I_p\|_{\mathrm{F}}^2 = \|\Sigma U^\top QD(U^\top Q)^\top\Sigma - I_p\|_{\mathrm{F}}^2 \tag{27}$$

where β_i, σ_i are the positive eigenvalues of B and the singular values of X respectively in decreasing order.

Denote $P = Q^\top U \in \mathbb{R}^{n\times p}$ that forms an orthogonal frame $P^\top P = I_p$. The bound in (27) implies

$$\varepsilon^2 \ge \sum_{i=1}^{p} \left(\sigma_i^2 \left(P^\top DP\right)_{ii} - 1\right)^2, \tag{28}$$

where $\left(P^\top DP\right)_{ii}$ marks the ith diagonal entry of the matrix $P^\top DP$. Consequently, we have that

$$1 - \varepsilon \le \sigma_i^2 \left(P^\top DP\right)_{ii} \le 1 + \varepsilon \tag{29}$$

for all $i = 1, \ldots, p$. We can bound

$$\beta_n = \inf_{\|x\|_2=1} x^\top Dx \le \left(P^\top DP\right)_{ii} \le \sup_{\|x\|_2=1} x^\top Dx = \beta_1, \tag{30}$$

since $P^\top P = I_p$. The inequality in (29) combined with (30) implies that

$$\sqrt{(1-\varepsilon)/\beta_1} \le \sigma_i \le \sqrt{(1+\varepsilon)/\beta_n}. \tag{31}$$

By the lower and the upper bounds on singular values of a matrix product, the above bound gives that the singular values of $\mathrm{D}h(X)^* = 2BX$ are in the interval $[2\sqrt{(1-\varepsilon)\beta_n\kappa_B^{-1}}, 2\sqrt{(1+\varepsilon)\beta_1\kappa_B}]$ which in turn gives the constants $C_h, \bar{C}_h$. □

D.3. Proof of Proposition 3.2

Proof. First consider $\Psi_B(X)$. For ease of notation we denote $G = \nabla f(X) \in \mathbb{R}^{n\times p}$. The first property Definition 2.1 *(i)* comes from

$$\left\langle \mathrm{skew}(GX^\top B)BX,\, BXS\right\rangle = 0, \tag{32}$$

which holds for a symmetric matrix S, since a skew-symmetric matrix is orthogonal in the Frobenius inner product to a symmetric matrix,

The second property *(ii)* is a consequence of the following

$$\langle \Psi_B(X),\, G\rangle = \left\langle \mathrm{skew}(GX^TB)BX,\, G\right\rangle = \|\mathrm{skew}(GX^TB)\|_{\mathrm{F}}^2 \ge \frac{1}{(1+\varepsilon)\beta_1\kappa_B}\|\Psi_B(X)\|_{\mathrm{F}}^2, \tag{33}$$

where we use the bounds on $\|BX\|_2 \leq \sqrt{(1+\varepsilon)\beta_1\kappa_B}$ derived in the proof of Proposition 3.1.

To show the third property *(iii)*, we first consider a critical point $X \in \mathrm{St}_B(p,n)$, for which it must hold that G is in the image of $\mathrm{D}h(X)^*$, i.e.,

$$G = BXS, \tag{34}$$

for some $S \in \mathrm{sym}(p)$ and that $X^\top BX = I_p$ by feasibility. We have that at the critical point defined in (34), the relative ascent direction is

$$\Psi_B(X) = \mathrm{skew}(GX^\top B)BX = \mathrm{skew}(BXSX^\top B)BX = 0, \tag{35}$$

where the second equality is the consequence of (34) and the third equality comes from the fact that $BXSX^\top B$ is symmetric.

To show the other side of the implication, that $\Psi_B(X) = 0$ combined with feasibility imply that X is a critical point, we consider

$$0 = \Psi_B(X) = \mathrm{skew}(GX^\top B)BX = GX^\top B^2 X - BXG^\top BX \tag{36}$$

which, since $X^\top B^2 X \in \mathbb{R}^{p\times p}$ is invertible, is equivalent to

$$G = BXG^\top BX\left(X^\top B^2 X\right)^{-1}. \tag{37}$$

It remains to show that the factor $G^\top BX\left(X^\top B^2 X\right)^{-1}$ in (37) is symmetric in order to get (34). To this end, multiply (36) on the left by $(X^\top B^2 X)^{-1}X^\top B$ and on the right by $(X^\top B^2 X)^{-1}$ and rearrange the terms.

For the other choice of relative gradient $\Psi_B^{\mathrm{R}}(X) = \mathrm{skew}(B^{-1}GX^\top)BX$, letting $M = B^{-1}GX^\top$, we find

$$\begin{aligned}
\langle \Psi_B^{\mathrm{R}}(X), G\rangle &= \langle \mathrm{skew}(M), BMB\rangle && (38)\\
&= \langle \mathrm{skew}(M), \mathrm{skew}(BMB)\rangle && (39)\\
&= \langle \mathrm{skew}(M), B\mathrm{skew}(M)B\rangle && (40)\\
&\geq \|\mathrm{skew}(M)\|_{\mathrm{F}}^2\beta_n^2 && (41)
\end{aligned}$$

and similarly as in (33), it holds $\|\Psi_B^{\mathrm{R}}(X)\|^2 \leq \|\mathrm{skew}(M)\|_{\mathrm{F}}^2(1+\varepsilon)\beta_1\kappa_B$ which in turn leads to $\langle \Psi_B^{\mathrm{R}}(X), G\rangle \geq \frac{\beta_n}{1+\varepsilon}\|\Psi_B^{\mathrm{R}}(X)\|^2$ □

D.4. Lipschitz constants for the GEVP

Lemma D.1 (Lipschitz constants for the generalized eigenvalue problem)**.** *Let $f = -\frac{1}{2}\mathrm{Tr}(X^\top AX)$ and $\mathcal{N}(X) = \frac{1}{2}\|X^\top BX - I_p\|_F^2$ as in the optimization problem corresponding to the generalized eigenvalue problem. We have that, for $X \in \mathrm{St}_B^\varepsilon(p,n)$, a Lipschitz constant for $\nabla\mathcal{N}$ is $L_{\mathcal{N}} = 2\beta_1\left(\varepsilon + 2(1+\varepsilon)\kappa_B\right)$ and the Lipschitz constant for ∇f is $L_f = \alpha_1$ where α_1 is the largest eigenvalue of A.*

Proof. Take $X, Y \in \mathrm{St}_B(p,n)$, we have that $\nabla\mathcal{N}(X) = 2BX(X^\top BX - I_p)$, thus

$$\begin{aligned}
\nabla\mathcal{N}(X) - \nabla\mathcal{N}(Y) &= 2B\left(X(X^\top BX - I_p) - Y(Y^\top BY - I_p)\right) && (42)\\
&= 2B\left((X-Y)(X^\top BX - I_p) + Y\left((X^\top BX - Y^\top BY)\right)\right) && (43)\\
&= 2B\left((X-Y)(X^\top BX - I_p) + Y\left((X-Y)^\top BX + Y^\top B(X-Y)\right)\right). && (44)
\end{aligned}$$

Taking the Frobenius norm and by the triangle inequality we get

$$\begin{aligned}
\|\nabla\mathcal{N}(X) - \nabla\mathcal{N}(Y)\| &\leq 2\left(\|B(X-Y)(X^\top BX - I_p)\| + \|BY(X-Y)^\top BX\| + \|BYY^\top B(X-Y)\|\right) && (45)\\
&\leq 2\left(\|X-Y\|\left\|B(X^\top BX - I_p)\right\|_2 + \|X-Y\|\,\|BYBX\|_2 + \|X-Y\|\left\|BYY^\top B\right\|_2\right) && (46)\\
&\leq 2\|X-Y\|\left(\|B\|_2\|X^\top BX - I_p\| + \|B\|_2^2\|X\|_2\|Y\|_2 + \|B\|_2^2\|Y\|_2^2\right) && (47)\\
&\leq 2\beta_1\left(\varepsilon + 2(1+\varepsilon)\kappa_B\right)\|X-Y\|, && (48)
\end{aligned}$$

where for the second inequality we used that $\|AB\| \leq \|A\|_2\|B\|$, the third inequality comes from submultiplicativity of the induced ℓ_2-norm for matrices, and the fourth inequality comes from $X, Y \in \mathrm{St}_B^\varepsilon(p,n)$ for which we have that $\|X\|_2 \leq \sqrt{(1+\varepsilon)/\beta_n}$, as in (31), and the same for Y.

When $f = \frac{1}{2}\mathrm{Tr}(X^\top AX)$, we have that $\|\nabla f(X) - \nabla f(Y)\| \leq \|A\|_2\|X-Y\|$. □

D.5. Proof of Proposition 3.3

Proof. We start by deriving the bound on the variance of the normalizing component $\nabla\mathcal{N}(X)$. Consider U and V to be two independent random matrices taking i.i.d. values from the distribution of B_ζ with variance σ_B^2. We have that

$$\mathrm{Var}\left[UX(X^\top VX - I_p)\right] = \mathbb{E}_{U,V}\left[\|UX(X^\top VX - I_p) - BX(X^\top BX - I_p)\|^2\right]. \tag{49}$$

Introducing the random marginal $BX(X^\top VX - I_p)$, we further decompose

$$\begin{aligned}\mathrm{Var}\left[UX(X^\top VX - I_p)\right] &= \mathbb{E}_{U,V}\left[\|UX(X^\top VX - I_p) - BX(X^\top VX - I_p)\|^2\right] &(50)\\ &\quad + \mathbb{E}_V\left[\|BX(X^\top VX - I_p) - BX(X^\top BX - I_p)\|^2\right]. &(51)\end{aligned}$$

The first term in the above is upper bounded as

$$\begin{aligned}\mathbb{E}_{U,V}\left[\|UX(X^\top VX - I_p) - BX(X^\top VX - I_p)\|^2\right] &\le \mathbb{E}_{U,V}\left[\|U - B\|^2\|X(X^\top VX - I_p)\|_2^2\right] &(52)\\ &= \sigma_B^2\mathbb{E}_V[\|X(X^\top VX - I_p)\|_2^2] &(53)\\ &\le \sigma_B^2\frac{1+\varepsilon}{\beta_n}\mathbb{E}_V[\|X^\top VX - I_p\|_2^2] &(54)\\ &\le \sigma_B^2\frac{1+\varepsilon}{\beta_n}\left(\sigma_B^2\frac{1+\varepsilon}{\beta_n} + \varepsilon^2\right), &(55)\end{aligned}$$

where we used $\|X\|^2 \le \frac{1+\varepsilon}{\beta_n}$, and we control $\mathbb{E}_V[\|X^\top VX - I_p\|_2^2] \le \mathbb{E}_V[\|X^\top VX - I_p\|^2] = \mathbb{E}_V[\|X^\top(V-B)X\|^2] + \|XBX^\top - I_p\|^2 \le \sigma_B^2\frac{1+\varepsilon}{\beta_n} + \varepsilon^2$. The second term is controlled by

$$\begin{aligned}\mathbb{E}_V\left[\|BX(X^\top VX - I_p) - BX(X^\top BX - I_p)\|^2\right] &= \mathbb{E}_V\left[\|BXX^\top(V-B)X\|^2\right] &(56)\\ &\le \sigma_B^2\|B\|_2^2\|X\|_2^6 &(57)\\ &\le \sigma_B^2\beta_1^2\frac{(1+\varepsilon)^3}{\beta_n^3}, &(58)\end{aligned}$$

where we used $\|X\|_2^2 \le \frac{1+\varepsilon}{\beta_n}$ and $\|B\|_2 = \beta_1$. Taking things together we obtain

$$\mathrm{Var}\left[UX(X^\top VX - I_p)\right] \le \sigma_B^2\left(\frac{1+\varepsilon}{\beta_n}\left(\sigma_B^2\frac{1+\varepsilon}{\beta_n} + \varepsilon^2\right) + \beta_1^2\frac{(1+\varepsilon)^3}{\beta_n^3}\right). \tag{59}$$

Similarly, the variance of the first term in the landing is controlled by introducing yet another random variable G that takes values from $\nabla f_\xi(X)$. We use the U-statistics variance decomposition twice to get

$$\begin{aligned}\mathrm{Var}[\mathrm{skew}\left(GX^\top U\right)VX] &= \mathbb{E}_{G,U,V}[\|\mathrm{skew}((G - \nabla f(X))X^\top U)VX\|^2]\\ &\quad + \mathbb{E}_{U,V}[\|\mathrm{skew}(\nabla f(X)X^\top(U-B))VX\|^2]\\ &\quad + \mathbb{E}_V[\|\mathrm{skew}(\nabla f(X)X^\top B)(V-B)X\|^2].\end{aligned}$$

The first term is upper bounded by doing

$$\begin{aligned}\mathbb{E}_{G,U,V}[\|\mathrm{skew}((G - \nabla f(X))X^\top U)VX\|^2] &\le \mathbb{E}_{G,U,V}[\|G - \nabla f(X)\|^2\|X^\top U\|_2^2\|VX\|_2^2] &(60)\\ &\le \sigma_G^2\mathbb{E}_U[\|U\|_2^2]^2\|X\|_2^4 &(61)\\ &\le \sigma_G^2 p_B^2\frac{(1+\varepsilon)^2}{\beta_n^2}, &(62)\end{aligned}$$

where we used $p_B = \mathbb{E}_U[\|U\|_2^2] = \mathbb{E}_{B_\zeta}[\|B_\zeta\|_2^2]$. The second term gives

$$\begin{aligned}\mathbb{E}_{U,V}[\|\mathrm{skew}(\nabla f(X)X^\top(U-B))VX\|^2] &\le \mathbb{E}_{U,V}[\|\nabla f(X)X^\top\|_2^2\|U-B\|^2\|VX\|_2^2] &(63)\\ &\le \sigma_B^2\|\nabla f(X)X^\top\|_2^2\mathbb{E}_U[\|U\|^2]\|X\|_2^2 &(64)\\ &\le \sigma_B^2\Delta p_B\frac{1+\varepsilon}{\beta_n}, &(65)\end{aligned}$$

where Δ upper-bounds $\|\nabla f(X)X^\top\|_2^2$. The third term gives

$$\mathbb{E}_V[\|\text{skew}(\nabla f(X)X^\top B)(V-B)X\|^2] \leq \mathbb{E}_V[\|\nabla f(X)X^\top\|_2^2\|B\|_2^2\|V-B\|^2\|X\|_2^2] \tag{66}$$
$$\leq \sigma_B^2\|\nabla f(X)X^\top\|_2^2\|B\|_2^2\|X\|_2^2 \tag{67}$$
$$\leq \sigma_B^2\Delta\beta_1^2\frac{1+\varepsilon}{\beta_n}, \tag{68}$$

which leads to the bound

$$\text{Var}[\text{skew}\left(GX^\top U\right)VX] \leq \sigma_G^2 p_B^2\frac{(1+\varepsilon)^2}{\beta_n^2} + \sigma_B^2\frac{1+\varepsilon}{\beta_n}\Delta\left(p_B+\beta_1^2\right).$$

Finally, we join these two bounds using the generic inequality $\text{Var}[a+b] \leq 2(\text{Var}[a]+\text{Var}[b])$, which gives

$$\mathbb{E}_\Xi[\|\tilde{E}(X,\Xi)\|^2] = \text{Var}[2\text{skew}\left(GX^\top U\right)VX + 2\omega VX(X^\top UX - I_p)] \tag{69}$$
$$\leq 8(\text{Var}[\text{skew}\left(GX^\top U\right)VX] + \omega^2\text{Var}[VX(X^\top UX - I_p)]) \tag{70}$$
$$\leq 8\left(\sigma_G^2 p_B^2\frac{(1+\varepsilon)^2}{\beta_n^2} + \sigma_B^2\frac{1+\varepsilon}{\beta_n}\Delta\left(p_B+\beta_1^2\right) + \omega^2\sigma_B^2\left(\frac{1+\varepsilon}{\beta_n}(\sigma_B^2\frac{1+\varepsilon}{\beta_n}+\varepsilon^2) + \beta_1^2\frac{(1+\varepsilon)^3}{\beta_n^3}\right)\right) \tag{71}$$
$$= 8\sigma_G^2 p_B^2\frac{(1+\varepsilon)^2}{\beta_n^2} + 8\sigma_B^2\frac{1+\varepsilon}{\beta_n}\left(\Delta\left(p_B+\beta_1^2\right) + \omega^2\left(\sigma_B^2\frac{1+\varepsilon}{\beta_n} + \varepsilon^2 + \beta_1^2\frac{(1+\varepsilon)^2}{\beta_n^2}\right)\right) \tag{72}$$
$$= \sigma_G^2\alpha_G + \sigma_B^2(\alpha_B + \omega^2\gamma_B), \tag{73}$$

with

$$\alpha_G = 8p_B^2\frac{(1+\varepsilon)^2}{\beta_n^2} \tag{74}$$
$$\alpha_B = 8\frac{1+\varepsilon}{\beta_n}\Delta\left(p_B+\beta_1^2\right) \tag{75}$$
$$\gamma_B = 8\frac{1+\varepsilon}{\beta_n}\left(\frac{1+\varepsilon}{\beta_n}\sigma_B^2 + \varepsilon^2 + \beta_1^2\frac{(1+\varepsilon)^3}{\beta_n^3}\right). \tag{76}$$

□

E. Riemannian Interpretation of $\Psi_B^{\mathrm{R}}(X)$ in Proposition 3.2

Similar to the work of Gao et al. (2022b), we provide a geometric interpretation of the relative ascent direction $\Psi_B^{\mathrm{R}}(X)$ as a Riemannian gradient in a metric induced by an isometry

$$\Phi_{B,M} : \mathrm{St}(p,n) \to \mathrm{St}_{B,M}(p,n) : Y \mapsto B^{-\frac{1}{2}}YM^{\frac{1}{2}}$$

between the standard Stiefel manifold $\mathrm{St}(p,n)$ and the doubly generalized Stiefel manifold

$$\mathrm{St}_{B,M}(p,n) := \{X : X^\top BX = M\},$$

for $B, M \succ 0$, which is a layered manifold (Goyens et al., 2024) of $h(X) := X^\top BX$.

The map $\Phi_{B,M}$ extends to a diffeomorphism of the set of the full rank $\mathbb{R}^{n\times p}$ matrices onto itself and maps the standard Stiefel manifold $\mathrm{St}(p,n)$ to the generalized Stiefel manifold $\mathrm{St}_{B,M}(p,n)$. The tangent space at $X \in \mathrm{St}_{B,M}(p,n)$ is the null space of $\mathrm{D}h(X)$:

$$\begin{aligned}\mathrm{T}_X\mathrm{St}_{B,M}(p,n) &= \{\xi \in \mathbb{R}^{n\times p} : \xi^T BX + X^T B\xi = 0\}\\ &= \{X(X^TBX)^{-1}\Omega + B^{-1}X_\perp K : \Omega^T + \Omega = 0, \Omega \in \mathbb{R}^{p\times p}, K \in \mathbb{R}^{(n-p)\times p}\}\\ &= \{WBX : W^T + W = 0, W \in \mathbb{R}^{n\times n}\}\\ &= \{\Phi_{B,M}(\zeta) : \zeta \in \mathrm{T}_{\Phi_{B,M}^{-1}(X)}\mathrm{St}(p,n)\},\end{aligned}$$

where $X_\perp \in \mathbb{R}^{n\times(n-p)}$ is any matrix such that $\mathrm{span}(X_\perp)$ is the orthogonal complement of $\mathrm{span}(X)$.

Consider the canonical metric on the standard Stiefel manifold $\mathrm{St}(p,n)$:

$$g_Y^{\mathrm{St}(p,n)}(Z_1,Z_2) = \left\langle Z_1,\, (I - \frac{1}{2}YY^T)Z_2 \right\rangle.$$

It turns out that the Riemannian gradient of $\tilde{f} : \mathrm{St}(p,n) \to \mathbb{R}$ is

$$\mathrm{grad}\tilde{f}(Y) = 2\,\mathrm{skew}\left(\nabla\tilde{f}(Y)Y^\top\right)Y.$$

Using the map $\Phi_{B,M}$, we define the metric $g^{\mathrm{St}_{B,M}(p,n)}$ which makes $\Phi_{B,M}$ an isometry. This metric is given by

$$\begin{aligned} g_X^{\mathrm{St}_{B,M}(p,n)}(\xi,\zeta) &= g^{\mathrm{St}(p,n)}_{\Phi^{-1}_{B,M}(X)}(\Phi^{-1}_{B,M}(\xi),\Phi^{-1}_{B,M}(\zeta)) \\ &= \left\langle \xi,\, (B - \frac{1}{2}BX(X^TBX)^{-1}X^TB)\zeta(X^TBX)^{-1} \right\rangle. \end{aligned}$$

This metric extends to arguments ξ and ζ in $\mathrm{T}_X\mathbb{R}^{n\times p} \simeq \mathbb{R}^{n\times p}$ using the same formula. With respect to this metric, the normal space of $\mathrm{St}_{B,M}(p,n)$ is

$$\mathrm{N}_X\mathrm{St}_{B,M}(p,n) = \{X(X^TBX)^{-1}S : S^T = S, S \in \mathbb{R}^{p\times p}\}.$$

The form of the derived tangent and normal spaces allow us to derive their projection operators P_X and $P_X^\perp$ respectively as

$$\begin{aligned} P_X^\perp(Y) &= X(X^TBX)^{-1}\mathrm{sym}(X^TBY), \\ P_X(Y) &= Y - X(X^TBX)^{-1}\mathrm{sym}(X^TBY). \end{aligned}$$

Since $\Phi_{B,M}$ is a linear isometric, and letting $\Phi^*_{B,M}$ denote the adjoint of $\Phi_{B,M}$ with respect to the Frobenius inner product, the Riemannian gradient w.r.t. $g^{\mathrm{St}_{B,M}(p,n)}$ can be computed directly by

$$\begin{aligned} \mathrm{grad}_{B,M}f(X) &= \Phi_{B,M}\left(\mathrm{grad}(f\circ\Phi_{B,M})(\Phi^{-1}_{B,M}(X))\right) \\ &= \Phi_{B,M}\left(2\,\mathrm{skew}\left(\nabla(f\circ\Phi_{B,M})(\Phi^{-1}_{B,M}(X))\right)(\Phi^{-1}_{B,M}(X))^\top\right) \\ &= \Phi_{B,M}\left(2\,\mathrm{skew}\left(\Phi^*_{B,M}\nabla f(X)(\Phi^{-1}_{B,M}(X))^\top\right)(\Phi^{-1}_{B,M}(X))^\top\right) \\ &= 2B^{-\frac{1}{2}}\mathrm{skew}\left(B^{-\frac{1}{2}}\nabla f(X)M^{\frac{1}{2}}(B^{\frac{1}{2}}XM^{-\frac{1}{2}})^\top\right)B^{\frac{1}{2}}XM^{-\frac{1}{2}}M^{\frac{1}{2}} \\ &= 2\,\mathrm{skew}(B^{-1}\nabla f(X)X^\top)BX, \end{aligned}$$

where the second and fourth equalities follow from the Riemannian gradient on the Stiefel manifold and the definition of $\Phi_{B,M}$, respectively. Alternatively, one can check that the obtained expression indeed satisfies the characteristic properties of the gradient, as was done in the proof of Gao et al. (2022b, Proposition 4).

Note that the formula for $\Psi^{\mathrm{R}}_B(X)$ involves computing an inverse of B and thus does not allow a simple unbiased estimator to be used in the stochastic case, as opposed to $\Psi_B(X)$.