A Constraint Dissolving Approach for Nonsmooth Optimization over the Stiefel Manifold

Xiaoyin Hu [*], Nachuan Xiao [†], Xin Liu [‡], and Kim-Chuan Toh [§]

May 21, 2022

Abstract

This paper focus on the minimization of a possibly nonsmooth objective function over the Stiefel manifold. The existing approaches either lack efficiency or can only tackle prox-friendly objective functions. We propose a constraint dissolving function named NCDF and show that it has the same first-order stationary points and local minimizers as the original problem in a neighborhood of the Stiefel manifold. Furthermore, we show that the Clarke subdifferential of NCDF is easy to achieve from the Clarke subdifferential of the objective function. Therefore, various existing approaches for unconstrained nonsmooth optimization can be directly applied to nonsmooth optimization problems on the Stiefel manifold. We propose a framework for developing subgradient-based methods and establish their convergence properties based on prior works. Preliminary numerical experiments further highlight that the proposed constraint dissolving approach enables the efficient and direct implementations of various unconstrained solvers to nonsmooth optimization problems over the Stiefel manifold.

1 Introduction

In this paper, we consider the following nonsmooth optimization problem

$$\begin{aligned} \min_{X\in\mathbb{R}^{n\times p}} \quad & f(X) \\ \text{s. t.} \quad & X^\top X = I_p, \end{aligned} \tag{OCP}$$

where the objective function $f : \mathbb{R}^{n\times p} \mapsto \mathbb{R}$ is locally Lipschitz continuous and possibly nonsmooth, while I_p denotes the $p \times p$ identity matrix. We denote the feasible region of OCP as $\mathcal{S}_{n,p} := \{X \in \mathbb{R}^{n\times p} : X^\top X = I_p\}$, which is also referred as the Stiefel manifold throughout this paper.

Our interest in OCP with nonsmooth objective functions comes from its various applications in different engineering fields, including sparse principal component analysis [15, 62], robust subspace recovery [41, 57], packing problems [2], and training orthogonally constrained neural networks [3, 5], where the objective function is only locally Lipschitz continuous and usually not weakly convex

[*]Zhejiang University City College, Hangzhou 310015, China (huxy@zucc.edu.cn). Research is partially supported by the Zhejiang Provincial Key Research and Development Program of China (2021C01164).

[†]The Institute of Operations Research and Analytics, National University of Singapore, Singapore (xnc@lsec.cc.ac.cn). The research of this author is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 3 grant call (MOE-2019-T3-1-010).

[‡]State Key Laboratory of Scientific and Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, and University of Chinese Academy of Sciences, China (liuxin@lsec.cc.ac.cn). Research is supported in part by the National Natural Science Foundation of China (No. 12125108, 11971466, 12288201, 12021001 and 11991021), Key Research Program of Frontier Sciences, Chinese Academy of Sciences (No. ZDBS-LY-7022).

[§]Department of Mathematics, and Institute of Operations Research and Analytics, National University of Singapore, Singapore 119076 (mattohkc@nus.edu.sg). The research of this author is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 3 grant call (MOE-2019-T3-1-010).

(i.e. $f(X) + \frac{\tau}{2}\|X\|_{\mathrm{F}}^2$ is convex for some constant $\tau \geq 0$ [19]). As an illustration, let us present some motivating applications arisen from training deep neural networks.

Training deep neural networks is usually thought to be challenging both theoretically and practically, for which the vanishing/exploding gradients is one of the most important reasons [29]. To address such issue, several recent works focus on imposing orthogonality constraints to the weights of the layers in these deep neural networks [3, 5, 42]. As orthogonality implies energy preservation properties [71], these existing works demonstrate that the orthogonal constraints can stabilize the distribution of activations over layers within convolutional neural networks and make their optimization more efficient. Moreover, some existing works [43, 42, 56] observe encouraging improvements in the accuracy and robustness of the networks with orthogonal constraints. However, when these neural networks are built from nonsmooth activation functions, their loss functions are usually not weakly convex. For example, the loss function for fully connected two-layer neural network with mean squared error (MSE) can be expressed as

$$f_{\mathrm{mse}(W_1,W_2,b_1,b_2)} = \frac{1}{N}\sum_{i=1}^{N} \|\tilde{\sigma}_2(W_2\tilde{\sigma}_1(W_1x_i + b_1) + b_2) - y_i\|_2^2,$$

where $\{W_1, W_2\}$ denote the weight matrices, $\{b_1, b_2\}$ denote the bias vectors, $\{\tilde{\sigma}_1, \tilde{\sigma}_2\}$ refer to the activation functions, and $\{(x_i, y_i) \mid i = 1, ..., N\}$ is the training data set. When $\tilde{\sigma}_i$ $(i = 1, 2)$ are chosen as nonsmooth activation functions, such as rectified linear unit (ReLU) or leaky ReLU [47], the loss function is clearly locally Lipschitz continuous but not weakly convex with respect to (W_1, W_2, b_1, b_2). Therefore, we only assume that the objective function f in OCP to be locally Lipschitz continuous throughout this paper.

1.1 Existing works

Minimizing smooth objective functions over the Stiefel manifold has been extensively studied in the past several years. Interested readers can refer to the books [1, 9], the recent survey paper [34] and the references therein for more details. Compared with smooth optimization over the Stiefel manifold, research on nonsmooth cases are limited [44]. In the following, we briefly mention some existing state-of-the-art approaches for nonsmooth optimization over the Stiefel manifold.

Riemannian subgradient methods Some existing Riemannian subgradient methods are specifically designed for minimizing geodsically convex objectives over a Riemannian manifold [26, 27, 68]. However, the geodesic convexity over a compact Riemannian manifold, such as the Stiefel manifold discussed in this paper, only holds for constant functions. Therefore, these methods are invalid for any nontrivial case of OCP. Very recently, [44] studies a class of Riemannian subgradient methods to minimize weakly convex functions over the Stiefel manifold. By extending the subgradient inequality [19] from $\mathbb{R}^{n\times p}$ to the tangent spaces of the Stiefel manifold, the theoretical analysis of these Riemannian subgradient methods can be implemented by the same methodologies as existing works for unconstrained optimization [22, 19, 45]. However, the analysis in [44] relies on the weak convexity of the objective function. Therefore, when we apply subgradient methods to solve OCP, the proposed frameworks in [44] are not capable of analyzing the related theoretical properties.

Riemannian gradient sampling methods Gradient sampling methods [10, 11] are originally proposed for unconstrained nonsmooth nonconvex optimization. For nonsmooth optimization problems over a Riemannian manifold, several recent works [33, 32] proposed Riemannian gradient sampling methods based on their unconstrained origins. When applied to the minimization over the Stiefel manifold, those Riemannian gradient sampling methods first take a few sampling points $\{X_{k,j} \mid j = 1, ..., N_J\} \subset \mathcal{S}_{n,p}$ in a neighborhood of the current iterate X_k. Then they aim to find a descent direction in the convex hull of the Riemannian gradient at points $\{X_{k,j} \mid j = 1, ..., N_J\}$. However, as mentioned in [32, 33], the number of sample points N_J should be much larger than the dimension of the manifold. Therefore, for optimization over the Stiefel manifold, the number of sampling points should be much larger than $np - p(p+1)/2$ as described in [44, 11]. As emphasized in [44], it is usually expensive

to generate a descent direction in existing Riemannian gradient sampling methods, especially in high dimensional cases.

Proximal gradient methods For problems with prox-friendly objective functions, for instance, the summation of a smooth function and a nonsmooth regularizer whose proximal mapping is easy-to compute, the Riemannian proximal gradient methods are developed by extending the proximal gradient methods from Euclidean space to Riemannian manifolds. Several existing works [27, 24] develop the Riemannian proximal gradient methods by computing the proximal mappings over the Riemannian manifold. Although their global convergence properties could be established by following the same techniques as their unconstrained counterparts, computing the proximal mappings in these approaches is as difficult as the original problem, hence they are generally inefficient in practice [44]. Recently, several Riemannian proximal gradient methods [15, 36, 72] are proposed by computing the proximal mapping in the tangent space rather than over the manifold. Therefore, computing the proximal mappings in each iteration is equivalent to solving a linearly constrained strongly convex optimization problem, for which various existing efficient solvers such as semi-smooth Newton methods [51] and Arrow-Hurwicz methods [13] can be applied.

Apart from these Riemannian proximal gradient methods, a variety of existing approaches are developed by regarding OCP as a constrained optimization problem in $\mathbb{R}^{n\times p}$. Some of these approaches are built upon splitting and alternating techniques, such as the splitting method for orthogonality constrained problem (SOC) [39], ALM-based variable splitting framework [53], proximal alternating minimization approach based on the augmented Lagrangian method (PAMAL) [16], etc. By splitting the objective function and constraints apart properly, the subproblems in these approaches are usually easy to compute. However, these approaches usually lack convergence guarantees and their performance is very sensitive to the parameters, which are difficult to tune in practice [15, 44, 36].

Recently, a novel class of efficient approaches for Stiefel manifold optimization are developed based on exact penalty models. Inspired by the exact penalty model (PenC) for smooth optimization over the Stiefel manifold [61, 35], [62] extends PenC to $\ell_{2,1}$-norm regularized cases and proposes a proximal gradient method called PenCPG. In PenCPG, the proximal subproblem has a closed-form solution, which leads to its numerical superiority over existing Riemannian proximal gradient approaches.

Furthermore, for generalized composite objective functions, [63] proposed a penalty-free infeasible approach named sequential linearized proximal gradient method (SLPG). In each iteration, SLPG alternatively takes the tangential and the normal steps, both of which do not involve any penalty parameter or orthonormalization procedure. Consequently, SLPG enjoys high scalability and avoids the numerical inefficiency from inappropriately selected penalty parameters.

However, the efficiencies of all the aforementioned proximal gradient methods heavily rely on having prox-friendly objective functions. For general nonsmooth objective functions, computing its corresponding proximal mappings can be extremely expensive [44]. As a result, we need new approaches for general nonsmooth optimization over the Stiefel manifold.

1.2 Motivation

Our motivation comes from the penalty function approaches for Riemannian optimization. For smooth optimization problems on the Stiefel manifold, [60] proposes an exact penalty function (ExPen) that takes the form as follows:

$$f\left(X\left(\frac{3}{2}I_p-\frac{1}{2}X^\top X\right)\right)+\frac{\beta}{4}\left\|X^\top X-I_p\right\|_{\mathrm{F}}^2. \tag{ExPen}$$

They show that ExPen is an exact penalty function for smooth optimization over the Stiefel manifold. Different from the existing Fletcher's penalty function [28], the objective function in ExPen does not involve ∇f. Therefore, ExPen has easy-to-compute derivatives and enables direct implementation of various existing unconstrained approaches for solving smooth optimization problems over the Stiefel manifold.

Very recently, [64] proposes constraint dissolving approaches for minimizing smooth functions over a closed Riemannian manifold. In their proposed approaches, solving a Riemannian optimization problem is transferred into the unconstrained minimization of a corresponding constraint dissolving function. According to [64], the constraint dissolving function for OCP can be expressed as

$$f\left(\mathcal{A}(X)\right) + \frac{\beta}{4}\left\|X^\top X - I_p\right\|_{\mathrm{F}}^2, \tag{CDF}$$

where the constraint dissolving mapping $\mathcal{A} : \mathbb{R}^{n\times p} \to \mathbb{R}^{n\times p}$ is locally Lipschitz smooth and satisfies the following conditions.

Condition 1.1. *1. $\mathcal{A}(X) = X$ holds for any $X \in \mathcal{S}_{n,p}$.*

2. The Jacobian of $\mathcal{A}(X)^\top \mathcal{A}(X) - I_p$ equals to 0 for any $X \in \mathcal{S}_{n,p}$.

Condition 1.1 grants great flexibility in designing the constraint dissolving functions for OCP. However, for nonsmooth optimization problems over the Stiefel manifold, the chain rule usually fails when computing $\partial(f \circ \mathcal{A})$. Therefore, the Clarke subdifferential of CDF is usually not easy to achieve from ∂f and the Jacobian of $\mathcal{A}$. Meanwhile, a sufficient condition to guarantee the validity of chain rule is that $\mathcal{A}$ is a *homeomorphism* (i.e. $\mathcal{A}^{-1}$ is well-defined and continuous over $\mathbb{R}^{n\times p}$) [17, Theorem 2.3.10].

1.3 Contribution

Taking both Condition 1.1 and the desirable homeomorphism property of $\mathcal{A}$ into account, we construct the following nonsmooth constraint dissolving function named NCDF for nonsmooth optimization over the Stiefel manifold,

$$h(X) := f\left(\mathcal{A}(X)\right) + \frac{\beta}{4}\left\|X^\top X - I_p\right\|_{\mathrm{F}}^2, \tag{NCDF}$$

where $\mathcal{A} : \mathbb{R}^{n\times p} \to \mathbb{R}^{n\times p}$ is defined as

$$\mathcal{A}(X) := \frac{1}{8} X\left(15I_p - 10X^\top X + 3(X^\top X)^2\right). \tag{1.1}$$

We prove that $\mathcal{A}$ is a homeomorphism and $\partial(f \circ \mathcal{A})$ follows the chain rule [17, Theorem 2.3.10], and $\partial h(X)$ has an explicit expression which is easy to achieve from the subdifferential of f . Based on the explicit expression of $\partial h(X)$, we prove that OCP and NCDF share the same local minimizers and stationary points under mild assumptions. These properties characterize the equivalence between OCP and NCDF and further illustrate that various unconstrained approaches can be directly implemented to solve OCP through NCDF.

Moreover, we present several examples to show that OCP can be solved by directly applying subgradient methods to minimize NCDF. In these examples, we propose a class of subgradient methods and show that their convergence properties can be established directly from existing rich theoretical results in unconstrained optimization. Preliminary numerical experiments on training neural networks with normalized weights further illustrate that NCDF can be minimized by direct and efficient implementations of various existing advanced unconstrained optimization solvers. These illustrative examples highlight the great potential of NCDF.

1.4 Notations

Throughout this paper, the Euclidean inner product of two matrices $X, Y \in \mathbb{R}^{n\times p}$ is defined as $\langle X, Y\rangle = \mathrm{tr}(X^\top Y)$, where $\mathrm{tr}(A)$ is the trace of a matrix $A \in \mathbb{R}^{n\times p}$. We use $\|\cdot\|_2$ and $\|\cdot\|_{\mathrm{F}}$ to represent the 2-norm and the Frobenius norm, respectively. The notations $\mathrm{diag}(A)$ and $\mathrm{Diag}(x)$ stand for the vector formed by the diagonal entries of the matrix A, and the diagonal matrix with the entries of

$x \in \mathbb{R}^n$ as its diagonal, respectively. We denote the smallest eigenvalue of a square matrix A by $\lambda_{\min}(A)$, and set

$$\Phi(M) := \frac{1}{2}(M + M^\top).$$

Moreover, we set the metric on the Stiefel manifold as the metric inherited from the inner product in $\mathbb{R}^{n\times p}$. Then we denote $\mathcal{T}_X$ as the tangent space of the Stiefel manifold at X, which can be expressed as

$$\mathcal{T}_X := \{D \in \mathbb{R}^{n\times p} : \Phi(D^\top X) = 0\},$$

while $\mathcal{N}_X$ denotes the normal space of the Stiefel manifold at X,

$$\mathcal{N}_X := \{D \in \mathbb{R}^{n\times p} : D = X\Lambda, \Lambda = \Lambda^\top\}.$$

Additionally, $\mathcal{P}_{\mathcal{S}_{n,p}}(X) = UV^\top$ denotes the orthogonal projection to Stiefel manifold, where $X = U\Sigma V^\top$ is the economical SVD of X with $U \in \mathcal{S}_{n,p}$, $V \in \mathcal{S}_{p,p}$ and Σ is a $p \times p$ diagonal matrix with the singular values of X on its diagonal. Moreover, the ball centered at X with radius δ is defined as $\mathcal{B}(X, \delta) := \{Y \in \mathbb{R}^{n\times p} : \|Y - X\|_{\mathrm{F}} \le \delta\}$. Furthermore, $\mathcal{J}_{\mathcal{A}}(X)[D]$ refers to the linear transform of D by the linear mapping $\mathcal{J}_{\mathcal{A}}(X)$, i.e., $\mathcal{J}_{\mathcal{A}}(X)[D] = \lim_{t\to 0} \frac{1}{t}(\mathcal{A}(X + tD) - \mathcal{A}(X))$. Additionally, for any subset $\mathcal{F} \subseteq \mathbb{R}^{n\times p}$, we denote

$$\mathcal{J}_{\mathcal{A}}(X)[\mathcal{F}] := \{\mathcal{J}_{\mathcal{A}}(X)[D] : D \in \mathcal{F}\}.$$

Finally, we denote $g(X) := f(\mathcal{A}(X))$, and

$$\Omega_r = \left\{X \in \mathbb{R}^{n\times p} \,:\, \left\|X^\top X - I_p\right\|_{\mathrm{F}} \le r\right\}.$$

1.5 Organization

The rest of this paper is organized as follows. Section 2 presents several preliminary notations and definitions. We analyze the relationships between OCP and NCDF in Section 3. We show how to implement subgradient methods for OCP by NCDF and prove its theoretical convergence directly from existing works in Section 4. In Section 5, we present preliminary numerical experiments to illustrate that NCDF enables efficient and direct implementation of various existing unconstrained solvers to solve OCP. We draw a brief conclusion in the last section.

2 Preliminaries

2.1 Definitions

Definition 2.1. *Given $X \in \mathbb{R}^{n\times p}$, the generalized directional derivative of f at X in the direction $D \in \mathbb{R}^{n\times p}$, denoted by $f^\circ(X, D)$, is defined as*

$$f^\circ(X, D) = \limsup_{Y\to X, t\downarrow 0} \frac{f(Y + tD) - f(Y)}{t}.$$

Then the generalized gradient or the Clarke subdifferential of f at $X \in \mathbb{R}^{n\times p}$, denoted by $\partial f(X)$, is defined as

$$\partial f(X) := \{W \in \mathbb{R}^{n\times p} : \langle W, D\rangle \le f^\circ(X, D), \text{ for all } D \in \mathbb{R}^{n\times p}\}.$$

Definition 2.2. *We say that f is regular at $X \in \mathbb{R}^{n\times p}$ if for every direction D, the one-sided directional derivative*

$$f^\star(X, D) = \lim_{t\downarrow 0} \frac{f(X + tD) - f(X)}{t}$$

exists and $f^\star(X, D) = f^\circ(X, D)$.

Next we follow the definition in [66, 33] to present the definition for Riemannian subdifferential on the Stiefel manifold.

Definition 2.3. *Given $X \in \mathcal{S}_{n,p}$, the generalized Riemannian directional derivative of f at X in the direction $D \in \mathcal{T}_X$, denoted by $f^\diamond(X, D)$, is given by*

$$f^\diamond(X, D) = \limsup_{\mathcal{S}_{n,p} \ni Y \to X, t \downarrow 0} \frac{f\left(\mathcal{P}_{\mathcal{S}_{n,p}}(Y + tD)\right) - f(Y)}{t}.$$

Then the Riemannian subdifferential, denoted as $\partial_{\mathcal{R}} f(X)$, is defined as

$$\partial_{\mathcal{R}} f(X) := \{W \in \mathcal{T}_X : \langle W, D\rangle \leq f^\diamond(X, D), \text{ for all } D \in \mathcal{T}_X\}.$$

Definition 2.4. *Given any $X \in \mathcal{S}_{n,p}$, the projected subdifferential of f at $X \in \mathcal{S}_{n,p}$ on the Stiefel manifold is defined as*

$$\partial_{\mathcal{P}} f(X) = \{W - X\Phi(X^\top W) : W \in \partial f(X)\}.$$

Moreover, we define the first-order stationary points of problem OCP as follows.

Definition 2.5 ([17, Theorem 6.11])**.** *Given $X \in \mathcal{S}_{n,p}$, we say X is a first-order stationary point of OCP if and only if*

$$0 \in \partial_{\mathcal{P}} f(X).$$

Proposition 2.6 ([66, Theorem 5.1])**.** *For any given $X \in \mathcal{S}_{n,p}$, it holds that*

$$\partial_{\mathcal{R}} f(X) \subseteq \partial_{\mathcal{P}} f(X).$$

Furthermore, the equality holds when f is Clarke regular.

Remark 2.7. *It is worth mentioning that Proposition 2.6 implies that any local minimizer $\tilde{X}$ of OCP satisfies $0 \in \partial_{\mathcal{P}} f(X)$. Moreover, Definition 2.5 coincides with the widely used optimality conditions [66] for regular objective functions. Additionally, since $\partial f(X)$ is usually easy to compute, the projected subdifferential $\partial_{\mathcal{P}} f(X)$ is usually much easier to compute than $\partial_{\mathcal{R}} f(X)$.*

Definition 2.8 ([17, 52])**.** *For any locally Lipschitz continuous function $\psi : \mathbb{R}^{n\times p} \to \mathbb{R}$, we say $X \in \mathbb{R}^{n\times p}$ is a first-order stationary point of function ψ if and only if*

$$0 \in \partial\psi(X).$$

The following theorem is a direct corollary from [17, Theorem 2.3.10] that characterizes $\partial g(X)$, where $g := f \circ \mathcal{A}$.

Theorem 2.9. *[17, Theorem 2.3.10] For any given $X \in \mathbb{R}^{n\times p}$, it holds that*

$$\partial(f \circ \mathcal{A})(X) \subseteq \mathcal{J}_{\mathcal{A}}(X)[\partial f(\mathcal{A}(X))].$$

Moreover, when $\mathcal{A}$ maps any neighborhood of X to a set which is dense in a neighborhood of $\mathcal{A}(X)$, we have

$$\partial(f \circ \mathcal{A})(X) = \mathcal{J}_{\mathcal{A}}(X)[\partial f(\mathcal{A}(X))].$$

Finally, we define the following constants for the theoretical analysis of NCDF,

- $M_0 := \sup_{X\in\Omega_1} f(X) - \inf_{X\in\Omega_1} f(X)$;
- $M_1 := \sup_{X\in\Omega_1, W\in\partial f(X)} \|W\|_{\mathrm{F}}$.

It is worth mentioning that both M_0 and M_1 are independent of the penalty parameter β.

3 Theoretical Properties

3.1 Preliminary lemmas

In this subsection, we first present the expression for the Jacobian of the nonlinear mapping $\mathcal{A}$, which could be directly derived from the expression of $\mathcal{A}$. Here we denote $\mathcal{J}_{\mathcal{A}}(X)$ as the Jacobian of the mapping $\mathcal{A}$ at X, which can be regarded as a linear mapping from $\mathbb{R}^{n\times p}$ to $\mathbb{R}^{n\times p}$.

Proposition 3.1. *For any given $X \in \mathbb{R}^{n\times p}$, it holds that*

$$\mathcal{A}(X)^\top \mathcal{A}(X) - I_p = \frac{1}{64}(X^\top X - I_p)^3\big(9(X^\top X)^2 - 33(X^\top X) + 64I_p\big). \tag{3.1}$$

Moreover, we have for any $X \in \Omega_1$,

$$\left\|\mathcal{A}(X)^\top \mathcal{A}(X) - I_p\right\|_{\mathrm{F}} \le \left\|X^\top X - I_p\right\|_{\mathrm{F}}^3.$$

Proof. By using the economical SVD of $X = U\Sigma V^\top$, we have that

$$\mathcal{A}(X)^\top \mathcal{A}(X) - I_p = \frac{1}{64} V\Big(\Sigma^2(15I_p - 10\Sigma^2 + 3\Sigma^4)^2 - 64I_p\Big)V^T.$$

By considering the polynomial function $p(x) = x(15 - 10x + 3x^2)^2 - 64$ and noting that $p(1) = p'(1) = p''(1) = 0$, we can factorize it as $p(x) = (x-1)^3(9x^2 - 33x + 64)$. From here, we can readily obtain the required result in (3.1).

Next, for any $X \in \Omega_1$ with the economical SVD of $X = U\Sigma V^\top$, we have that $\Sigma^2 \preceq 2I_p$, and

$$\begin{aligned}
\left\|\mathcal{A}(X)^\top \mathcal{A}(X) - I_p\right\|_{\mathrm{F}} &\le \frac{1}{64}\left\|X^\top X - I_p\right\|_{\mathrm{F}}^3 \left\|9(X^\top X)^2 - 33(X^\top X) + 64I_p\right\|_2 \\
&= \frac{1}{64}\left\|X^\top X - I_p\right\|_{\mathrm{F}}^3 \max_{0\le x\le 2}\{|9x^2 - 33x + 64|\} \\
&= \left\|X^\top X - I_p\right\|_{\mathrm{F}}^3.
\end{aligned}$$

This completes the proof. □

Proposition 3.2. *For any $X \in \mathbb{R}^{n\times p}$, it holds that for any $D \in \mathbb{R}^{n\times p}$*

$$\mathcal{J}_{\mathcal{A}}(X)[D] = \frac{1}{8}D\left(15I_p - 10X^\top X + 3(X^\top X)^2\right) - X\Phi(X^\top D) + \frac{3}{2}X\Phi(\Phi(X^\top D)(X^\top X - I_p)).$$

Moreover, for any $X \in \mathcal{S}_{n,p}$, we have

$$\mathcal{J}_{\mathcal{A}}(X)[D] = D - X\Phi(D^\top X).$$

The statements in Proposition 3.2 can be verified by straightforward calculations, and hence its proof is omitted.

Lemma 3.3. *For any given $X \in \mathbb{R}^{n\times p}$, the mapping $\mathcal{J}_{\mathcal{A}}(X)$ is self-adjoint.*

Proof. For any $D, W \in \mathbb{R}^{n\times p}$, it is easy to verify that

$$\begin{aligned}
&\mathrm{tr}\left(W^\top D\left(15I_p - 10X^\top X + 3(X^\top X)^2\right)\right) = \mathrm{tr}\left(D^\top W\left(15I_p - 10X^\top X + 3(X^\top X)^2\right)\right), \\
&\mathrm{tr}\left(W^\top X\Phi(X^\top D)\right) = \mathrm{tr}\left(D^\top X\Phi(X^\top W)\right).
\end{aligned}$$

Moreover, through direct calculation, we achieve

$$\begin{aligned}
&\mathrm{tr}\left(W^\top X\Phi(\Phi(X^\top D)X^\top X)\right) = \mathrm{tr}\left(\Phi(X^\top W)\Phi(X^\top D)X^\top X\right) \\
&= \mathrm{tr}\left(X^\top X\Phi(X^\top W)\Phi(X^\top D)\right) = \mathrm{tr}\left(D^\top X\Phi(\Phi(X^\top W)X^\top X)\right).
\end{aligned}$$

Therefore, we conclude that

$$\begin{aligned}
&\langle \mathcal{J}_{\mathcal{A}}(D), W\rangle \\
&= \left\langle \frac{1}{8}D\left(15I_p - 10X^\top X + 3(X^\top X)^2\right) - \frac{5}{2}X\Phi(X^\top D) + \frac{3}{2}X\Phi(\Phi(X^\top D)X^\top X), W\right\rangle \\
&= \left\langle \frac{1}{8}W\left(15I_p - 10X^\top X + 3(X^\top X)^2\right) - \frac{5}{2}X\Phi(X^\top W) + \frac{3}{2}X\Phi(\Phi(X^\top W)X^\top X), D\right\rangle \\
&= \langle \mathcal{J}_{\mathcal{A}}(W), D\rangle,
\end{aligned}$$

and this completes the proof. □

Lemma 3.4. *$\mathcal{A}$ is a homeomorphism from $\mathbb{R}^{n\times p}$ to $\mathbb{R}^{n\times p}$. Moreover, for any given $X \in \mathbb{R}^{n\times p}$ and any $\varepsilon > 0$, there exists a $\delta > 0$ such that $\mathcal{B}(\mathcal{A}(X), \delta) \subset \mathcal{A}(\mathcal{B}(X, \varepsilon))$.*

Proof. Consider the auxiliary function $\psi(t) = t(\frac{15}{8} - \frac{5}{4}t^2 + \frac{3}{8}t^4)$. Since $\psi'(t) = \frac{15}{8}(t^2-1)^2$, we can conclude that $\psi(t)$ is non-decreasing in $\mathbb{R}$ and $\psi'(t)$ is positive except for $t = \pm 1$. As a result, $\psi(t)$ is an injection from $\mathbb{R}$ to $\mathbb{R}$, and it is easy to verify that ψ is a bijection and ψ^{-1} is continuous in $\mathbb{R}$.

Let $X = U\Sigma V^\top$ be the rank-revealing singular value decomposition of X, where both $U \in \mathbb{R}^{n\times p}$ and $V \in \mathbb{R}^{p\times p}$ are orthogonal matrices, and $\Sigma \in \mathbb{R}^{p\times p}$ is diagonal. Then from the definition of ψ, we can conclude that $\mathcal{A}(X) = U\psi(\Sigma)V^\top$, and $\mathcal{A}$ is an injection. Moreover, for any given $Y \in \mathbb{R}^{n\times p}$ and its singular value decomposition $Y = U_Y\Sigma_Y V_Y^\top$, from the definition of ψ^{-1}, it holds that $Y = \mathcal{A}(U_Y\psi^{-1}(\Sigma_Y)V_Y^\top)$. As a result, $\mathcal{A}$ is a bijection and $\mathcal{A}^{-1}$ is well-defined and continuous in $\mathbb{R}^{n\times p}$.

Therefore, $\mathcal{A}$ maps any open set to an open set due to the continuity of $\mathcal{A}^{-1}$. As a result, we can conclude that for any $X \in \mathbb{R}^{n\times p}$ and any $\varepsilon > 0$, $\mathcal{A}(\mathcal{B}(X, \varepsilon))$ is an open set that contains $\mathcal{A}(X)$. Therefore, there exists $\delta > 0$ such that $\mathcal{B}(\mathcal{A}(X), \delta) \subset \mathcal{A}(\mathcal{B}(X, \varepsilon))$ and thus we complete the proof. □

Proposition 3.5. *For any given $X \in \mathbb{R}^{n\times p}$,*

$$\partial g(X) = \mathcal{J}_{\mathcal{A}}(X)[\partial f(\mathcal{A}(X))].$$

Moreover, we have

$$\partial h(X) = \mathcal{J}_{\mathcal{A}}(X)[\partial f(\mathcal{A}(X))] + \beta X(X^\top X - I_p).$$

Proof. From Lemma 3.4, and Theorem 2.9 (Theorem 2.3.10 in [17]), we have that the chain rule holds for $f(\mathcal{A}(X))$ and this leads to the desired result. □

Proposition 3.5 illustrates that the Clarke subdifferential of g is easy to achieve from ∂f even if f is only assumed to be locally Lipschitz continuous.

Remark 3.6. *From the definition of $h(X)$ and Lemma 3.5, we can easily verify that for any $X \in \mathcal{S}_{n,p}$, it holds that*

$$h(X) = f(X), \quad \textit{and} \quad \partial_{\mathcal{P}} f(X) = \partial h(X). \tag{3.2}$$

3.2 Relationships on stationary points

In this section, we build up the relationships between the first-order stationary points of OCP and NCDF.

Lemma 3.7. *For any given $X \in \mathbb{R}^{n\times p}$ and any $W \in \mathbb{R}^{n\times p}$, it holds that*

$$\left\langle J_{\mathcal{A}}(X)[W], X(X^\top X - I_p)\right\rangle = \frac{15}{8}\left\langle \Phi(X^\top W), (X^\top X - I_p)^3\right\rangle.$$

Proof. From Proposition 3.2, we can obtain after some tedious calculations that

$$\begin{aligned}&\left\langle J_{\mathcal{A}}(X)[W], X(X^\top X - I_p)\right\rangle = \left\langle X^\top J_{\mathcal{A}}(X)[W], (X^\top X - I_p)\right\rangle\\ =&\left\langle \Phi(X^\top W), \frac{15}{8}((X^\top X)^2 - 2X^\top X + I_p)(X^\top X - I_p)\right\rangle\\ =&\frac{15}{8}\left\langle \Phi(X^\top W), (X^\top X - I_p)^3\right\rangle,\end{aligned}$$

and the proof is completed. □

Theorem 3.8. *For any given $\beta > 0$, $r \in \left(0, \frac{\beta}{2\beta+8M_1}\right)$, and any $X \in \Omega_r$, it holds that*

$$\operatorname{dist}(0, \partial h(X)) \geq \frac{\beta}{4}\left\|X^\top X - I_p\right\|_{\mathrm{F}}.$$

Furthermore, when $X^* \in \Omega_r$ is a first-order stationary point of NCDF, then X^* is a first-order stationary point of OCP.

Proof. Suppose X is not feasible, then $\left\|X^\top X - I_p\right\|_{\mathrm{F}} > 0$. Since $X \in \Omega_r$, it holds that $\left\|X^\top X - I_p\right\|_{\mathrm{F}} \leq r < 1$ and

$$\sigma_{\min}(X^\top X) \geq 1 - \sigma_{\max}(X^\top X - I_p) \geq 1 - \left\|X^\top X - I_p\right\|_{\mathrm{F}}.$$

Then for any $D \in \partial h(X)$, by Proposition 3.5 there exists $W \in \partial f(\mathcal{A}(X))$ such that $D = J_{\mathcal{A}}(X)[W] + \beta X(X^\top X - I_p)$. Moreover, by Proposition 3.1, $\mathcal{A}(X) \in \Omega_{r^3} \subset \Omega_1$, and hence $\|W\|_{\mathrm{F}} \leq M_1$. In addition, since $\|X\|_2 \leq 2$, we get $\left\|\Phi(X^\top W)\right\|_{\mathrm{F}} \leq \|X\|_2 \|W\|_{\mathrm{F}} \leq 2M_1$. Now we can conclude that

$$\begin{aligned}&2\|D\|_{\mathrm{F}}\left\|X^\top X - I_p\right\|_{\mathrm{F}} \geq \|D\|_{\mathrm{F}}\|X\|_2\left\|X^\top X - I_p\right\|_{\mathrm{F}} \geq \left\langle D, X(X^\top X - I_p)\right\rangle\\ \geq&\beta\left\langle X(X^\top X - I_p), X(X^\top X - I_p)\right\rangle - \left|\left\langle J_{\mathcal{A}}(X)[W], X(X^\top X - I_p)\right\rangle\right|\\ \geq&\beta\sigma_{\min}(X^\top X)\left\|X^\top X - I_p\right\|_{\mathrm{F}}^2 - \frac{15}{8}\left|\left\langle \Phi(X^\top W), (X^\top X - I_p)^3\right\rangle\right|\\ \geq&\beta\sigma_{\min}(X^\top X)\left\|X^\top X - I_p\right\|_{\mathrm{F}}^2 - \frac{15}{8}\left\|\Phi(X^\top W)\right\|_2\left\|X^\top X - I_p\right\|_{\mathrm{F}}^3\\ \geq&\beta\left\|X^\top X - I_p\right\|_{\mathrm{F}}^2 - \left(\frac{15M_1}{4} + \beta\right)\left\|X^\top X - I_p\right\|_{\mathrm{F}}^3\\ \geq&\beta\left\|X^\top X - I_p\right\|_{\mathrm{F}}^2\left(1 - \frac{\beta + 4M_1}{\beta}\left\|X^\top X - I_p\right\|_{\mathrm{F}}\right)\\ \geq&\frac{\beta}{2}\left\|X^\top X - I_p\right\|_{\mathrm{F}}^2.\end{aligned}$$

Therefore, it holds that

$$\|D\|_{\mathrm{F}} \geq \frac{\beta}{4}\left\|X^\top X - I_p\right\|_{\mathrm{F}}.$$

By the arbitrariness of $D \in \partial h(X)$, we get

$$\operatorname{dist}(0, \partial h(X)) \geq \frac{\beta}{4}\left\|X^\top X - I_p\right\|_{\mathrm{F}}.$$

Furthermore, when $X^* \in \Omega_r$ is a first-order stationary point of NCDF, then

$$0 = \operatorname{dist}(0, \partial h(X^*)) \geq \frac{\beta}{4}\left\|X^{*\top} X^* - I_p\right\|_{\mathrm{F}} \geq 0,$$

which implies that $X^{*\top} X^* = I_p$ and this completes the proof. □

3.3 Estimating stationarity

Directly applying unconstrained optimization approaches to solve NCDF usually yields an infeasible sequence and terminates at an infeasible point. However, sometimes we expect mild accuracy in the substationarity, but pursue high accuracy in the feasibility in minimizing NCDF. To this end, we impose an orthonormalization post-process after obtaining the solution X by applying unconstrained approaches to solve NCDF. Namely,

$$X_{\text{orth}} := \mathcal{P}_{\mathcal{S}_{n,p}}(X), \tag{3.3}$$

where $\mathcal{P}_{\mathcal{S}_{n,p}} : \mathbb{R}^{n\times p} \to \mathcal{S}_{n,p}$ is the projection on Stiefel manifold defined in Section 1.4.

Lemma 3.9. *For any given $X \in \mathbb{R}^{n\times p}$, it holds that*

$$\left\|X - \mathcal{P}_{\mathcal{S}_{n,p}}(X)\right\|_{\text{F}} \le \left\|X^\top X - I_p\right\|_{\text{F}}.$$

Proof. By the economical SVD of X, that is, $X = U\Sigma V^\top$ with $\Sigma = \text{diag}(\sigma)$, we can achieve that

$$\left\|X - \mathcal{P}_{\mathcal{S}_{n,p}}(X)\right\|_{\text{F}} = \left\|\Sigma - I_p\right\|_{\text{F}} \le \left\|\Sigma^2 - I_p\right\|_{\text{F}} = \left\|X^\top X - I_p\right\|_{\text{F}}.$$

The inequality above holds simply because $|x-1| \le |x+1||x-1|$ for any $x \ge 0$. □

Lemma 3.10. *For any given $X \in \Omega_{1/2}$, it holds that*

$$\left\|\mathcal{A}(X) - \mathcal{P}_{\mathcal{S}_{n,p}}(X)\right\|_{\text{F}} \le 4\left\|X^\top X - I_p\right\|_{\text{F}}^3. \tag{3.4}$$

Moreover

$$|f(\mathcal{A}(X)) - f(\mathcal{P}_{\mathcal{S}_{n,p}}(X))| \;\le\; M_1\left\|\mathcal{A}(X) - \mathcal{P}_{\mathcal{S}_{n,p}}(X)\right\|_{\text{F}}. \tag{3.5}$$

Proof. By the economical SVD of X, that is, $X = U\Sigma V^\top$, we obtain that

$$\mathcal{A}(X) - \mathcal{P}_{\mathcal{S}_{n,p}}(X) = \frac{1}{8}U\left(3\Sigma^5 - 10\Sigma^3 + 15\Sigma - 8I_p\right)V^T.$$

Note that $X \in \Omega_{1/2}$ implies that $0 \prec \Sigma \preceq \frac{3}{2}I_p$. Next, we have that

$$\begin{aligned}
&\left\|\mathcal{A}(X) - \mathcal{P}_{\mathcal{S}_{n,p}}(X)\right\|_{\text{F}} = \frac{1}{8}\left\|3\Sigma^5 - 10\Sigma^3 + 15\Sigma - 8I_p\right\|_{\text{F}} \\
=\ &\frac{1}{8}\left\|\left(\Sigma - I_p\right)^3\left(3\Sigma^2 + 9\Sigma + 8I_p\right)\right\|_{\text{F}} \\
\le\ &\frac{1}{8}\left\|3\Sigma^2 + 9\Sigma + 8I_p\right\|_2\left\|\left(\Sigma - I_p\right)^3\right\|_{\text{F}} \\
\le\ &\left(\frac{1}{8}\max_{0\le x\le 3/2}\{|3x^2 + 9x + 8|\}\right)\left\|\Sigma - I_p\right\|_{\text{F}}^3 \le 4\left\|\Sigma^2 - I_p\right\|_{\text{F}}^3 = 4\left\|X^\top X - I_p\right\|_{\text{F}}^3.
\end{aligned}$$

Note that we used Lemma 3.9 in the last inequality.

Next we prove (3.5). Since f is locally Lipschitz continuous, by [17, Theorem 2.3.7], there exists $t \in (0,1)$ such that $f(\mathcal{A}(X)) - f(\mathcal{P}_{\mathcal{S}_{n,p}}(X)) \in \{\langle W, \mathcal{A}(X) - \mathcal{P}_{\mathcal{S}_{n,p}}(X)\rangle : W \in \partial f(Z_t)\}$, where $Z_t = t\mathcal{A}(X) + (1-t)\mathcal{P}_{\mathcal{S}_{n,p}}(X)$. Thus

$$|f(\mathcal{A}(X)) - f(\mathcal{P}_{\mathcal{S}_{n,p}}(X))| \le \left\|\mathcal{A}(X) - \mathcal{P}_{\mathcal{S}_{n,p}}(X)\right\|_{\text{F}}\sup_{W\in\partial f(Z_t)}\|W\|_{\text{F}}$$

By some tedious calculations and using the result in (3.4), we can show that $Z_t \in \Omega_1$, and hence $\sup_{W\in\partial f(Z_t)}\|W\|_{\text{F}} \le M_1$. This completes the proof.

□

The following proposition illustrates that the post-processing procedure (3.3) can further reduce the function value simultaneously if the current point is close to the Stiefel manifold.

Proposition 3.11. *For any given $X \in \Omega_{1/2}$, it holds that*

$$h(\mathcal{P}_{\mathcal{S}_{n,p}}(X)) \leq h(X) - \left(\frac{\beta}{4} - 4M_1 \left\|X^\top X - I_p\right\|_{\mathrm{F}}\right) \left\|X^\top X - I_p\right\|_{\mathrm{F}}^2.$$

Proof.

$$\begin{aligned} h(X) - h(\mathcal{P}_{\mathcal{S}_{n,p}}(X)) &= f(\mathcal{A}(X)) - f(\mathcal{P}_{\mathcal{S}_{n,p}}(X)) + \frac{\beta}{4}\left\|X^\top X - I_p\right\|_{\mathrm{F}}^2 \\ &\overset{(i)}{\geq} -M_1 \left\|\mathcal{A}(X) - \mathcal{P}_{\mathcal{S}_{n,p}}(X)\right\|_{\mathrm{F}} + \frac{\beta}{4}\left\|X^\top X - I_p\right\|_{\mathrm{F}}^2 \\ &\geq \left(\frac{\beta}{4} - 4M_1 \left\|X^\top X - I_p\right\|_{\mathrm{F}}\right) \left\|X^\top X - I_p\right\|_{\mathrm{F}}^2. \end{aligned}$$

Here (i) follows from (3.5) in Lemma 3.10. □

Theorem 3.12. *For any given $X^* \in \Omega_{1/2}$, suppose X^* is a local minimizer of NCDF, then X^* is a local minimizer of OCP. Moreover, when $\beta \geq 16M_1$, any local minimizer X^* of OCP is a local minimizer of NCDF.*

Proof. Suppose $X^* \in \Omega_{1/2}$ is a local minimizer of NCDF. Then $0 \in \partial h(X^*)$ and Theorem 3.8 further implies that $X^* \in \mathcal{S}_{n,p}$. Notice that $h(X) = f(X)$ holds for any $X \in \mathcal{S}_{n,p}$, we immediately obtain that X^* is a local minimizer of OCP.

On the other hand, when $X^* \in \mathcal{S}$ is a local minimizer of OCP, then there exists a $\gamma \in (0,1)$ such that $f(Z) \geq f(X^*)$ holds for any $Z \in \mathcal{S}_{n,p}$ satisfying $\|Z - X^*\|_{\mathrm{F}} \leq \gamma$. Then for any $Y \in \mathbb{R}^{n\times p}$ such that $\|Y - X^*\|_{\mathrm{F}} \leq \frac{\gamma}{2}$ and $\left\|Y^\top Y - I_p\right\|_{\mathrm{F}} \leq \frac{\gamma}{2}$, we first have that

$$\left\|\mathcal{P}_{\mathcal{S}_{n,p}}(Y) - X^*\right\|_{\mathrm{F}} \leq \left\|\mathcal{P}_{\mathcal{S}_{n,p}}(Y) - Y\right\|_{\mathrm{F}} + \|Y - X^*\|_{\mathrm{F}} \leq \frac{\gamma}{2} + \frac{\gamma}{2} = \gamma.$$

Then from Proposition 3.11 we can conclude that

$$\begin{aligned} h(Y) - h(X^*) &= h(Y) - h(\mathcal{P}_{\mathcal{S}_{n,p}}(Y)) + h(\mathcal{P}_{\mathcal{S}_{n,p}}(Y)) - h(X^*) \\ &\geq \left(\frac{\beta}{4} - 4M_1 \left\|Y^\top Y - I_p\right\|_{\mathrm{F}}\right) \left\|Y^\top Y - I_p\right\|_{\mathrm{F}}^2 \geq 0, \end{aligned}$$

and complete the proof. □

Remark 3.13. *For the following optimization problem that minimizes a nonsmooth objective function over product of multiple Stiefel manifolds and Euclidean spaces,*

$$\begin{aligned} \min_{Y_i \in \mathbb{R}^{n_i \times p_i}, Z \in \mathbb{R}^l} \quad & f(Y_1, ..., Y_N, Z) \\ \text{s. t.} \quad & Y_i^\top Y_i = I_{p_i}, \text{ for } i = 1, ..., N, \end{aligned} \tag{3.6}$$

the corresponding constraint dissolving function can be formulated as

$$f(\mathcal{A}_1(Y_1), ..., \mathcal{A}_N(Y_N), Z) + \frac{\beta}{4}\left(\sum_{i=1}^N \left\|Y_i^\top Y_i - I_{p_i}\right\|_{\mathrm{F}}^2\right), \tag{3.7}$$

where $\mathcal{A}_i(Y_i) := \frac{1}{8} Y_i \left(15 I_{p_i} - 10 Y_i^\top Y_i + 3(Y_i^\top Y_i)^2\right)$. The relationship between (3.6) and (3.7) can be established in the same way as in the proofs in Section 3.

Moreover, the exactness of NCDF can be further extended to optimization problems over the generalized Stiefel manifold,

$$\begin{aligned} \min_{X \in \mathbb{R}^{n\times p}} \quad & f(X) \\ \text{s. t.} \quad & X^\top B X = I_p, \end{aligned} \tag{3.8}$$

where $B \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix. For (3.8)*, we can consider the following exact penalty function $h^\sharp$,*

$$h^\sharp(X) := f(\mathcal{A}^\sharp(X)) + \frac{\beta}{4}\left\|X^\top BX - I_p\right\|_{\mathrm{F}}^2, \tag{3.9}$$

where

$$\mathcal{A}^\sharp(X) := \frac{1}{8}X\left(15I_p - 10X^\top BX + 3(X^\top BX)^2\right). \tag{3.10}$$

Similar to the proof of Lemma 3.4, it is easy to verify that $\mathcal{A}^\sharp$ is a homeomorphism from $\mathbb{R}^{n\times p}$ to $\mathbb{R}^{n\times p}$. Therefore, we can the prove the exactness of (3.9) *together with the results in Section 2 and Section 3 through the same approach as the proofs in this paper.*

4 Subgradient Methods

In this section, we discuss how to develop subgradient methods for OCP and establish its convergence properties. Subgradient method and its variants play important roles in minimizing nonsmooth functions that are not regular, in particular, not prox-friendly. Recently, [23] shows the global convergence for subgradient methods in minimizing those functions that *admit a chain rule*, which is defined later in Definition 4.3. Such properties are essential for the so-called "descent condition" [23, Assumption D] for subgradient methods, hence are fundamental for establishing their convergence properties for solving nonconvex nonsmooth optimization, as illustrated in various existing works [23, 7, 12]. Without assuming such property, as far as we know, there is no technique to analyze these theoretical properties of subgradient methods in such general situations. Therefore, we first introduce the concept of functions that admit a chain rule in Section 4.1. Moreover, we show that functions satisfying such property can cover all the problems considered in this paper, and remarkably NCDF admits a chain rule whenever the objective function f admits a chain rule. Furthermore, we prove several important preliminary lemmas in Section 4.2, and propose a unified framework of subgradient methods in Section 4.3.

Remark 4.1. *By transforming OCP into NCDF, various existing approaches in unconstrained nonsmooth nonconvex optimization can be directly implemented to solve OCP. These approaches include gradient sampling methods [10, 38, 20], normalized subgradient methods [69], nonsmooth second-order approaches [50, 4, 65], etc. Furthermore, the convergence properties of those approaches, including the global convergence and iteration complexity, directly follow the existing related works [19, 69, 21, 20, 30].*

4.1 Preliminaries

In this subsection, we first present the concept of the functions that admit a chain rule. Then we introduce the concept of Whitney stratifiable functions, which plays an important role in nonsmooth analysis and optimization [8]. We show that all the Whitney stratifiable functions and Clarke regular functions admit a chain rule. Furthermore, we provide the definition for several important classes of functions, including semi-algebraic functions, semi-analytic functions, and functions that are definable in an *o*-minimal structure, all of which are contained in the class of Whitney stratifiable functions.

Definition 4.2. *An absolutely continuous curve is a continuous $\gamma : [0,1] \to \mathbb{R}^{n\times p}$ whose derivative exists almost everywhere in $\mathbb{R}$, and $\gamma(t) - \gamma(0)$ is Lebesgue integral of γ' between 0 and t for all $t \in \mathbb{R}$, i.e.,*

$$\gamma(t) = \gamma(0) + \int_0^t \gamma'(\tau)\mathrm{d}\tau \qquad \text{for all } t \in \mathbb{R}. \tag{4.1}$$

Definition 4.3 (Deinfition 5.1 in [23])**.** *We say a locally Lipschitz continuous function f admits a chain rule if for any absolutely continuous curve $\gamma : \mathbb{R}_+ \to \mathbb{R}^{n\times p}$, the equality*

$$(f\circ\gamma)'(t) = \langle \partial f(\gamma(t)), \gamma'(t)\rangle \tag{4.2}$$

holds for a.e. $t \geq 0$.

Lemma 4.4 ([7]). *Suppose f admits a chain rule, then for any absolutely continuous curve $\gamma : [0,1] \to \mathbb{R}^{n\times p}$, it holds that*

$$f(\gamma(1)) - f(\gamma(0)) = \int_0^1 \max_{V\in\partial f(\gamma(t))} \langle V, \gamma'(t)\rangle \,\mathrm{d}t = \int_0^1 \min_{V\in\partial f(\gamma(t))} \langle V, \gamma'(t)\rangle \,\mathrm{d}t. \tag{4.3}$$

The authors in [23, 7] prove the global convergence for applying a subgradient method to minimize locally Lipschitz continuous function that admits a chain rule. And it is also proved that Clarke regular functions admit a chain rule[23].

Proposition 4.5 (Lemma 5.4 in [23]). *Any locally Lipschitz function that is Clarke regular admits a chain rule.*

In the rest of this subsection, we focus on another broad class of functions known as Whitney stratifiable functions. Before giving a formal definition, let us fix some notations. A set $M \subset \mathbb{R}^{n\times p}$ is a C^s-smooth manifold if there is an integer $r \in \mathbb{N}$ such that around any point $X \in M$, there is a neighborhood U and a C^s-smooth mapping $F : U \mapsto \mathbb{R}^{np-r}$ such that the Jacobian of F is full rank at any $Y \in U$ and $M \cap U = \{V \in U : F(V) = 0\}$. And we denote the tangent space and normal space of M at X as $\mathcal{T}_M(X)$ and $\mathcal{N}_M(X)$, respectively.

Definition 4.6. *[23, Definition 5.6]*
A Whitney $\mathcal{C}^s$-stratification $\mathcal{E}$ of a set $Q \subset \mathbb{R}^{n\times p}$ is a partition of Q into finitely many nonempty C^s manifolds, called strata, satisfying the following compatibility conditions.

- **Frontier condition:** *For any two strata L and M, the following implication holds*

$$L \cap \mathrm{cl}(M) \neq \emptyset \quad \Longrightarrow \quad L \subset \mathrm{cl}(M).$$

- **Whitney condition:** *For any sequence of points $\{Z_k\}$ in stratum M converging to a point $\bar{Z}$ in stratum L, if the corresponding normal vectors $\{V_k\}$ chosen by $V_k \in \mathcal{N}_M(Z_k)$ converge to a vector V, then $V \in \mathcal{N}_L(\bar{Z})$.*

Moreover, we say that a function $f : \mathbb{R}^{n\times p} \to \mathbb{R}$ is Whitney $\mathcal{C}^s$-stratifiable if its graph admits a Whitney $\mathcal{C}^s$-stratification. The following theorem illustrates that Whitney stratification implies the chain rule for locally Lipschitz continuous functions.

Theorem 4.7 (Theorem 5.8 in [23]). *Any locally Lipschitz function $f : \mathbb{R}^{n\times p} \to \mathbb{R}$ that is Whitney C^1-stratifiable admits the chain rule.*

It is worth mentioning that verifying the path differentiability for nonsmooth functions is usually an easy task. As shown in [23], any Clarke regular function admits a chain rule. Moreover, Whitney $\mathcal{C}^1$-stratifiable functions contain a broad class of functions, including semi-algebraic and semi-analytic functions. A closed set Q is called semi-analytic if it can be written as a finite union of sets, each takes the form as

$$\{x \in \mathbb{R}^{n\times p} : p_i(x) \leq 0 \quad \text{for } i = 1, ...l\}, \tag{4.4}$$

for some real-analytic functions $p_1, ..., p_l$ defined on $\mathbb{R}^{n\times p}$. If all the function p_i are polynomials, then we say that Q is semi-algebraic. As mentioned in [40], any semi-analytic set admits a Whitney C^∞ stratification.

Another important class of functions is characterized based on the concept of o-minimal structure. We first present the definition for o-minimal structure, which follows the definitions in [55, 23]:

Definition 4.8. *An o-minimal structure is a collection of set $\mathcal{O} = \{\mathcal{O}_k\}_{k=1,2,...}$ where each $\mathcal{O}_k$ is a family of subsets of $\mathbb{R}^d$ such that for each $k \geq 0$:*

1. *$\mathcal{O}_k$ is closed under complementation, finite union, finite intersection and contains $\mathbb{R}^k$.*
2. *If A belongs to $\mathcal{O}_k$, then $A \times \mathbb{R}$ and $\mathbb{R} \times A$ belongs to $\mathcal{O}_{k+1}$.*

3. *If $\pi : \mathbb{R}^k \times \mathbb{R} \to \mathbb{R}^k$ denotes the coordinate projection onto $\mathbb{R}^k$, then for any A in $\mathcal{O}_{k+1}$, the set $\pi(A)$ belongs to $\mathcal{O}_k$.*

4. *$\mathcal{O}_k$ contains the family of real algebraic subsets of $\mathbb{R}^k$, that is, the set of the form $\{x \in \mathbb{R}^k : p(x) = 0\}$ for a polynomial $p(x)$ on $\mathbb{R}^k$.*

5. *The elements of $\mathcal{O}_1$ are exactly the finite unions of intervals.*

Moreover, the sets A belonging to $\mathcal{O}_d$ for some $d \geq 1$ are called definable in the o-minimal structure.

Definition 4.9. *A locally Lipschitz function f is said to be definable if its graph is definable in an o-minimal structure.*

As illustrated in [55], any function definable in an *o*-minimal structure admits a Whitney $\mathcal{C}^p$- stratification for any $p \geq 1$, and hence admits a chain rule. It straightforwardly holds from Tarski–Seidenberg theorem [6] that any semi-algebraic function is definable. Moreover, beyond semialgebraicity, [58] shows that there is an o-minimal structure that simultaneously contains both the graph of the exponential function and all semi-algebraic sets. As a result, various common activation functions and loss functions, such as sigmoid, softplus, RELU, ℓ_1-loss, MSE loss, hinge loss, logistic loss and cross-entropy loss, are all definable. Additionally, as shown in [23, 7], any composition of two definable functions is definable, which further illustrates that the loss function of a neural network built from definable activation functions is definable and thus admits a chain rule.

Therefore, the objective functions for the applications mentioned in [2, 15, 41, 54, 14, 62, 63] are Whitney $\mathcal{C}^1$-stratifiable. As a result, we can develop subgradient methods to minimize those functions and prove their convergence properties based on the framework presented in [23].

Assumption 4.10. *For the objective function f in OCP, we assume,*

1. *The objective function f in OCP admits a chain rule;*

2. *The set $\{f(X) : 0 \in \partial_{\mathcal{P}} f(X),\ X \in \mathcal{S}_{n,p}\}$ has empty interior in $\mathbb{R}$.*

It is worth mentioning that the mapping $\mathcal{A}$ is semi-algebraic, then it follows from [7] that NCDF admits a chain rule whenever the objective function f in OCP admits a chain rule.

4.2 Preliminary lemmas

We can conclude from Section 3.2 that when an unconstrained optimization approach is applied to minimize NCDF, if the generated sequence $\{X_k\}$ is contained in $\Omega_{1/6}$ and converges to some stationary points of NCDF, then $\{X_k\}$ converges to some stationary points of OCP.

In this subsection, we present two preliminary lemmas to illustrate that keeping the generated sequence restricted in $\Omega_{1/6}$ only requires mild conditions. The following lemma shows that when starting from an initial point in $\Omega_{1/12}$, a specific algorithm yields a sequence $\{X_k\}$ such that $h(X_k) \leq h(X_0)$ holds for any $k \geq 0$, then we can conclude that the sequence $\{X_k\}$ is restricted in $\Omega_{1/6}$.

Lemma 4.11. *For any $Y \in \Omega_{1/12}$ and $Z \in \Omega_1 \setminus \Omega_{1/6}$, suppose $\beta \geq 192M_0$, then it holds that $h(Y) \leq h(Z)$.*

Proof. For any $Y \in \Omega_{1/12}$, $Z \in \Omega_1 \setminus \Omega_{1/6}$, we have

$$h(Y) - h(Z) \leq \sup_{W \in \Omega_1} f(W) - \inf_{W \in \Omega_1} f(W) + \frac{\beta}{576} - \frac{\beta}{144} \leq M_0 - \frac{\beta}{192} \leq 0, \tag{4.5}$$

and complete the proof. □

Lemma 4.11 provides theoretical guarantees for a large number of algorithms, such as the gradient sampling methods [10, 38, 20], the nonsmooth second-order approaches [50, 65], etc.

Moreover, when applying subgradient-based algorithms to minimize the penalty function, we usually need to choose D_k to approximate $\partial g(X_k)$ at every iterate X_k. The direction D_k is directly chosen from $\partial g(X_k)$ in deterministic subgradient methods, or with noise in stochastic settings. In the following lemmas, we show that under mild conditions, the generated sequence $\{X_k\}$ is restricted in $\Omega_{1/6}$. Therefore, the convergence properties of deterministic subgradient methods could be established without assuming that h is bounded below in $\mathbb{R}^{n\times p}$.

Lemma 4.12. *Suppose Y_0 satisfies $Y_0 \in \Omega_{1/6}$ and Y_k is generated by*

$$Y_{k+1} = Y_k - \eta_k(D_k + \beta Y_k(Y_k{}^\top Y_k - I_p)),$$

where $D_k \in \mathbb{R}^{n\times p}$, and there exists a constant M_2 such that $\|D_k\|_{\mathrm{F}} \le M_2$ holds for any $k \ge 0$. Then when $\beta \ge 60M_2$ and $\eta_k \le \frac{1}{2\beta}$, $Y_k \in \Omega_{1/6}$ holds for any $k \ge 0$.

Proof. Note that when $\left\|Y_k{}^\top Y_k - I_p\right\|_{\mathrm{F}} \le \frac{1}{6}$, we have that $5/6 I_p \preceq Y_k{}^\top Y_k \preceq 7/6 I_p$, and hence $\|Y_k\|_2 \le 7/6$. By the definition of Y_k, let $\hat{D}_k := D_k + \beta Y_k(Y_k{}^\top Y_k - I_p)$. We have that

$$\begin{aligned}
&\left\|Y_{k+1}{}^\top Y_{k+1} - I_p\right\|_{\mathrm{F}} = \left\|Y_k{}^\top Y_k - I_p - \eta_k \hat{D}_k^\top Y_k - \eta_k Y_k{}^\top \hat{D}_k + \eta_k{}^2 \hat{D}_k^\top \hat{D}_k\right\|_{\mathrm{F}} \\
\le\ &\left\|Y_k{}^\top Y_k - I_p - 2\eta_k\beta Y_k{}^\top Y_k(Y_k{}^\top Y_k - I_p) + \eta_k{}^2\beta^2 Y_k{}^\top Y_k(Y_k{}^\top Y_k - I_p)^2\right\|_{\mathrm{F}} \\
&+ \eta_k\left\|-D_k{}^\top Y_k - Y_k{}^\top D_k + \eta_k\beta(Y_k{}^\top Y_k - I_p)Y_k{}^\top D_k + \eta_k\beta D_k{}^\top Y_k(Y_k{}^\top Y_k - I_p) + \eta_k D_k{}^\top D_k\right\|_{\mathrm{F}} \\
\le\ &\left\|(Y_k{}^\top Y_k - I_p)\Big((I_p - \eta_k\beta Y_k{}^\top Y_k)^2 - \eta_k{}^2\beta^2 Y_k{}^\top Y_k\Big)\right\|_{\mathrm{F}} \\
&+ \eta_k\Big(2\|D_k\|_{\mathrm{F}}\|Y_k\|_2 + 2\eta_k\beta\left\|Y_k{}^\top Y_k - I_p\right\|_{\mathrm{F}}\|D_k\|_{\mathrm{F}}\|Y_k\|_2 + \eta_k\|D_k\|_{\mathrm{F}}^2\Big) \\
\le\ &\left\|(Y_k{}^\top Y_k - I_p)\right\|_{\mathrm{F}}\left\|(I_p - \eta_k\beta Y_k{}^\top Y_k)^2 - \eta_k{}^2\beta^2 Y_k{}^\top Y_k\right\|_2 + \eta_k\Big(3M_2 + \eta_k M_2^2\Big) \\
\le\ &\left(1 - \frac{5}{6}\eta_k\beta\right)\left\|Y_k{}^\top Y_k - I_p\right\|_{\mathrm{F}} + 3\eta_k M_2 + \eta_k{}^2 M_2^2.
\end{aligned}$$

As a result, when $\left\|Y_k{}^\top Y_k - I_p\right\|_{\mathrm{F}} \ge \frac{1}{12}$,

$$\left\|Y_{k+1}{}^\top Y_{k+1} - I_p\right\|_{\mathrm{F}} - \left\|Y_k{}^\top Y_k - I_p\right\|_{\mathrm{F}} \le -\frac{5\eta_k\beta}{72} + 3\eta_k M_2 + \eta_k{}^2 M_2^2 \le 0.$$

Moreover, when $\left\|Y_k{}^\top Y_k - I_p\right\|_{\mathrm{F}} \le \frac{1}{12}$,

$$\left\|Y_{k+1}{}^\top Y_{k+1} - I_p\right\|_{\mathrm{F}} \le \left\|Y_k{}^\top Y_k - I_p\right\|_{\mathrm{F}} + 3\eta_k M_2 + \eta_k{}^2 M_2^2 \le \frac{1}{6}.$$

Then, we have $\left\|Y_{k+1}{}^\top Y_{k+1} - I_p\right\|_{\mathrm{F}} \le \frac{1}{6}$. Therefore, $\left\|Y_k{}^\top Y_k - I_p\right\|_{\mathrm{F}} \le \frac{1}{6}$ holds for $k = 0, 1, 2, \ldots$ by induction. □

Lemma 4.12 illustrates that when we choose the initial point in $\Omega_{1/6}$, the sequence generated by a class of subgradient-based methods are uniformly restricted in $\Omega_{1/6}$.

4.3 Convergence properties of subgradient methods

In this subsection, we consider developing subgradient methods for solving OCP by applying existing Euclidean subgradient methods to solve NCDF. We first present a unified framework (Algorithm 1) based on the framework proposed by [23]. Then we present a basic framework for subgradient methods and its stochastic version. Based on the framework presented in Algorithm 1, we analyze their convergence properties and show that these subgradient methods are globally convergent.

Condition 4.13. *In Algorithm 1, we assume*

1. *$D_k = G_k + E_k$, where $G_k \in \mathcal{J}_{\mathcal{A}}(X_k)[\partial f(X_k)]$ and $\lim\limits_{N\to+\infty}\sum\limits_{k=0}^{N} E_k\eta_k$ exists.*
2. *There exists a constant $\tilde{M}$ such that $\sup_{k\ge0}\|D_k\|_{\mathrm{F}} \le \tilde{M}$, and we choose $\beta \ge \max\{16M_1, 60\tilde{M}\}$.*

Algorithm 1 A framework of subgradient methods for solving OCP.

Require: functions f and penalty parameter β;

1: Choose an initial guess $X_0 \in \Omega_{1/6}$, set $k = 0$;
2: **while** not terminate **do**
3: Compute D_k as an approximated evaluation for $\mathcal{J}_{\mathcal{A}}(X_k)[\partial f(X_k)]$;
4: $X_{k+1} = X_k - \eta_k \left(D_k + \beta X_k (X_k{}^\top X_k - I_p) \right)$;
5: $k = k + 1$;
6: **end while**
7: Return X_k.

3. *The sequence $\{\eta_k\}$ satisfies*

$$\eta_k > 0, \quad \sum_{k=0}^{+\infty} \eta_k = \infty, \quad \sum_{k=0}^{+\infty} \eta_k^2 < +\infty, \quad \sup_{k \geq 0} \eta_k \leq \frac{1}{2\beta}.$$

Theorem 4.14. *Suppose that Assumption 4.10 holds, and Algorithm 1 satisfies Condition 4.13. Then $\{h(X_k)\}$ converges and every limit point X^* of $\{X_k\}$ is a stationary point of OCP.*

Proof. Firstly, Lemma 4.12 illustrates that when $\sup_{k\geq 0} \|D_k\|_{\mathrm{F}} \leq \tilde{M}$, $\beta \geq \max\{16M_1, 60\tilde{M}\}$, and $\eta_k \leq \frac{1}{2\beta}$ for any $k \geq 0$ in Algorithm 1, it holds that $\{X_k\}$ is restricted in $\Omega_{1/6}$. Then every cluster point of $\{X_k\}$ lies in $\Omega_{1/6}$.

Now we check the validity of Assumption A and Assumption B in [23]. The fact that $\{X_k\}$ is restricted in $\Omega_{1/6}$ directly shows that Assumption A(1) and A(2) hold. The Assumption A(3) and A(4) directly follow Condition 4.13(2) and 4.13(3), respectively. In addition, Assumption A.5 follows quickly from the fact that ∂h is outer-semicontinuous as mentioned in [23, Lemma 4.1].

Furthermore, Assumption B(1) in [23] is guaranteed by Assumption 4.10(2) and Theorem 3.8, while Assumption B(2) directly follows Assumption 4.10(1) and [23, Lemma 5.2].

Therefore, from [23, Theorem 3.2] it holds that any cluster point of $\{X_k\}$ lies in the set $\{X \in \Omega : 0 \in \partial h(X)\}$ and $h(X_k)$ converges. Then it follows from Theorem 3.8 that any point in $\{X \in \Omega : 0 \in \partial h(X)\}$ is a stationary point of OCP. Hence we complete the proof. □

It is worth mentioning that we can apply Algorithm 1 to analyze the convergence for stochastic subgradient methods for solving OCP, where the stochasticity is modeled by the sequence $\{E_k\}$ in Condition 4.13(2). We follow the standard assumptions that conditioned on the past, each random variable $\{E_k\}$ has mean zero and bounded covariance. We present detailed descriptions in Condition 4.15.

Condition 4.15. *In Algorithm 1, we assume*

1. *$D_k = G_k + E_k$, where $G_k \in \mathcal{J}_{\mathcal{A}}(X_k)[\partial f(X_k)]$ and $\{E_k\}$ is a martingale difference sequence with respect to the increasing σ-fields $\mathcal{F}_k := \sigma(X_k, G_j, E_j : j < k)$. That is, there exists a constant $\hat{M}$, and for each $k \geq 0$, E_k is measurable with respect to $\mathcal{F}_k$ and satisfies*

$$\mathbb{E}[E_k | \mathcal{F}_k] = 0, \quad \mathbb{E}[\|E_k\|_{\mathrm{F}}^2 | \mathcal{F}_k] < \hat{M}. \tag{4.6}$$

2. *There exists a constant $\tilde{M}$ such that $\sup_{k\geq 0} \|D_k\|_{\mathrm{F}} \leq \tilde{M}$ holds almost surely. Moreover, we choose $\beta \geq \max\{16M_1, 60\tilde{M}\}$.*

3. *The deterministic sequence $\{\eta_k\}$ satisfies*

$$\eta_k > 0, \quad \sum_{k=0}^{+\infty} \eta_k = \infty, \quad \sum_{k=0}^{+\infty} \eta_k^2 < +\infty, \quad \sup_{k \geq 0} \eta_k \leq \frac{1}{2\beta}. \tag{4.7}$$

Condition 4.15(2) guarantees that sequence $\sum_{k=0}^{N} \eta_k E_k$ converges almost surely as mentioned in [23]. Therefore, Condition 4.15 directly implies Condition 4.13, which yields the following corollary on the convergence of the proposed stochastic gradient method.

Corollary 4.16. *Suppose that Assumption 4.10 holds, and Algorithm 1 satisfies Condition 4.15. Then with probability* 1, $\{h(X_k)\}$ *converges and every limit point* X^* *of* $\{X_k\}$ *is a stationary point of OCP.*

5 Numerical Experiments

In this section, we apply OCP to train an orthogonally constrained convolutional neural network (OCNN) and compare its numerical performance with existing practical approaches. All the numerical experiments in this section are run on a server with Intel Xeon Silver 4110 CPU @ 2.10GHz CPU and NVIDIA Tesla P40 GPU running Pytorch 1.9.0.

5.1 Test settings

In our numerical experiments, to illustrate the practical performance of NCDF, we choose the well-known Fashion-MNIST [59] and EMNIST [18] image classification datasets. We develop the network based on the PyTorch examples on image classification. The detailed structure of the tested network is presented in Figure 1, where we enforce the orthogonality of the weights in the first full-connected layer. Let the weight of the first full-connected layer and all the other weights be denoted as $W_1 \in \mathbb{R}^{1936\times 128}$ and W_2, respectively, then we train the neural network with the constraints $W_1^\top W_1 = I_{128}$. Therefore, training this OCNN can be expressed as solving the following optimization problem,

$$\min_{W_1, W_2} \; f_{NN}(W_1, W_2), \qquad \text{s. t. } W_1^\top W_1 = I_{128}. \tag{5.1}$$

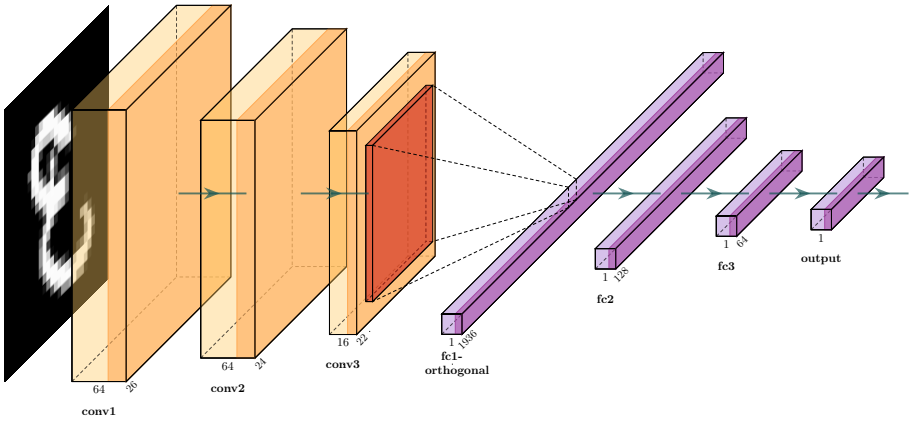

Figure 1: Structure of the OCNN. The neural network has 3 convolution layers. An average pooling and 3 fully connected layers follow the convolution layers, where we enforce the orthogonality of the weights in the first fully connected layer. The negative log-likelihood loss is used as the loss function.

Then based on the formulation of NCDF, we can reshape (5.1) as the following unconstrained optimization problem

$$\min_{W_1, W_2} \; f_{NN}(\mathcal{A}(W_1), W_2) + \frac{\beta}{4} \left\| W_1^\top W_1 - I_p \right\|_{\mathrm{F}}^2, \tag{5.2}$$

which can be solved by existing unconstrained solvers provided in the PyTorch platform.

Moreover, we compare the proposed approach with existing trivalization approaches [42]. Based on matrix exponential mapping, [42] reformulates (5.1) as

$$\min_{Z_1, W_2} f_{NN}(\exp(Z_1)\tilde{Z}, W_2), \qquad \text{s. t. } Z_1 \in \mathbb{R}^{1936\times 1936}, Z_1 + Z_1^\top = 0, \tag{5.3}$$

where exp is the matrix exponential and $\tilde{Z} \in \mathcal{S}_{1936\times 128}$ is a randomly prefixed matrix to ensure that $\exp(Z_1)\tilde{Z}$ is orthogonal. However, as mentioned in [31, 48], computing the exponential function for matrices is extremely costly. Therefore, they proposed an improved trivalization approach based on Cayley transform,

$$\min_{Z_1, W_2} f_{NN}(\mathrm{cay}(Z_1)\tilde{Z}, W_2), \qquad \text{s. t. } Z_1 \in \mathbb{R}^{1936\times 1936}, Z_1 + Z_1^\top = 0, \tag{5.4}$$

where $\mathrm{cay}(Z_1) := (I_{1936} + Z_1)^{-1}(I_{1936} - Z_1)$ and $\tilde{Z} \in \mathcal{S}_{1936\times 128}$ is a prefixed matrix on the Stiefel manifold. Although these approaches require computing the exponential function or inverse of square matrices, they transform (5.1) into an unconstrained optimization problem which can be solved directly by existing unconstrained solvers provided by the widely-used PyTorch platform .

Recently, a collection of PyTorch-based solvers, named McTorch [49], provides pre-defined full-connected and convolution layers. Therefore, we could define the model (5.1) on PyTorch platform based on McTorch and employ several specialized Riemannian solvers to train it. However, based on the framework provided in [1], transforming existing unconstrained solvers to their Riemannian versions requires deep modification to the corresponding unconstrained solvers. Therefore, only Riemannian stochastic gradient method, Riemannian version of adagrad, and a specialized Riemannian adaptive stochastic gradient method are provided in McTorch. Various other efficient solvers, including Adam [37], Nesterov accelerated Adam [25], Lamb [67], Lookahead [70], and AdamW [46] are still absent for existing Riemannian optimization packages.

5.2 Numerical results

In our numerical tests, we compare the numerical performance of training orthogonally constrained CNN by NCDF, trivialization by exponential mapping (Exp) [42], trivialization by Cayley transformation (Cayley) [31, 48], nonsmooth penalty function (L1), quadratic penalty function (Courant) and Riemannian solvers from McTorch package. We choose the penalty parameter $\beta = 0.01$ in NCDF, L1 and Courant, then we train the orthogonally constrained CNN by PyTorch build-in unconstrained solvers Adam, AdamW, NAdam, and SGD. Furthermore, we also test the Lamb solver from the torch-optimizer package, which provides various extended solvers for PyTorch. However, among all the aforementioned solvers, only Riemannian SGD is provided in McTorch package. Therefore, we only use the Riemannian SGD from the McTorch package to train the orthogonally constrained CNN in (5.1). For Adam, NAdam, AdamW, and Lamb, we set the learning rate as 0.0005 while fix all the other parameters as default. For SGD, we set the learning rate as 0.05, the momentum factor as 0.9, and fix all the other parameters as default. Furthermore, we decay the learning rate of each parameter group by 0.8 after every epoch.

All the experiments are repeated 5 times, and the batch size is set as 64. Table 1 presents the averaged performance of the selected approaches in training the proposed network with different datasets and different solvers. In Table 1, compared with the existing penalty function approaches, NCDF achieve similar training speed. Moreover, the solutions obtained by minimizing NCDF has much higher accuracy in feasibility than the solutions from other penalty-based approaches, while they all have comparable test accuracy.

Moreover, compared with the existing trivialization approaches, NCDF achieve similar accuracy as the trivialization approaches by an exponential function and Cayley transform. However, as computing the matrix exponential function and matrix inverse are expensive in GPU, those trivialization approaches take longer CPU time than NCDF significantly. Moreover, since GPU utilizes half-precision floating-point numbers, training in GPU by trivialization approaches introduces higher round-off errors, resulting in lower feasibility of the weights in practice.

Test instances		Fashion-MNIST			EMNIST-Letters			EMNIST-Balanced		
		Acc	Feas	Time (s)/epoch	Acc	Feas	Time (s)/epoch	Acc	Feas	Time (s)/epoch
ADAM	NCDF	92.79	1.68e-04	14.85	94.47	1.37e-04	29.99	88.73	1.60e-04	27.13
	Exp	92.75	9.03e-03	195.63	94.60	1.37e-02	403.23	88.71	1.43e-02	364.84
	Cayley	92.37	3.65e-03	109.76	94.47	9.65e-03	229.44	88.84	1.04e-02	206.92
	McTorch	-	-	-	-	-	-	-	-	-
	L1	92.69	2.54e-02	14.68	94.24	5.16e-02	30.10	88.71	5.41e-02	27.39
	Courant	92.50	1.38e+01	14.55	94.21	2.02e+01	30.09	88.46	2.14e+01	27.04
Adam-W	NCDF	92.74	3.13e-04	14.40	94.53	1.07e-04	29.84	88.84	1.83e-04	26.79
	Exp	92.87	9.58e-03	196.27	94.39	1.38e-02	405.06	88.84	4.00e-03	358.81
	Cayley	92.33	3.87e-03	110.18	94.45	1.04e-02	228.32	88.90	1.08e-02	209.47
	McTorch	-	-	-	-	-	-	-	-	-
	L1	92.51	2.01e-02	14.78	94.08	1.53e-02	30.17	88.37	2.48e-02	27.20
	Courant	92.08	1.29e+01	14.48	94.15	1.89e+01	29.76	88.21	2.05e+01	26.83
SGD	NCDF	91.50	6.26e-05	14.40	93.94	5.57e-05	29.50	87.89	1.14e-04	26.64
	Exp	91.46	2.30e-03	165.16	93.26	2.20e-03	343.20	87.49	2.27e-03	311.08
	Cayley	89.61	2.57e-03	108.58	93.28	2.38e-03	226.84	86.98	2.35e-03	204.88
	McTorch	91.43	2.72e-06	23.37	93.89	2.68e-06	48.30	87.92	2.60e-06	44.17
	L1	90.86	1.37e-02	14.78	93.67	1.13e-02	30.07	87.31	1.29e-02	27.29
	Courant	89.61	7.20e+00	14.52	92.93	7.42e+00	30.04	86.05	7.50e+00	26.90
Lamb	NCDF	92.96	6.87e-05	14.43	94.45	9.16e-05	30.07	88.65	4.58e-05	26.99
	Exp	93.10	2.34e-03	170.74	94.26	2.52e-03	354.44	88.58	2.59e-03	322.00
	Cayley	92.42	2.61e-03	112.94	94.40	2.46e-03	235.13	88.63	2.56e-03	207.35
	McTorch	-	-	-	-	-	-	-	-	-
	L1	93.02	1.47e-04	14.80	94.22	6.22e-03	30.58	88.62	4.73e-03	27.85
	Courant	91.98	7.06e+00	14.59	94.28	6.89e+00	31.41	88.64	6.66e+00	27.36
NAdam	NCDF	92.49	8.39e-05	14.46	94.52	1.98e-04	29.68	88.86	3.05e-04	26.80
	Exp	92.72	2.12e-02	191.58	94.41	4.43e-02	403.93	88.23	4.60e-02	356.74
	Cayley	92.70	9.59e-03	111.35	94.45	6.88e-03	229.76	88.89	7.06e-03	201.21
	McTorch	-	-	-	-	-	-	-	-	-
	L1	92.46	4.61e-02	14.76	94.13	7.17e-02	30.14	88.50	2.57e-02	27.47
	Courant	92.68	9.27e+00	14.42	94.15	1.58e+01	29.78	88.48	1.45e+01	27.06

Table 1: A comparison between NCDF and other approaches on training orthogonally constrained neural networks. "Acc" refers to the test accuracy of the final results. "Feas" denotes the feasibility of the final weights in the orthogonally constrained layer. "Time (s)/epoch" refers to the averaged running time for each epoch.

Furthermore, compared with existing Riemannian solvers, NCDF achieves similar accuracy. However, training orthogonally constrained neural networks by NCDF only requires matrix-matrix multiplication, while the solvers in McTorch compute the retractions in each iteration by singular value decomposition or QR factorization. Therefore, running SGD to minimize NCDF can be much faster than the Riemannian SGD provided by McTorch. Moreover, in the presence of difficulties in developing efficient Riemannian solvers, existing PyTorch-based Riemannian solvers are limited to SGD and Adagrad. In contrast, training orthogonally constrained neural networks by NCDF is highly adaptive to existing unconstrained solvers, and various existing efficient solvers can be applied to minimize OCP through NCDF.

6 Conclusion

Nonsmooth nonconvex optimization problems over the Stiefel manifold have real-world applications in various areas. The difficulties in handling such problems often arise from the failure of the chain rule once the required regularity conditions of the objective are missing. We present a novel constraint dissolving function NCDF for optimization problems over the Stiefel manifold OCP, whose objective functions are only assumed to be locally Lipschitz continuous. We show that OCP and NCDF share the same first-order stationary points and local minimizers in a neighborhood of the Stiefel manifold. The relationships between OCP and NCDF illustrate that minimizing a locally Lipschitz smooth function over the Stiefel manifold can be transferred into minimizing NCDF without any constraint. Moreover, we show that the Clarke subdifferential of NCDF is easy to achieve from ∂f. Therefore, the exactness and accessibility of Clarke subdifferential enable the direct implementation of various unconstrained optimization approaches for solving OCP through NCDF.

We present a representative example to demonstrate that NCDF admits direct implementation of existing unconstrained optimization approaches, while their theoretical properties are simultane-

ously retained. We then introduce a generalized framework for a class of subgradient methods and establish their convergence theories based on [23]. Furthermore, numerical examples on an orthogonally constrained neural network for classification problems demonstrate that the problem can solved via NCDF by efficient and direct implementation of existing unconstrained optimization solvers.

References

[1] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

[2] PA Absil and S Hosseini. A collection of nonsmooth Riemannian optimization problems. Technical report, Technical Report UCL-INMA-2017.08, Université catholique de Louvain, 2017.

[3] Martin Arjovsky, Amar Shah, and Yoshua Bengio. Unitary evolution recurrent neural networks. In *International Conference on Machine Learning*, pages 1120–1128. PMLR, 2016.

[4] Azam Asl and Michael L Overton. Analysis of limited-memory BFGS on a class of nonsmooth convex functions. *IMA Journal of Numerical Analysis*, 41(1):1–27, 2021.

[5] Nitin Bansal, Xiaohan Chen, and Zhangyang Wang. Can we gain more from orthogonality regularizations in training deep networks? In *Advances in Neural Information Processing Systems*, pages 4261–4271, 2018.

[6] Edward Bierstone and Pierre D Milman. Semianalytic and subanalytic sets. *Publications Mathématiques de l'IHÉS*, 67:5–42, 1988.

[7] Jérôme Bolte and Edouard Pauwels. Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning. *Mathematical Programming*, 188(1):19–51, 2021.

[8] Jérôme Bolte, Aris Daniilidis, Adrian Lewis, and Masahiro Shiota. Clarke subgradients of stratifiable functions. *SIAM Journal on Optimization*, 18(2):556–572, 2007.

[9] Nicolas Boumal. An introduction to optimization on smooth manifolds. *Available online, May*, 2020.

[10] James V Burke, Adrian S Lewis, and Michael L Overton. A robust gradient sampling algorithm for nonsmooth, nonconvex optimization. *SIAM Journal on Optimization*, 15(3):751–779, 2005.

[11] James V Burke, Frank E Curtis, Adrian S Lewis, Michael L Overton, and Lucas EA Simões. Gradient sampling methods for nonsmooth optimization. *Numerical nonsmooth optimization: State of the art algorithms*, pages 201–225, 2020.

[12] Camille Castera, Jérôme Bolte, Cédric Févotte, and Edouard Pauwels. An inertial Newton algorithm for deep learning. *Journal of Machine Learning Research*, 22(134):1–31, 2021.

[13] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[14] Shixiang Chen, Zengde Deng, Shiqian Ma, and Anthony Man-Cho So. Manifold proximal point algorithms for dual principal component pursuit and orthogonal dictionary learning. In *Asilomar Conference on Signals, Systems, and Computers*, 2019.

[15] Shixiang Chen, Shiqian Ma, Anthony Man-Cho So, and Tong Zhang. Proximal gradient method for nonsmooth optimization over the Stiefel manifold. *SIAM Journal on Optimization*, 30(1):210–239, 2020.

[16] Weiqiang Chen, Hui Ji, and Yanfei You. An augmented lagrangian method for 1-regularized optimization problems with orthogonality constraints. *SIAM Journal on Scientific Computing*, 38 (4):B570–B592, 2016.

[17] Frank H Clarke. *Optimization and nonsmooth analysis*, volume 5. Siam, 1990.

[18] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. EMNIST: Extending MNIST to handwritten letters. In *2017 international joint conference on neural networks (IJCNN)*, pages 2921–2926. IEEE, 2017.

[19] Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

[20] Damek Davis and Dmitriy Drusvyatskiy. A gradient sampling method with complexity guarantees for general Lipschitz functions. *arXiv preprint arXiv:2112.06969*, 2021.

[21] Damek Davis and Dmitriy Drusvyatskiy. Proximal methods avoid active strict saddles of weakly convex functions. *Foundations of Computational Mathematics*, pages 1–46, 2021.

[22] Damek Davis, Dmitriy Drusvyatskiy, Kellie J MacPhee, and Courtney Paquette. Subgradient methods for sharp weakly convex functions. *Journal of Optimization Theory and Applications*, 179 (3):962–982, 2018.

[23] Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D Lee. Stochastic subgradient method converges on tame functions. *Foundations of Computational Mathematics*, 20(1):119–154, 2020.

[24] Glaydston de Carvalho Bento, João Xavier da Cruz Neto, and Paulo Roberto Oliveira. A new approach to the proximal point method: convergence on general Riemannian manifolds. *Journal of Optimization Theory and Applications*, 168(3):743–755, 2016.

[25] Timothy Dozat. Incorporating Nesterov momentum into Adam. *ICLR*, 2016.

[26] OP Ferreira and PR Oliveira. Subgradient algorithm on Riemannian manifolds. *Journal of Optimization Theory and Applications*, 97(1):93–104, 1998.

[27] OP Ferreira and PR Oliveira. Proximal point algorithm on Riemannian manifolds. *Optimization*, 51(2):257–270, 2002.

[28] Roger Fletcher. A class of methods for nonlinear programming with termination and convergence properties. *Integer and nonlinear programming*, pages 157–173, 1970.

[29] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

[30] Serge Gratton, Ehouarn Simon, and Ph L Toint. An algorithm for the minimization of nonsmooth nonconvex functions using inexact evaluations and its worst-case complexity. *Mathematical Programming*, pages 1–24, 2020.

[31] Kyle Helfrich, Devin Willmott, and Qiang Ye. Orthogonal recurrent neural networks with scaled Cayley transform. In *International Conference on Machine Learning*, pages 1969–1978. PMLR, 2018.

[32] Seyedehsomayeh Hosseini and André Uschmajew. A Riemannian gradient sampling algorithm for nonsmooth optimization on manifolds. *SIAM Journal on Optimization*, 27(1):173–189, 2017.

[33] Seyedehsomayeh Hosseini, Wen Huang, and Rohollah Yousefpour. Line search algorithms for locally Lipschitz functions on Riemannian manifolds. *SIAM Journal on Optimization*, 28(1):596–619, 2018.

[34] Jiang Hu, Xin Liu, Zai-Wen Wen, and Ya-Xiang Yuan. A brief introduction to manifold optimization. *Journal of the Operations Research Society of China*, 8(2):199–248, 2020.

[35] Xiaoyin Hu and Xin Liu. An efficient orthonormalization-free approach for sparse dictionary learning and dual principal component pursuit. *Sensors*, 20(3041), 2020.

[36] Wen Huang and Ke Wei. Riemannian proximal gradient methods. *Mathematical Programming*, pages 1–43, 2021.

[37] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[38] Kiwiel and C. Krzysztof. Convergence of the gradient sampling algorithm for nonsmooth nonconvex optimization. *SIAM Journal on Optimization*, 18(2):379–388, 2007.

[39] Rongjie Lai and Stanley Osher. A splitting method for orthogonality constrained problems. *Journal of Scientific Computing*, 58(2):431–449, 2014.

[40] Stanis law Lojasiewicz. Ensembles semi-analytiques. *IHES notes*, 1965.

[41] Gilad Lerman and Tyler Maunu. An overview of robust subspace recovery. *Proceedings of the IEEE*, 106(8):1380–1410, 2018.

[42] Mario Lezcano-Casado. Trivializations for gradient-based optimization on manifolds. *arXiv preprint arXiv:1909.09501*, 2019.

[43] Mario Lezcano-Casado and David Martınez-Rubio. Cheap orthogonal constraints in neural networks: A simple parametrization of the orthogonal and unitary group. In *International Conference on Machine Learning*, pages 3794–3803. PMLR, 2019.

[44] Xiao Li, Shixiang Chen, Zengde Deng, Qing Qu, Zhihui Zhu, and Anthony Man Cho So. Weakly convex optimization over Stiefel manifold using Riemannian subgradient-type methods. *arXiv preprint arXiv:1911.05047*, 2019.

[45] Xiao Li, Zhihui Zhu, Anthony Man-Cho So, and Jason D Lee. Incremental methods for weakly convex optimization. *arXiv preprint arXiv:1907.11687*, 2019.

[46] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[47] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3. Citeseer, 2013.

[48] Jonathan H Manton. Optimization algorithms exploiting unitary constraints. *IEEE Transactions on Signal Processing*, 50(3):635–650, 2002.

[49] Mayank Meghwanshi, Pratik Jawanpuria, Anoop Kunchukuttan, Hiroyuki Kasai, and Bamdev Mishra. Mctorch, a manifold optimization library for deep learning. Technical report, arXiv preprint arXiv:1810.01811, 2018.

[50] Andre Milzarek, Xiantao Xiao, Shicong Cen, Zaiwen Wen, and Michael Ulbrich. A stochastic semismooth Newton method for nonsmooth nonconvex optimization. *SIAM Journal on Optimization*, 29(4):2916–2948, 2019.

[51] Liqun Qi and Jie Sun. A nonsmooth version of Newton's method. *Mathematical programming*, 58 (1-3):353–367, 1993.

[52] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[53] Guy Rosman, Xuecheng Tai, Ron Kimmel, and Alfred M Bruckstein. Augmented-lagrangian regularization of matrix-valued maps. *Methods and applications of analysis*, 21(1):105–122, 2014.

[54] Manolis C Tsakiris and René Vidal. Dual principal component pursuit. *The Journal of Machine Learning Research*, 19(1):684–732, 2018.

[55] Lou Van den Dries and Chris Miller. Geometric categories and o-minimal structures. *Duke Mathematical Journal*, 84(2):497–540, 1996.

[56] Jiayun Wang, Yubei Chen, Rudrasis Chakraborty, and Stella X Yu. Orthogonal convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11505–11515, 2020.

[57] Peng Wang, Huikang Liu, and Anthony Man-Cho So. Globally convergent accelerated proximal alternating maximization method for l1-principal component analysis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8147–8151. IEEE, 2019.

[58] Alex J Wilkie. Model completeness results for expansions of the ordered field of real numbers by restricted Pfaffian functions and the exponential function. *Journal of the American Mathematical Society*, 9(4):1051–1094, 1996.

[59] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms, 2017.

[60] Nachuan Xiao and Xin Liu. Solving optimization problems over the Stiefel manifold by smooth exact penalty function. *arXiv preprint arXiv:2110.08986*, 2021.

[61] Nachuan Xiao, Xin Liu, and Ya-xiang Yuan. A class of smooth exact penalty function methods for optimization problems with orthogonality constraints. *Optimization Methods and Software*, pages 1–37, 2020.

[62] Nachuan Xiao, Xin Liu, and Ya-xiang Yuan. Exact penalty function for 2,1 norm minimization over the Stiefel manifold. *SIAM Journal on Optimization*, 31(4):3097–3126, 2021.

[63] Nachuan Xiao, Xin Liu, and Ya-xiang Yuan. A penalty-free infeasible approach for a class of nonsmooth opimtization problems over the Stiefel manifold. *arXiv preprint arXiv:2103.03514*, 2021.

[64] Nachuan Xiao, Xin Liu, and Kim-Chuan Toh. Constraint dissolving approaches for Riemannian optimization. *arXiv preprint arXiv:2203.10319*, 2022.

[65] Minghan Yang, Andre Milzarek, Zaiwen Wen, and Tong Zhang. A stochastic extra-step quasi-Newton method for nonsmooth nonconvex optimization. *Mathematical Programming*, pages 1–47, 2021.

[66] Wei Hong Yang, Lei-Hong Zhang, and Ruyi Song. Optimality conditions for the nonlinear programming problems on Riemannian manifolds. *Pacific Journal of Optimization*, 10(2):415–434, 2014.

[67] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962*, 2019.

[68] Hongyi Zhang and Suvrit Sra. First-order methods for geodesically convex optimization. In *Conference on Learning Theory*, pages 1617–1638. PMLR, 2016.

[69] Jingzhao Zhang, Hongzhou Lin, Stefanie Jegelka, Ali Jadbabaie, and Suvrit Sra. Complexity of finding stationary points of nonsmooth nonconvex functions. *arXiv preprint arXiv:2002.04130*, 2020.

[70] Michael R Zhang, James Lucas, Geoffrey Hinton, and Jimmy Ba. Lookahead optimizer: k steps forward, 1 step back. *arXiv preprint arXiv:1907.08610*, 2019.

[71] Jianping Zhou, Minh N Do, and Jelena Kovacevic. Special paraunitary matrices, Cayley transform, and multidimensional orthogonal filter banks. *IEEE Transactions on Image Processing*, 15(2): 511–519, 2006.

[72] Yuhao Zhou, Chenglong Bao, Chao Ding, and Jun Zhu. A semi-smooth Newton based augmented Lagrangian method for nonsmooth optimization on matrix manifolds. *arXiv preprint arXiv:2103.02855*, 2021.