J. Appl. Numer. Optim. 3 (2021), No. 1, pp. 43-60
Available online at http://jano.biemdas.com
https://doi.org/10.23952/jano.3.2021.1.04

NON-EUCLIDEAN PROXIMAL METHODS FOR CONVEX-CONCAVE SADDLE-POINT PROBLEMS

EYAL COHEN[1], SHOHAM SABACH[2], MARC TEBOULLE[1,*]

[1] *School of Mathematical Sciences, Tel-Aviv University, Ramat-Aviv 69978, Israel*
[2] *Faculty of Industrial Engineering, The Technion, Haifa, 32000, Israel*

Abstract. Motivated by the flexibility of the Proximal Alternating Predictor Corrector (PAPC) algorithm which can tackle a broad class of structured constrained convex optimization problems via their convex-concave saddle-point reformulation, in this paper, we extend the scope of the PAPC algorithm to include non-Euclidean proximal steps. This allows for adapting to the geometry of the problem at hand to produce simpler computational steps. We prove a sublinear convergence rate of the produced ergodic sequence, and under additional natural assumptions on the non-Euclidean distances, we also prove that the algorithm globally converges to a saddle-point. We demonstrate the performance and simplicity of the proposed algorithm through its application to the multinomial logistic regression problem.

Keywords. Saddle-point problems, nonsmooth convex minimization, Non-Euclidean proximal distances and algorithms, Bregman and φ-divergences, convergence rate, iteration complexity.

1. INTRODUCTION

In this paper, we are interested in solving *convex-concave min-max* problems of the following form:

$$\text{(M)} \quad \min_{u\in\mathbb{U}} \max_{v\in\mathbb{V}} \{K(u,v) := f(u) + \langle u, \mathcal{A}v\rangle - g(v)\},$$

where $\mathbb{U}$ and $\mathbb{V}$ are Euclidean vector spaces (see Section 2 for precise assumptions on the involved functions f, g and the linear mapping $\mathcal{A}$). Nonsmooth convex-concave saddle-point problems can be used to model most of the convex optimization problems which are prevalent in a wide range of modern applications. Solving them efficiently is a very challenging task, in particular, due to the very large scale nature of the problems, and this task can be tackled in several ways: (i) via techniques of variational inequalities starting with the classical Extra-gradient method [1], see also [2, 3] for some modern extensions; (ii) splitting methods and related Lagrangian-based decomposition schemes, see the pioneering works [4, 5, 6, 7], and the recent review paper [8] and references therein; and (iii) via smoothing techniques combined with fast gradient algorithms [9, 10]. This topic of research has been studied very intensively in the last decade, with a focus on first order methods, due to the growing demand for simple and

*Corresponding author.
E-mail addresses: eyalcohen1@mail.tau.ac.il (E. Cohen), ssabach@ie.technion.ac.il (S. Sabach), teboulle@post.tau.ac.il (M. Teboulle).

Received November 8, 2020; Accepted February 2, 2021.

efficient optimization algorithms that are able to tackle various structures in large scale convex and non-smooth optimization problems. For an excellent overview which includes a plethora of relevant algorithms and modern imaging applications, see the recent comprehensive review [11], which also contains an extensive list of references.

Motivated by this trend of research, a few years ago, we have developed in [12] the Proximal Alternating Predictor Corrector (PAPC) algorithm to tackle saddle-point problems (M). By its name, the PAPC algorithm involves proximal steps, which are based on the classical proximal map of Moreau [13]. The PAPC algorithm, like most other first order methods in this domain, achieves a non-asymptotic efficiency estimate of $O(1/\varepsilon)$, where $\varepsilon > 0$ is the desired accuracy. However, the novel feature of the PAPC algorithm is in its simplicity (e.g., no matrix invesion is involved), and in its flexibility. Indeed, the PAPC method can tackle a broad class of structured optimization models that include, for example, block linear constraints (in particular with more than two blocks), as well as models with a finite sum of the composition of non-smooth functions with linear mappings in the objective or in the constraints. As shown in [12], the PAPC algorithm tackles such models by fully decomposing them into simple algorithmic steps that avoid the difficult task of computing the proximal map of the composition of a convex function with a linear map, and instead requires the proximal map of the given function.

Building on the aforementioned advantages of the PAPC algorithm, in this paper we pursue this line of research to further expand the scope of PAPC. Indeed, as alluded above, the main computational step in PAPC relies on computing the proximal map of a given convex function, which in many instances can also be a difficult task. To overcome this difficulty, we propose a *Non-Euclidean* version of PAPC, which on the one hand is proven to preserve the same convergence properties of PAPC, but on the other hand, has the additional advantage of allowing the exploitation of the geometry of the objective or the constraints in a given saddle-point problem. In turn, this allows to simplify the proximal computational step, through the use of suitable proximal distances and maps which better adapt to the problem's data at hand. By non-Euclidean, we mean replacing the classical Moreau's proximal mapping with a general proximal map as defined in [14], whereby the classical squared norm can be replaced by a general family of proximal distances (which includes the classical squared norm, as a special case). In Section 3, we present the relevant definitions and results on these proximal distances and their corresponding generalized proximal maps. These are used in Section 4, where we develop the Non-Euclidean PAPC (NEPAPC) algorithm, and prove its theoretical guarantees. We establish two main results: (i) an $O(1/\varepsilon)$ rate of convergence result of NEPAPC, which extends (and covers) the result of [12] for PAPC that was restricted to using the classical proximal map based on the squared Euclidean norm, and (ii) a global convergence analysis which ensures that NEPACP converges to a saddle-point. Finally, in Section 5, we demonstrate the applicability of NEPAPC in training a regularized multinomial logistic regression model. We show the benefits of using a non-Euclidean distance over the classical proximal map used in PAPC. Indeed, for this problem, NEPAPC produces a simple scheme with explicit iterative formula, and the numerical illustration confirms the better performance of NEPAPC over its counterpart PAPC.

Notations. Throughout this work, we employ standard convex analysis notation, as can be found in [15]. Capital letters $\mathbb{E}, \mathbb{U}, \mathbb{V}$, and $\mathbb{W}$ stand for finite dimensional vector spaces. For a given set $C \subseteq \mathbb{E}$, $\bar{C}$ denotes its closure and $\operatorname{int} C$ denotes its interior. We use the notation, $\nabla_1 \phi(\cdot, v)$, for the gradient map of the function $\phi(\cdot, v)$ with respect to its first variable. Similarly,

$\partial_1\phi(\cdot,v)$ denotes the subgradient map of the function $\phi(\cdot,v)$ with respect to its first variable. We use $\operatorname{dom}\phi$ to denote the effective domain of ϕ when ϕ is an extended valued function. For a set valued map ψ, $\operatorname{dom}\psi$ denotes its domain.

2. The Saddle-Point Model and Preliminaries

We consider the following *convex-concave min-max* problem

$$\text{(M)} \quad \min_{u\in\mathbb{U}}\max_{v\in\mathbb{V}}\{K(u,v) := f(u)+\langle u,\mathcal{A}v\rangle - g(v)\},$$

where $f:\mathbb{U}\to\mathbb{R}$ is a convex and continuously differentiable function with an L-Lipschitz continuous gradient, $g:\mathbb{V}\to(-\infty,\infty]$ is a proper, lower-semicontinuous (lsc), and convex function, and $\mathcal{A}:\mathbb{V}\to\mathbb{U}$ is a linear mapping. Unless otherwise stated, the inner-product $\langle\cdot,\cdot\rangle$ stands for the dot product, and $\|\cdot\|$ for the usual ℓ_2 norm. For the simplicity of the presentation below we denote $\mathbb{W}=\mathbb{U}\times\mathbb{V}$.

Throughout the rest of the paper, our blanket assumption is that the convex-concave function $K(\cdot,\cdot)$ has a saddle-point, i.e., there exists a feasible pair $(u^*,v^*)\in W\equiv\mathbb{U}\times\operatorname{dom}g$, which satisfies (see [15, Section 36])

$$K(u^*,v)\le K(u^*,v^*)\le K(u,v^*), \quad \forall\ (u,v)\in\mathbb{W}. \tag{2.1}$$

We denote by W^* the set of all saddle-points of $K(\cdot,\cdot)$, which is assumed to be non-empty. It is well-known that the existence of a saddle-point for problem (M) is equivalent to having a zero duality gap for the induced primal and dual problems:

$$\text{(P)} \qquad \min_{u\in\mathbb{U}}\left\{p(u) := \sup_{v\in\mathbb{V}}K(u,v) = f(u)+g^*(\mathcal{A}^Tu)\right\}$$

and

$$\text{(D)} \qquad \max_{v\in\mathbb{V}}\left\{d(v) := \inf_{u\in\mathbb{U}}K(u,v) = -g(v)-f^*(-\mathcal{A}v)\right\},$$

where ψ^* denotes the Fenchel conjugate function of ψ (see [15]). Let S_p and S_d be the optimal solution sets of the primal and dual problems, respectively. Then, the saddle-point condition (2.1) is equivalent to $p(u^*)=d(v^*)$ with $(u^*,v^*)\in S_p\times S_d$, see [15, Lemma 36.2]. For constraint qualification conditions which ensure the existence of a saddle-point see, e.g., [16, Chapter 11] and [17, Chapter 5].

Since the main goal of this paper is to find saddle-points, it will be very convenient to use the function $\Lambda:\mathbb{W}\times\mathbb{W}\to[-\infty,\infty]$, which characterizes saddle-points of $K(\cdot,\cdot)$, and will be essential for the notion of approximated saddle-points as defined below in Definition 2.1. Given two pairs $z=(x,y)\in\mathbb{W}$ and $w=(u,v)\in\mathbb{W}$, we define

$$\Lambda(z,w) := K(x,v)-K(u,y) = f(x)+g(y)+\langle x,\mathcal{A}v\rangle-\langle u,\mathcal{A}y\rangle-f(u)-g(v). \tag{2.2}$$

Thus, we obviously have the following equivalence

$$w^*\in W^* \quad\Leftrightarrow\quad \Lambda(w^*,w)\le 0, \quad \forall\, w\in W. \tag{2.3}$$

Moreover, we can easily derive a sufficient condition for a limit point to be a saddle-point of problem (M) using the function $\Lambda(\cdot,\cdot)$.

Lemma 2.1. *Let $\{w^n\}_{n\in\mathbb{N}} \subseteq W$ be a convergent sequence with a limit $\hat{w}$. Assume, for all $w \in W$, that*

$$\liminf_{n\to\infty} \Lambda(w^n, w) \leq 0. \tag{2.4}$$

Then, $\hat{w} \in W^*$.

Proof. Since $\Lambda(\cdot, w)$ is lsc, using (2.4), we get

$$\Lambda(\hat{w}, w) \leq \liminf_{n\to\infty} \Lambda(w^n, w) \leq 0,$$

and therefore the result follows from (2.3). □

Following [18], we will use the following concept of approximated saddle-points.

Definition 2.1 (ε-saddle-point)**.** Given $\varepsilon > 0$, a point $w^\varepsilon = (u^\varepsilon, v^\varepsilon) \in W$ is called *ε-saddle-point* of $K(\cdot,\cdot)$ if

$$\sup\left\{\Lambda(w^\varepsilon, w) \equiv K(u^\varepsilon, v) - K(u, v^\varepsilon) : w = (u,v) \in S_p \times S_d\right\} \leq \varepsilon. \tag{2.5}$$

We conclude this section with two classical results that will be used in our developments below. Given $x, y, z \in \mathbb{U}$, the well-known *three points Pythagoras identity* is

$$\langle y - z, x - y\rangle = \frac{1}{2}\left(\|x - z\|^2 - \|x - y\|^2 - \|y - z\|^2\right). \tag{2.6}$$

We also recall the *three points descent lemma* (cf. [12, Fact 1])

$$f\left(u^+\right) \leq f(u) + \langle \nabla f(\tilde{u}), u^+ - u\rangle + \frac{L}{2}\|u^+ - \tilde{u}\|^2, \quad \forall\, u, \tilde{u}, u^+ \in \mathbb{U}. \tag{2.7}$$

3. Non-Euclidean Distances and Proximal Mappings

In this paper, we are focusing on non-Euclidean proximal distances and their associated proximal maps in the more general sense as introduced in [14]. When considering non-Euclidean distances, Bregman distances (also called divergences) [19] are a common choice, as observed in many papers over the last decades. For initial works, we refer interested readers to the works [20, 21, 22, 23], and for modern advances to the very recent work [24] and references therein. Bregman distances are not the only representatives of non-Euclidean distance-like functions. Another popular choice is the so-called φ-divergence proximal distance, see, e.g., [21, 25, 26]. For more examples, see [14] and the references therein.

As mentioned above, in this paper we follow [14], which proposed a general framework for *proximal distances*, that covers Bregman distances, φ-divergences and others. We use here a variant of their definition as recorded below. Before doing so, in order to properly define non-Euclidean proximal distances, we first recall the notions of *essential smoothness* and *Legendre functions*, as defined in [15, Section 26].

Definition 3.1 (Essential smoothness and Legendre type)**.** Let $\psi : \mathbb{E} \to (-\infty, \infty]$ be a proper, lsc, and convex function. Then, ψ is said to be *essentially smooth* if it is differentiable on $\operatorname{int}\operatorname{dom}\psi$ and

$$\operatorname{dom}\partial\psi := \{x : \partial\psi(x) \neq \emptyset\} = \operatorname{int}\operatorname{dom}\psi. \tag{3.1}$$

If, in addition, ψ is strictly convex on $\operatorname{int}\operatorname{dom}\psi$, then it is considered *of Legendre type* or a *Legendre function*.

A list of popular choices of Legendre functions and corresponding Bregman distances and φ-divergences can be found in [14] and references therein.

Definition 3.2 (Proximal distance)**.** Let $C \subseteq \mathbb{V}$ be a nonempty, open, and convex set, and let S be a convex set such that $\overline{C} \supseteq S$ and $C \cap S$ is nonempty. A *proximal distance* with respect to (C,S) is defined as a function $\mathcal{D} : \mathbb{V} \times C \to (-\infty,\infty]$, where, for each $y \in C$, $\mathcal{D}(\cdot,y)$ is proper, lsc, convex and *essentially smooth* with $\operatorname{int}\operatorname{dom}\mathcal{D}(\cdot,y) = C$. In addition, for any $(x,y) \in S \times (C \cap S)$, we have $\mathcal{D}(x,y) \geq 0$ and $\mathcal{D}(y,y) = 0$.

Note that we have omitted the condition that $\mathcal{D}(\cdot,y)$ is level bounded, for all $y \in C$, as suggested in [14, Definition 2.1(P3)]. In [14], this additional requirement ensures that $\operatorname{prox}^{\mathcal{D}}_{\rho f}(y)$, to be precisely defined below in (3.6), is nonempty (and compact) as it is assumed that f is lower bounded on $\overline{C}$ and that $C \cap \operatorname{dom} f \neq \emptyset$, see [14, Proposition 2.1]. When f is not necessarily lower bounded, which is the case in this work, level boundedness of $\mathcal{D}(\cdot,y)$ is not enough. Thus, the nonemptiness of the proximal map can either be explicitly assumed, or ensured by other sufficient conditions, see, e.g., [24].

In [14], the notion of *induced proximal distance* was also introduced (see [14, Definition 2.2]). It is associated with each *proximal distance* and plays a key role in the convergence analysis. This notion was proposed as a natural generalization of the well-known *three points identity*, which holds for Bregman distances, see [22, Lemma 3.1]. We recall this notion below, while adapting it to our goals.

Definition 3.3 (Induced proximal distance)**.** Let $\mathcal{D}$ be a *proximal distance* with respect to (C,S). Given $\lambda \geq 0$, a function $\mathcal{R} : \mathbb{E} \times C \to (-\infty,\infty]$ is called an *λ-induced proximal distance* to $\mathcal{D}$ with respect to (C,S), if for any $y \in C \cap S$,

$$\infty > \mathcal{R}(x,y) \geq \frac{\lambda}{2}\|x-y\|^2 \geq 0, \quad \forall\, x \in S, \tag{3.2}$$

and

$$\langle \nabla_1 \mathcal{D}(z,y), x-z\rangle \leq \mathcal{R}(x,y) - \mathcal{R}(x,z) - \frac{\lambda}{2}\|z-y\|^2, \quad \forall\, x \in S, \quad \forall\, z \in C \cap S. \tag{3.3}$$

For brevity, we write $(\mathcal{D},\mathcal{R}) \in \mathcal{F}^{\lambda}(C,S)$ to denote that $[\mathcal{D},\mathcal{R},C,S]$ satisfy the conditions of Definition 3.3, with $\lambda \geq 0$. Note that, given $\lambda > 0$ and $(\mathcal{D},\mathcal{R}) \in \mathcal{F}^{\lambda}(C,S)$, $(\lambda^{-1}\mathcal{D}, \lambda^{-1}\mathcal{R}) \in \mathcal{F}^{1}(C,S)$.

Given $\lambda \geq 0$, following [14], we say that $\mathcal{D}$ is a *λ-self-proximal* with respect to (C,S), if $(\mathcal{D},\mathcal{D}) \in \mathcal{F}^{\lambda}(C,S)$. For example, taking a Legendre function $\psi : \mathbb{E} \to (-\infty,\infty]$, and the associated *Bregman distance* [19], which is defined on $\mathbb{E} \times \operatorname{int}\operatorname{dom}\psi$ by

$$D_{\psi}(x,y) := \psi(x) - \psi(y) - \langle \nabla\psi(y), x-y\rangle, \tag{3.4}$$

it can be easily verified using [22, Lemma 3.1], that the corresponding Bregman distance D_{ψ} is 0-self-proximal with respect to $(\operatorname{int}\operatorname{dom}\psi, S)$, where S is a convex set such that $S \subset \operatorname{dom}\psi$ and $S \cap \operatorname{int}\operatorname{dom}\psi \neq \emptyset$. Furthermore, for any $\lambda > 0$, $(D_{\psi}, D_{\psi}) \in \mathcal{F}^{\lambda}(\operatorname{int}\operatorname{dom}\psi, S)$ if and only if ψ is λ-strongly convex on S, i.e., $\psi + \delta_S$ is λ-strongly convex. For more important examples and interesting results dealing with proximal distances and the corresponding induced proximal distances, see [14] and the references therein.

Non-Euclidean Proximal Mappings

Let $f:\mathbb{E}\to(-\infty,\infty]$ be a proper, lsc, and convex function. Given $\rho>0$, the classical Moreau proximal mapping [13] $\text{prox}_{\rho f}:\mathbb{E}\to\mathbb{E}$, is defined by

$$\text{prox}_{\rho f}(y):=\text{argmin}_{x\in\mathbb{E}}\left\{f(x)+\frac{1}{2\rho}\|x-y\|^2\right\}. \tag{3.5}$$

The computation of the Moreau proximal mapping is not always tractable. Therefore, one way to overcome this difficulty, is to replace the quadratic proximal term with a distance-like function which better adapts to the geometry of the function f. This leads to the following extension of the Moreau proximal mapping. Given a proximal distance $\mathcal{D}$, we define the mapping $\text{prox}^{\mathcal{D}}_{\rho f}:\mathbb{E}\to\mathbb{E}$ by

$$\text{prox}^{\mathcal{D}}_{\rho f}(y):=\text{argmin}_{x\in\mathbb{E}}\left\{f(x)+\rho^{-1}\mathcal{D}(x,y)\right\}. \tag{3.6}$$

The following result, which follows [14, Proposition 2.1], establishes two important properties of this extension of the proximal mapping (cf. the proof of [14, Theorem 2.1]).

Proposition 3.1 (A well-defined proximal map and the proximal inequality)**.** *Let $f:\mathbb{E}\to(-\infty,\infty]$ be a proper, lsc, and convex function, $(\mathcal{D},\mathcal{R})\in\mathcal{F}^{\lambda}(C,\text{dom}\,f)$, with $\lambda\geq 0$, and $\rho>0$. Set $Y=C\cap\text{dom}\,f$ and assume that $\text{dom}\,\text{prox}^{\mathcal{D}}_{\rho f}\supseteq Y$. Then,*

(i) $\text{prox}^{\mathcal{D}}_{\rho f}$ maps Y to Y, and, for every $y\in Y$, $\text{prox}^{\mathcal{D}}_{\rho f}(y)$ is nonempty and closed;

(ii) for any $x\in\text{dom}\,f$, $y\in Y$, and every $z\in\text{prox}^{\mathcal{D}}_{\rho f}(y)$

$$\rho\left(f(z)-f(x)\right)\leq\langle\nabla_1\mathcal{D}(z,y),x-z\rangle\leq\mathcal{R}(x,y)-\mathcal{R}(x,z)-\frac{\lambda}{2}\|z-y\|^2. \tag{3.7}$$

Proof. Fix $\rho>0$ and $y\in Y$. As $\overline{C}\supseteq\text{dom}\,f$, it follows that

$$\text{prox}^{\mathcal{D}}_{\rho f}(y)=\text{argmin}_{x\in\mathbb{E}}\{\varphi(x):=\rho f(x)+\mathcal{D}(x,y)+\delta_{\overline{C}}(x)\}.$$

Since φ is lsc and convex, as a sum of lsc and convex functions, we derive that $\text{prox}^{\mathcal{D}}_{\rho f}(y)$, which is assumed to be nonempty, is closed. In addition, as C is open and $C\cap\text{dom}\,f\neq\emptyset$, it follows that $\text{ri}\,\text{dom}\,f\cap\text{ri}\,\text{dom}\,\mathcal{D}(\cdot,y)\cap\text{ri}\,\text{dom}\,\delta_{\overline{C}}$ is nonempty. Thus, by applying [15, Theorem 23.8], we obtain, for any $z\in\mathbb{E}$, that

$$\partial\varphi(z)=\rho\partial f(z)+\partial_1\mathcal{D}(z,y)+N_{\overline{C}}(z),$$

where $N_{\overline{C}}:\mathbb{E}\rightrightarrows\mathbb{E}$ is the normal cone of $\overline{C}$, which is defined for all $z\in\overline{C}$ by

$$N_{\overline{C}}(z)=\left\{w\in\mathbb{E}:\langle w,x-z\rangle\leq 0,\quad\forall\,x\in\overline{C}\right\},$$

and $\text{dom}\,N_{\overline{C}}=\overline{C}$. On the other hand, since C is an open set, it follows that $N_{\overline{C}}(z)=\{0\}$ for all $z\in C$. Therefore, using the *essential smoothness* of $\mathcal{D}(.,y)$, it follows from the Fermat's optimality condition, for any $z\in\text{prox}^{\mathcal{D}}_{\rho f}$, that $z\in Y=C\cap\text{dom}\,f$ with

$$0\in\rho\partial f(z)+\nabla_1\mathcal{D}(z,y). \tag{3.8}$$

Thus, $-\rho^{-1}\nabla_1\mathcal{D}(z,y)\in\partial f(z)$. By applying the subgradient inequality for the convex function f, we obtain the left inequality in (3.7). The right inequality is a direct result of Definition 3.3. □

4. Non-Euclidean Proximal Alternating Predictor Corrector

4.1. **The Algorithm NEPAPC.** We follow the description of the Proximal Alternating Predictor Corrector (PAPC) algorithm as introduced in [12], and extends the algorithm's applicability by replacing the classical Moreau's proximal mapping with the general proximal mapping that corresponds to a well-chosen proximal distance $\mathcal{D}$. This requires the following assumption, which will be assumed throughout the rest of the work.

Assumption 4.1. Given $g : \mathbb{V} \to (-\infty,\infty]$ a proper, lsc and convex function, let $(\mathcal{D},\mathcal{R}) \in \mathcal{F}^1(C,\operatorname{dom} g)$ be an induced proximal distance such that, for every $\sigma \leq 1/(\tau\|\mathcal{A}\|^2)$ and every $z \in \mathbb{V}$, $\operatorname{dom}\operatorname{prox}^{\mathcal{D}}_{\sigma(g+\langle z,\cdot\rangle)} \supseteq C \cap \operatorname{dom} g$.

The Non-Euclidean extension of PAPC is described as follows. Note that with $\mathcal{D}$ being the classical squared Euclidean distance, NEPAPC reduces to PAPC [12].

Algorithm 1 Non-Euclidean Proximal Alternating Predictor Corrector (NEPAPC)

Initialization. $\tau \leq 1/L$, $\sigma \leq 1/(\tau\|\mathcal{A}\|^2)$, $u^0 \in \mathbb{U}$ and $v^0 \in C \cap \operatorname{dom} g$.
General step. For $k = 1,2,\ldots$ compute:

$$p^k = u^{k-1} - \tau(\nabla f(u^{k-1}) + \mathcal{A}v^{k-1}), \tag{4.1}$$

$$\begin{aligned} v^k &\in \operatorname{prox}^{\mathcal{D}}_{\sigma(g-\langle \mathcal{A}^T p^k,\cdot\rangle)}(v^{k-1}) \\ &= \operatorname{argmin}_{v\in\mathbb{V}}\left\{ g(v) - \langle \mathcal{A}^T p^k, v\rangle + \frac{1}{\sigma}\mathcal{D}\left(v, v^{k-1}\right)\right\}, \end{aligned} \tag{4.2}$$

$$u^k = u^{k-1} - \tau(\nabla f(u^{k-1}) + \mathcal{A}v^k). \tag{4.3}$$

Assumption 4.1 together with Proposition 3.1 guarantees the validity for the proximal step (4.2) and therefore the algorithm NEPAPC is well-defined as recorded in the following result.

Proposition 4.1. *Let $\{w^k = (u^k, v^k)\}_{k\in\mathbb{N}}$ be a sequence generated by NEPAPC. Then,*

$$\{w^k\}_{k\in\mathbb{N}} \subseteq \mathbb{U} \times (C \cap \operatorname{dom} g).$$

A key advantage of PAPC, which is also relevant to the non-Euclidean variant NEPAPC, is that it decomposes well according to the problem's structure in terms of *block separability*. Similarly to PAPC, the step (4.2) of NEPAPC may be computationally challenging. However, when the model's data is block separable it can be decomposed as discussed next.

Consider the block variant of problem (M) given by

$$\text{(SM)} \quad \min_{u\in\mathbb{U}} \max_{\substack{v_i\in\mathbb{V}_i \\ i=1,2,\ldots,m}} \left\{ f(u) + \langle u, \sum_{i=1}^m A_i v_i\rangle - \sum_{i=1}^m g_i(v_i)\right\},$$

where each $g_i : \mathbb{V}_i \to (-\infty,\infty]$ is a proper, lsc, and convex function, and $A_i : \mathbb{V}_i \to \mathbb{U}$ is a linear mapping, for all $i = 1,2,\ldots,m$. Following Assumption 4.1, we assume here that there exist induced proximal distances $(\mathcal{D}_i,\mathcal{R}_i) \in \mathcal{F}^1(C_i,\operatorname{dom} g_i)$ such that $\operatorname{dom}\operatorname{prox}^{\mathcal{D}_i}_{\sigma(g_i+\langle z_i,\cdot\rangle)} \supseteq C_i \cap \operatorname{dom} g_i$, for every $\sigma > 0$ and $z_i \in \mathbb{V}_i$. It is easy to verify that this block model can be captured as a

particular instance of model (M), which implies that the step (4.2) can be decomposed and parallelized, for all $i = 1,2,\dots,m$, as follows

$$v_i^k \in \operatorname{prox}^{\mathcal{D}_i}_{\sigma(g_i - \langle A_i^T p^k, \cdot\rangle)}(v_i^{k-1}).$$

As can be seen from the above parallel updating steps, the proximal parameter σ is shared by all the proximal steps, and is bounded by $1/\tau\|\mathcal{A}\|^2$ where $\mathcal{A} = (A_1, A_2, \dots, A_m)$. When $\mathcal{A}$ is *ill-conditioned* this may cause the proximal steps to be small. Therefore, in order to allow flexibility for each block, we propose the following *preconditioning* scheme. Let $\omega_i > 0$, $i = 1,2,\dots,m$, be the precondition coefficient for block i. First, we change the variables v_i, $i = 1,2,\dots,m$, as follows $z_i = \omega_i^{-1} v_i$. By defining $\mathcal{D}^{\omega_i}(x,y) := \omega_i^{-2}\mathcal{D}_i(\omega_i x, \omega_i y)$ and $\mathcal{R}^{\omega_i}(x,y) := \omega_i^{-2}\mathcal{R}_i(\omega_i x, \omega_i y)$, we have that $(\mathcal{D}_i^{\omega_i}, \mathcal{R}_i^{\omega_i}) \in \mathcal{F}^1(\omega_i^{-1}C_i, \operatorname{dom} g_i(\omega_i \cdot))$. Therefore, the proximal step for updating the new variable z_i, $i = 1,2,\dots,m$, is given by

$$\begin{aligned} z_i^k &\in \operatorname{prox}^{\mathcal{D}^{\omega_i}}_{\sigma(g_i(\omega_i\cdot) - \langle \omega_i A_i^T p^k, \cdot\rangle)}(z_i^{k-1}) \\ &= \operatorname{argmin}_{z_i \in \mathbb{V}_i}\left\{ g_i(\omega_i z_i) - \langle \omega_i A_i^T p^k, z_i\rangle + \frac{1}{\sigma\omega^2}\mathcal{D}_i(\omega_i z_i, \omega_i z_i^{k-1})\right\} \\ &= \omega_i^{-1}\operatorname{argmin}_{v_i \in \mathbb{V}_i}\left\{ g_i(v_i) - \langle A_i^T p^k, v_i\rangle + \frac{1}{\sigma\omega^2}\mathcal{D}_i(v_i, v_i^{k-1})\right\}, \end{aligned}$$

where the last equality uses the fact that $z_i^k = \omega_i^{-1} v_i^k$, $k \in \mathbb{N}$. Thus, for all $i = 1,2,\dots,m$,

$$v_i^k \in \operatorname{prox}^{\mathcal{D}_i}_{\sigma\omega_i^2(g_i - \langle A_i p^k, \cdot\rangle)}(v_i^{k-1}).$$

Therefore, the NEPAPC for block separable problems with preconditioning is recorded next.

Algorithm 2 NEPAPC for the block model (SM) with preconditioning

Initialization. Let $u^0 \in \mathbb{U}$ and for each $i = 1,2,\dots,m$, $\omega_i > 0$ and $v_i^0 \in C_i \cap \operatorname{dom} g_i$. Set $\tau \le 1/L$ and $\sigma \le 1/(\tau\|\mathcal{A}_\omega\|^2)$, where $\mathcal{A}_\omega = (\omega_1 A_1, \omega_2 A_2, \dots, \omega_m A_m)$.
General step. For $k = 1,2,\dots$ compute:

$$p^k = u^{k-1} - \tau\left(\nabla f(u^{k-1}) + \sum_{i=1}^m A_i v_i^{k-1}\right), \tag{4.4}$$

$$v_i^k \in \operatorname{prox}^{\mathcal{D}_i}_{\sigma\omega_i^2(g_i - \langle A_i^T p^k, \cdot\rangle)}(v_i^{k-1}), \quad i = 1,2,\dots,m, \tag{4.5}$$

$$u^k = u^{k-1} - \tau\left(\nabla f(u^{k-1}) + \sum_{i=1}^m A_i v_i^k\right). \tag{4.6}$$

4.2. **Convergence Analysis.** In this section, following PAPC, we describe the two main results: (i) a convergence of NEPAPC (cf. Algorithm 1) to a saddle-point of problem (M), (ii) a sublinear rate of convergence result of NEPAPC for the *ergodic sequence*. We begin with the following technical lemma that collects some useful properties of NEPAPC that will be essential in proving the main results (see Theorems 4.1 and 4.2).

Lemma 4.1. *Let $\{w^k = (u^k, v^k)\}_{k\in\mathbb{N}}$ be a sequence generated by NEPAPC and suppose that Assumption 4.1 holds. Then, the following statements hold.*

(i) For every $v \in \operatorname{dom} g$ and every $k \in \mathbb{N}$,

$$K(u^k, v) - K(u^k, v^k) \le \frac{1}{\sigma}\left(\hat{\mathcal{R}}(v, v^{k-1}) - \hat{\mathcal{R}}(v, v^k) - \frac{1}{2}\left(1 - \sigma\tau\|\mathcal{A}\|^2\right)\|v^k - v^{k-1}\|^2\right), \tag{4.7}$$

with $\hat{\mathcal{R}}(x,y) := \mathcal{R}(x,y) - \frac{1}{2}\sigma\tau\|\mathcal{A}(x-y)\|^2$.

(ii) For every $u \in \mathbb{U}$ and every $k \in \mathbb{N}$,

$$K(u^k, v^k) - K(u, v^k) \le \frac{1}{2\tau}\left(\|u - u^{k-1}\|^2 - \|u - u^k\|^2 - (1 - \tau L)\|u^k - u^{k-1}\|^2\right). \tag{4.8}$$

(iii) For every $w = (u, v) \in \mathbb{U} \times \operatorname{dom} g$ and $k \in \mathbb{N}$, we have

$$\Lambda\left(w^k, w\right) \le \Gamma\left(w, w^{k-1}\right) - \Gamma\left(w, w^k\right) - \frac{\beta}{2}\|w^k - w^{k-1}\|^2, \tag{4.9}$$

where $\beta = \min\{\sigma^{-1}(1 - \sigma\tau\|\mathcal{A}\|^2), \tau^{-1}(1 - \tau L)\} \ge 0$ and for $\tilde{w} = (\tilde{u}, \tilde{v}) \in \mathbb{U} \times C$

$$\Gamma(w, \tilde{w}) := \frac{1}{2\tau}\|u - \tilde{u}\|^2 + \frac{1}{\sigma}\hat{\mathcal{R}}(v, \tilde{v}). \tag{4.10}$$

Moreover, for all $y \in Y$, we have with $\alpha = \min\{\sigma^{-1}(1 - \sigma\tau\|\mathcal{A}\|^2), \tau^{-1}\}$, that

$$\Gamma(w, y) \ge \frac{\alpha}{2}\|w - y\|^2. \tag{4.11}$$

Proof. Fix $k \in \mathbb{N}$ and $(u, v) \in \mathbb{U} \times \operatorname{dom} g$. We have

$$\begin{aligned} K(u^k, v) - K(u^k, v^k) &= g(v^k) - g(v) + \langle u^k, \mathcal{A}(v - v^k)\rangle \\ &= g(v^k) - g(v) - \langle p^k, \mathcal{A}(v^k - v)\rangle + \langle p^k - u^k, \mathcal{A}(v^k - v)\rangle \\ &= g(v^k) - g(v) - \langle \mathcal{A}^T p^k, v^k - v\rangle - \tau\langle \mathcal{A}(v^{k-1} - v^k), \mathcal{A}(v^k - v)\rangle, \end{aligned} \tag{4.12}$$

where the last equality is due to steps (4.1) and (4.3). Applying Proposition 3.1 to the step (4.2) yields

$$g(v^k) - g(v) - \langle \mathcal{A}^T p^k, v^k - v\rangle \le \frac{1}{\sigma}\left(\mathcal{R}(v, v^{k-1}) - \mathcal{R}(v, v^k) - \frac{1}{2}\|v^k - v^{k-1}\|^2\right). \tag{4.13}$$

Finally, due to the Pythagoras identity (2.6), we have

$$\begin{aligned} -\tau\langle \mathcal{A}(v^{k-1} - v^k), \mathcal{A}(v^k - v)\rangle &= \frac{\tau}{2}\left(-\|\mathcal{A}(v - v^{k-1})\|^2 + \|\mathcal{A}(v - v^k)\|^2 + \|\mathcal{A}(v^k - v^{k-1})\|^2\right) \\ &\le -\frac{\tau}{2}\|\mathcal{A}(v - v^{k-1})\|^2 + \frac{\tau}{2}\|\mathcal{A}(v - v^k)\|^2 + \frac{\tau}{2}\|\mathcal{A}\|^2\|v^k - v^{k-1}\|^2. \end{aligned} \tag{4.14}$$

Thus, combining (4.12), (4.13), and (4.14) completes the proof of item (i).

The proof of the second item is actually identical to that of [12, Lemma 3.1(i)]. For completeness, we repeat its simple proof:

$$\begin{aligned} K(u^k, v^k) - K(u, v^k) &= f(u^k) - f(u) + \langle \mathcal{A}v^k, u^k - u\rangle \\ &= f(u^k) - f(u) - \langle \nabla f(u^{k-1}), u^k - u\rangle + \frac{1}{\tau}\langle u^{k-1} - u^k, u^k - u\rangle, \end{aligned} \tag{4.15}$$

where the second equality is due to step (4.3). Applying now the three points descent lemma (see (2.7)) and the Pythagoras identity (2.6), we complete the proof of item (ii).

The third item easily follows from the definition of β by summing (4.7) and (4.8). Finally, by recalling that $\sigma\tau\|\mathcal{A}\|^2 < 1$ with $y = (y_1, y_2)$, we have

$$\hat{\mathcal{R}}(y_1, y_2) = \mathcal{R}(y_1, y_2) - \frac{1}{2}\sigma\tau\|\mathcal{A}(y_1 - y_2)\|^2 \geq \frac{1}{2}\left(1 - \sigma\tau\|\mathcal{A}\|^2\right)\|y_1 - y_2\|^2,$$

and so, with $\alpha = \min\{\sigma^{-1}(1 - \sigma\tau\|\mathcal{A}\|^2), \tau^{-1}\}$, we easily obtain that $\Gamma(w, y) \geq (\alpha/2)\|w - y\|^2$. This completes the proof. □

Now, we can immediately obtain the first main result: a rate of convergence of NEPAPC in the ergodic sense.

Theorem 4.1 (Convergence rate for the ergodic sequence)**.** *Let $\{w^k = (u^k, v^k)\}_{k\in\mathbb{N}}$ be a sequence generated by NEPAPC and suppose that Assumption 4.1 holds. Then, for any $w \in \mathbb{U} \times \operatorname{dom} g$ and $N \in \mathbb{N}$, the following holds for the ergodic sequence $\overline{w}^N = (1/N)\sum_{k=1}^N w^k$*

$$\Lambda\left(\overline{w}^N, w\right) \leq \frac{1}{N}\left(\frac{1}{2\tau}\|u - u^0\|^2 + \frac{1}{\sigma}\hat{\mathcal{R}}\left(v, v^0\right)\right). \tag{4.16}$$

In addition, assume that the primal and dual optimal solution sets, S_p and S_d, are compact and that $\mathcal{R}(\cdot, y)$ is bounded on any compact subset of $\operatorname{dom} g$, for every $y \in C \cap \operatorname{dom} g$. Then, for any $\varepsilon > 0$, $\overline{w}^N$ is an ε-saddle-point with $\varepsilon = O(1/N)$.

Proof. Recalling that $\Gamma(w, \tilde{w}) = (1/(2\tau))\|u - \tilde{u}\|^2 + (1/\sigma)\hat{\mathcal{R}}(v, \tilde{v})$. Since $\Lambda(\cdot, w)$ is convex (cf. Section 2), by Jensen's inequality we have

$$\Lambda\left(\overline{w}^N, w\right) = \Lambda\left(\frac{1}{N}\sum_{k=1}^N w^k, w\right) \leq \frac{1}{N}\sum_{k=1}^N \Lambda\left(w^k, w\right) \leq \frac{1}{N}\sum_{k=1}^N \left(\Gamma\left(w, w^{k-1}\right) - \Gamma\left(w, w^k\right)\right),$$

where the last inequality follows from (4.9) (after omiting the non-negative term $(\beta/2)\|w^k - w^{k-1}\|^2$). Therefore, by combining now with (4.9), we obtain

$$\Lambda\left(\overline{w}^N, w\right) \leq \frac{1}{N}\left(\Gamma\left(w, w^0\right) - \Gamma\left(w, w^N\right)\right) \leq \frac{1}{N}\Gamma\left(w, w^0\right),$$

where the last inequality follows from the fact that $\Gamma\left(w, w^N\right) \geq 0$ thanks to (4.11) of Lemma 4.1(iii). The first result now follows from the definition of $\Gamma(\cdot, \cdot)$. The second result follows now immediately from the definition of ε-saddle-point (see Definition 2.1). □

We proceed to our second main result, which states the conditions for asymptotic convergence to a saddle-point of NEPAPC. To this end, we need some additional assumptions on the induced proximal distance.

Assumption 4.2. Given an induced proximal distance $(\mathcal{D}, \mathcal{R}) \in \mathcal{F}^1(C, S)$ with $S = \operatorname{dom} g$.

(i) For any two convergent sequences $\{x^n\}_{n\in\mathbb{N}}, \{z^n\}_{n\in\mathbb{N}} \subseteq C \cap S$, if $\lim_{n\to\infty} x^n = \lim_{n\to\infty} z^n$ then, for all $v \in S$, we have

$$\lim_{n\to\infty}\left(\mathcal{R}(v, x^n) - \mathcal{R}(v, z^n)\right) = 0. \tag{4.17}$$

(ii) For any convergent sequence $\{v^n\}_{n\in\mathbb{N}} \subseteq C \cap S$, if $\lim_{n\to\infty} v^n = v^* \in S$ then $\lim_{n\to\infty} \mathcal{R}(v^*, v^n) = 0$.

Theorem 4.2 (Pointwise convergence to saddle-points)**.** *Let $\{w^k = (u^k, v^k)\}_{k\in\mathbb{N}}$ be a sequence generated by NEPAPC and suppose that Assumption 4.1 holds. Then, the following assertions hold.*

(i) Assume that $\sigma\tau\|\mathcal{A}\|^2 < 1$. Then, the sequence $\{w^k\}_{k\in\mathbb{N}}$ is bounded.
(ii) Assume that $\tau < 1/L$ and $\sigma\tau\|\mathcal{A}\|^2 < 1$. If Assumption 4.2(i) holds true, then, any limit point of the sequence $\{w^k\}_{k\in\mathbb{N}}$ is a saddle-point of problem (M).
(iii) Assume $\tau < 1/L$ and $\sigma\tau\|\mathcal{A}\|^2 < 1$. If Assumption 4.2 holds true, then, the sequence $\{w^k\}_{k\in\mathbb{N}}$ converges to a saddle-point of problem (M).

Proof. Let w^* be a saddle-point of problem (M). From (2.2) and (2.3) it follows that $\Lambda\left(w^k, w^*\right) = -\Lambda\left(w^*, w^k\right) \geq 0$. Thus, we obtain form (4.9), for all $k \in \mathbb{N}$, that

$$0 \leq \frac{\beta}{2}\|w^k - w^{k-1}\|^2 \leq \Gamma\left(w^*, w^{k-1}\right) - \Gamma\left(w^*, w^k\right). \tag{4.18}$$

Therefore, the sequence $\left\{\Gamma\left(w^*, w^k\right)\right\}_{k\in\mathbb{N}}$ is a nonincreasing and nonnegative sequence (see (4.11)), i.e., it is monotone and bounded. Thus, with $B = \sup_{k\in\mathbb{N}}\Gamma(w^*, w^k) \in \mathbb{R}_+$, we have for all $k \in \mathbb{N}$

$$\|w^k\| \leq \|w^* - w^k\| + \|w^*\| \leq \sqrt{2B/\alpha} + \|w^*\| \in \mathbb{R}_+,$$

which proves Item (i).

Next, under the conditions of (ii), we have that $\beta > 0$. From the first item, it follows that $\left\{\Gamma\left(w^*, w^k\right)\right\}_{k\in\mathbb{N}}$ convergent and therefore

$$\lim_{k\to\infty}\left(\Gamma\left(w^*, w^{k-1}\right) - \Gamma\left(w^*, w^k\right)\right) = 0.$$

Thus, it also follows from (4.18) that

$$\lim_{k\to\infty}\|w^k - w^{k-1}\| = 0. \tag{4.19}$$

Let $\{w^{j_n}\}_{j_n\in\mathbb{N}}$ be a convergent subsequence, where $w^\infty = \lim_{n\to\infty} w^{j_n}$. Hence,

$$0 \leq \lim_{n\to\infty}\|w^{j_n-1} - w^\infty\| \leq \lim_{n\to\infty}\|w^{j_n-1} - w^{j_n}\| + \lim_{n\to\infty}\|w^{j_n} - w^\infty\| = 0.$$

From Assumption 4.2(i) combined with (4.9), it follows, for all $w \in W$, that

$$\liminf_{n\to\infty}\Lambda\left(w^{j_n}, w\right) \leq \lim_{n\to\infty}\left(\Gamma\left(w, w^{j_n-1}\right) - \Gamma\left(w, w^{j_n}\right)\right) = 0.$$

Applying Lemma 2.1, we obtain that w^∞ is a saddle-point, i.e., item (ii) is proved.

It remains to show that under the assumptions of item (iii) the sequence $\{w^k\}_{k\in\mathbb{N}}$ has a unique limit point. Indeed, assume that w_a^∞ and w_b^∞ are two limit points of $\{w^k\}_{k\in\mathbb{N}}$, i.e., $w^{i_n} \to w_a^\infty$ and $w^{j_n} \to w_b^\infty$ as $n \to \infty$. Then, due to the previous item, we have that both $w_a^\infty \in W^*$ and $w_b^\infty \in W^*$ are saddle-point of problem (M). Hence, by the same arguments that were made above for w^*, we have that $\{\Gamma\left(w_a^\infty, w^k\right)\}_{k\in\mathbb{N}}$ is a convergent sequence. Specifically, we have

$$\lim_{n\to\infty}\Gamma\left(w_a^\infty, w^{j_n}\right) = \lim_{n\to\infty}\Gamma\left(w_a^\infty, w^{i_n}\right) = 0, \tag{4.20}$$

where the second equality is due to the Assumption 4.2(ii). Therefore, with (4.11), we obtain $\lim_{n\to\infty}\|w^{j_n} - w_a^\infty\| = 0$, which implies that $w_b^\infty = \lim_{n\to\infty} w^{j_n} = w_a^\infty$. This completes the proof of item (iii). □

5. Application: Regularized multinomial logistic regression

To illustrate the advantage and relevance of NEPAPC over the classical PAPC, we consider the Multinomial Logistic Regression (MLR) model and the associated training problem, see, e.g., [27]. We show (after a proper saddle-point reformulation of the problem), the benefits of using a non-Euclidean proximal mapping over the classical prox used in PAPC. Indeed, within NEPAPC a simple explicit formula is obtained, and the numerical illustration confirms the efficiency of NEPAPC over its classical counterpart PAPC.

5.1. **The Problem.** Before describing the problem, we recall some basic notations that will be used below. Given a matrix $M \in \mathbb{R}^{n\times m}$, M_i denotes its ith row and m_j denotes its jth column. The identity matrix is denoted as I. The columns of I, i.e., the standard basis vectors, are denoted as e_i, for $i = 1,2,\ldots,n$. For $A,B \in \mathbb{R}^{n\times m}$, $\langle A,B\rangle = \mathrm{Tr}(A^TB) = \sum_{i=1}^n\sum_{j=1}^m A_{ij}B_{ij}$. When we apply a scalar function φ to a vector $\xi \in \mathbb{R}^n$ (or to any multidimensional array), it is applied elmentwise, e.g., $\varphi(\xi) := (\varphi(\xi_1),\varphi(\xi_2),\ldots,\varphi(\xi_n))^T$.

We consider the Multinomial Logistic Regression (MLR) model and the associated training problem, see, e.g., [27]. Given an observation with a feature vector $\hat{x} \in \mathbb{R}^n$, the MLR model, parameterized by $U \in \mathbb{R}^{n\times q}$, models the conditional probability of the observation's class $\hat{c}$ to be $\hat{l} \in \{1,2,\ldots,q\}$ by

$$P(\hat{c} = \hat{l}|\hat{x};U) = \frac{\exp(\hat{x}^Tu_{\hat{l}})}{\sum_{j=1}^q \exp(\hat{x}^Tu_j)} = \exp\left(\hat{x}^TU\hat{y} - \log\sum_{j=1}^q \exp(\hat{x}^Tu_j)\right)$$
$$= \exp\left(\langle \hat{x}\hat{y}^T, U\rangle - \log\sum_{j=1}^q \exp\left((\hat{x}^TU)_j\right)\right),$$

where $\hat{y} \equiv e_{\hat{l}} \in \mathbb{R}^q$ is the $\hat{l}^{th}$ standard basis vector.

Let $\{(x_i,l_i)\}_{i=1}^m$ be a set of m independent samples, where $x_i \in \mathbb{R}^n$ is the feature vector of sample i, its class c_i equals $l_i \in \{1,2,\ldots,q\}$, and we denote $y_i = e_{l_i} \in \mathbb{R}^q$, for $i = 1,2,\ldots,m$. We set $X = (x_1,x_2,\ldots,x_m) \in \mathbb{R}^{n\times m}$, $c = (c_1,c_2,\ldots,c_m)$, $l = (l_1,l_2,\ldots,l_m)$. Then, the *log-likelihood* of the model parameters U is given by

$$\begin{aligned}\log P(c = l|X;U) &= \log\prod_{i=1}^m P(c_i = l_i|x_i;U)\\ &= \sum_{i=1}^m \log P(c_i = l_i|x_i;U)\\ &= \sum_{i=1}^m \langle x_iy_i^T, U\rangle - \sum_{i=1}^m \log\sum_{j=1}^q \exp\left((x_i^TU)_j\right).\end{aligned} \tag{5.1}$$

We formulate the problem of estimating the model's parameters U in the following standard form

$$\min_{U\in\mathbb{R}^{n\times q}} \{\mu r(U) + \mathrm{loss}(U)\}.$$

As we aim to maximize the *log-likelihood*, we set the loss to be

$$\mathrm{loss}(U) = -\frac{1}{m}\log P(c = l|X;U),$$

where $1/m$ acts as a scaling factor. The regularizer $r(\cdot)$ and the regularization parameter $\mu > 0$ are added in order to impose prior assumptions on U and to cope with the overfitting issues caused by the high dimension of the feature space. In this example, we use a regularizer $r : \mathbb{R}^{n\times q} \to \mathbb{R}_+$, which is defined by

$$r(U) = \sum_{j=1}^{q} \left(\alpha \|Du_j\|_1 + \frac{1-\alpha}{2}\|u_j\|_2^2 \right) = \alpha\|DU\|_1 + \frac{1-\alpha}{2}\|U\|_2^2,$$

where $\alpha \in (0,1)$, $D \in \mathbb{R}^{(n-1)\times n}$ is the matrix of the forward difference linear operator $\mathfrak{D} : \mathbb{R}^n \to \mathbb{R}^{n-1}$ defined by $(\mathfrak{D}z)_i = z_{i+1} - z_i$, $i = 1,2,\ldots,n-1$, and where the norms $\|\cdot\|_1$ and $\|\cdot\|_2$ are the vector norms, i.e., the entrywise l_1 and l_2 norms, respectively. The chosen regularizer can be viewed as a hybrid between the elastic net [28] and a penalized version of the fused lasso [29].

Thus, the training problem translates to the following convex optimization problem, which we refer to as (RMLR),

$$\min_{U\in\mathbb{R}^{n\times q}} \left\{ \Phi(U) := \mu_1\|DU\|_1 + \frac{\mu_2}{2}\|U\|_2^2 - \frac{1}{m}\sum_{i=1}^{m}\langle x_i y_i^T, U\rangle + \frac{1}{m}\sum_{i=1}^{m}\log\sum_{j=1}^{q}\exp\left((x_i^T U)_j\right) \right\},$$

where $\mu_1 = \mu\alpha$ and $\mu_2 = \mu(1-\alpha)$.

5.2. **Min-Max Reformulations of (RMLR) and Our Algorithm.** Throughout, we will use the following notations:

$$\begin{aligned} g(\zeta) :=& \log\sum_{j=1}^{q}\exp(\zeta_j), \\ \mathcal{W} :=& \{W = (w_1,\ldots,w_q) \in \mathbb{R}^{(n-1)\times q} : \quad |W_{ij}| \le 1,\ i = 1,2,\ldots,n-1, j = 1,\ldots,q\}, \\ \mathcal{V} :=& \{V \in \mathbb{R}^{m\times q} : \ \sum_{j=1}^{q} V_{ij} = 1,\ i = 1,\ldots,m,\ V_{ij} \ge 0, i = 1,\ldots,n, j = 1,\ldots,q\}. \end{aligned}$$

First, noting that $\|DU\|_1 = \max\{\langle W, DU\rangle : \ W \in \mathcal{W}\}$, we obtain the following saddle-point formulation of (RMLR), which we refer to as (SRMLR1),

$$\min_{U\in\mathbb{R}^{n\times q}} \max_{W\in\mathcal{W}} \left\{ \frac{\mu_2}{2}\|U\|_2^2 - \frac{1}{m}\sum_{i=1}^{m}\langle x_i y_i^T, U\rangle + \mu_1\langle D^T W, U\rangle + \frac{1}{m}\sum_{i=1}^{m} g(x_i^T U) \right\}.$$

It is well-known that the function g has a Lipschitz continuous gradient, and thus within this formulation we can apply the classical PAPC on the formulation (SRMLR1) by taking the smooth function in model (M) (cf. Section 2) to be

$$f(U) := \frac{\mu_2}{2}\|U\|_2^2 - \frac{1}{m}\sum_{i=1}^{m}\langle x_i y_i^T, U\rangle + \frac{1}{m}\sum_{i=1}^{m} g(x_i^T U).$$

However, within this formulation, the inherent m blocks separable structure of the model is not exploited. As a result, the preconditioning (cf. Section 4) cannot be applied within this formulation, and as we shall see in the numerical experiment given below, this negatively affect the computational performance of PAPC.

An alternative formulation that can exploit the block separable structure of the problem is as follows. The main observation toward this task is based on the well-known fact that the log sum of exponent function is the conjugate of the (negative) entropy function when defined on the

unit simplex. More precisely, let us denote by $\Delta_d := \{x \in \mathbb{R}^d_+ : \sum_{i=1}^n \xi_i = 1\}$, the unit simplex on $\mathbb{R}^d$, and its relative interior by $\Delta_d^+ = \{x \in \mathbb{R}^d_{++} : \sum_{i=1}^n \xi_i = 1\}$. The (negative)-entropy function $h : \mathbb{R}^d_+ \to \mathbb{R}$ is defined as $h(\xi) := \sum_{i=1}^d \xi_i \log(\xi_i)$, (with $0 \log 0 = 0$). Then, the following result follows.

Lemma 5.1. *For any $z \in \mathbb{R}^d$, one has*

$$\log \sum_{j=1}^d \exp(z_j) = \max\{\langle \xi, z\rangle - \sum_{j=1}^d \xi_j \log \xi_j : \ \xi \in \Delta_d\}, \tag{5.2}$$

with the maximum attained at $\mathcal{S}(z) := \left(\sum_{j=1}^d \exp(z_j)\right)^{-1} \exp(z)$.

Proof. See, e.g., [15, p.148] or it simply follows by writing the optimality condition for the convex problem (5.2). □

Applying Lemma 5.1 for the separable sum of functions g in (SRMLR1), we then obtain the following alternative saddle-point reformulation, which we refer to as (SRMLR2),

$$\min_{U \in \mathbb{R}^{n\times q}} \max_{V \in \mathcal{V}, W \in \mathcal{W}} \left\{ \frac{\mu_2}{2}\|U\|_2^2 - \frac{1}{m}\sum_{i=1}^m \langle x_i y_i^T, U\rangle + \langle \frac{1}{m}\sum_{i=1}^m x_i V_i + \mu_1 D^T W, U\rangle - \frac{1}{m}\sum_{i=1}^m \sum_{j=1}^q V_{i,j} \log V_{i,j} \right\}.$$

The V-block now decomposes nicely, but applying the classical PAPC would require to compute for each block V_i the usual Euclidean prox of the entropy function over the simplex Δ_q. However, this task cannot be done explicitly and thus would imply a nested optimization loop requiring to implement a numerical procedure. On the other hand, exploiting the geometry of the constraint $\mathcal{V}$ described here by a unit simplex, we can instead naturally apply NEPAPC, by using the so-called *Kullback-Leibler (KL)* distance which is obtained by using the entropy function h to define the corresponding the Bregman distance $D_h(\cdot,\cdot)$ on $\Delta_d \times \Delta_d^+$, by

$$D_h(\xi, \eta) = \sum_{i=1}^d \xi_i \log\left(\frac{\xi_i}{\eta_i}\right) = h(\xi) - \langle \log(\eta), \xi\rangle. \tag{5.3}$$

Indeed, equipped with this D_h, computing the V_i step in NEPAPC, reduces to solving the following optimization problem, which is shown to admit a simple explicit formula.

Lemma 5.2. *For any $z \in \mathbb{R}^d, \eta \in \Delta_d^+$, and D_h as defined in (5.3), we have*

$$v^+ := \operatorname{argmin}_{\xi \in \Delta_d} \left\{h(\xi) + \langle z, \xi\rangle + \rho^{-1} D_h(\xi, \eta)\right\} = \mathcal{S}\left((t-1)z + t\log(\eta)\right),$$

with $t = (1+\rho)^{-1}$.

Proof. Simple algebra shows that finding v^+ consists of solving the convex minimization problem

$$\min\left\{\sum_{i=1}^d \xi_i \log \xi_i - (\frac{1}{\rho+1}\log \eta_i - \frac{\rho}{\rho+1} z_i)\xi_i : \ \xi \in \Delta_d\right\}.$$

Invoking Lemma 5.1, and setting $t = (1+\rho)^{-1}$, we immediately obtain the claimed formula for the minimizer v^+. □

Thus, we can apply NEPAPC. More precisely, we will apply the block version of NEPAPC with preconditioning as described in Algorithm 2. For $V_1, V_2, \ldots, V_m$, we set the *proximal distance* to be the Bregman distance D_h, with $h(x) = \sum_{j=1}^{q} x_j \log(x_j)$ and $\operatorname{dom} h = \mathbb{R}^q_+$. For W we use the standard squared Euclidean distance. Note that in this case we have $L = \mu_2$. We find that it is computationally effective to set the preconditioning parameters to be m for V and $1/\mu_1$ for W; that is (cf. Section 4), we set here $\mathcal{A}_\omega = (X, D^T)$, and obtain the following algorithm.

Algorithm 3 NEPAPC for SRMLR2 with preconditioning

Initialization. $\mathcal{A}_\omega = (X, D^T)$, $\tau \le 1/\mu_2$, $\sigma \le 1/(\tau \|\mathcal{A}_\omega\|^2)$, $t = 1/(1+m\sigma)$, $U^0 \in \mathbb{R}^{n\times q}$, $V_i^0 \in \Delta^+_{1\times q}$, for $i = 1,2,\ldots,m$, $W_{i,j}^0 \in [-1,1]$, for $i = 1,2,\ldots,n-1$ and $j = 1,2,\ldots,q$.
General step. For $k = 1,2,\ldots$ compute:

$$
\begin{aligned}
P^k &= (1-\tau\mu_2)U^{k-1} + \frac{\tau}{m}\sum_{i=1}^{m} x_i y_i^T - \frac{\tau}{m}\sum_{i=1}^{m} x_i V_i^{k-1} - \tau\mu_1 D^T W^{k-1},\\
V_i^k &= \mathcal{S}\left((1-t)x_i^T P^k + t\log V_i^{k-1}\right), \quad i = 1,2,\ldots,m,\\
W_{i,j}^k &= \mathcal{P}_{[-1,1]}\left(W_{i,j}^{k-1} + \frac{\sigma}{\mu_1}(DP^k)_{i,j}\right), \quad i = 1,2,\ldots,m,\\
U^k &= (1-\tau\mu_2)U^{k-1} + \frac{\tau}{m}\sum_{i=1}^{m} x_i y_i^T - \frac{\tau}{m}\sum_{i=1}^{m} x_i X V_i^k - \tau\mu_1 D^T W^k,
\end{aligned}
$$

where $\mathcal{P}_{[-1,1]}(s)$ denotes the projection of $s \in \mathbb{R}$ onto the interval $[-1,1]$.

5.3. **Numerical Experiment.** We have conducted a synthetic numerical experiment, similar to that of [12, Example 2]. The parameters of the model and the algorithm were set as follows: $n = 1000$, $q = 50$, $m = 5000$, where for each class we have generated 100 samples, $\mu = 1e-6$, and $\alpha = 0.5$. The model's parameter matrix U was predetermined, the samples x_i were randomly generated and each sample was randomly assigned a class according to probabilities which were computed by the MLR model.

We applied PAPC [12] on problem (SRMLR1) and the block version of NEPAPC (see Algorithm 3) on problem (SRMLR2). We measured the objective value of the primal problem (RMLR). The lowest generated value was regarded as the optimal value. The results are summarized in the following two figures. Figure 1 demonstrates that NEPAPC, that has the ability to adapt to the structure of the problem, clearly outperforms PAPC, and shows the advantage of using Non-Euclidean distances. In Figure 2, we show the performance of the main and ergodic sequences generated by NEPAPC. As can be seen, the main sequence performs much better and seems to converge at a linear rate. This phenomena was also observed in PAPC [12], and therefore it would be interesting, in a future research, to tackle the theoretical guarantees of the main sequence of these algorithms.

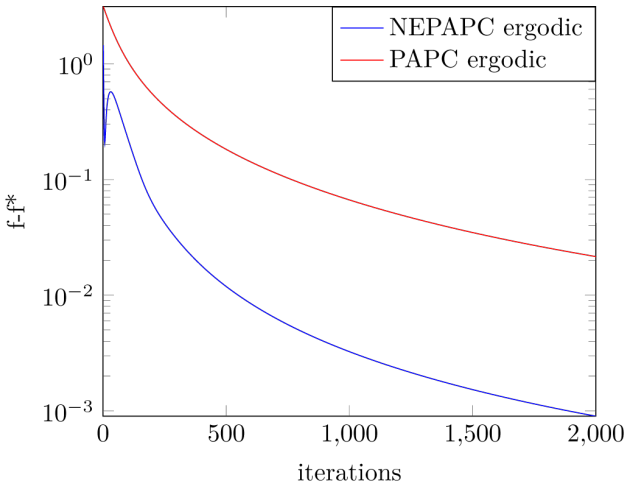

FIGURE 1. Objective values: ergodic PAPC vs. ergodic NEPAPC

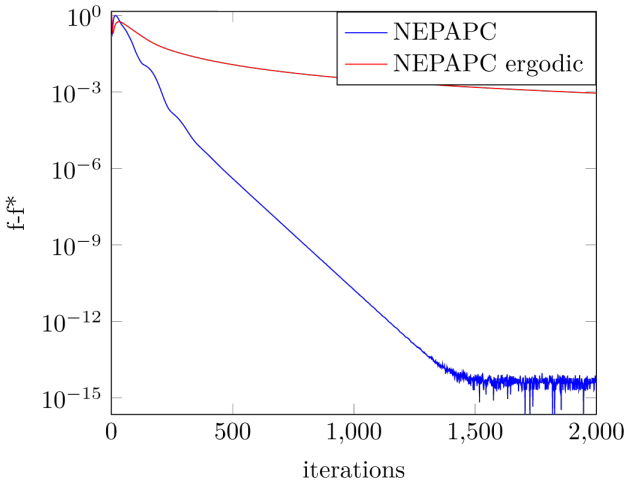

FIGURE 2. Objective values: Sequence vs. ergodic sequence in NEPAPC

6. CONCLUDING REMARKS

The PAPC algorithm can tackle a broad class of structured optimization models that include, for example, block linear constraints as well as models with a finite sum of the composition of non-smooth functions with linear mappings in the objective or in the constraints. The PAPC algorithm tackles such models by fully decomposing them into simple algorithmic steps that avoid the difficult task of computing the proximal map of the composition of a convex function with a linear map, and instead requires computing the proximal map of the given function, and this is the main computational step involved in PAPC. However, in many instances, computing the proximal map of a given convex function, can also be a difficult task. To overcome this difficulty, in this paper we proposed a *Non-Euclidean* version of PAPC (NEPAPC), which includes PAPC as a special case. By non-Euclidean here, we mean that the classical Moreau's proximal mapping is replaced with a general proximal map whereby the classical squared norm can be replaced by a broad family of proximal distances (which includes the classical squared

norm, as a special case.) NEPAPC is proven to preserve the same convergence properties of PAPC, but has the ability to adapt well to the geometry of the objective or the constraints in a given saddle-point problem. Indeed, in such cases, NEPAPC allows to simplify the proximal computational step through the use of suitable proximal distances and maps which exploit the problem's data at hand. It produces a simple and efficient scheme with explicit iterative formula which could not be obtained otherwise, and thus further expands the scope of applications of the PAPC method.

Acknowledgments

This research was partially supported by the German Research Foundation under DFG Grant 800240. The first author was supported by a Ph.D fellowship under ISF Grants 1844-16 and 2619-20. The third author was partially supported by the Israel Science Foundation, under ISF Grants 1844-16 and 2619-20.

REFERENCES

[1] G. M. Korpelevič, An extragradient method for finding saddle-points and for other problems, Èkonom. i Mat. Metody, 12 (1976), 747–756.

[2] A. Nemirovski, Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle-point problems, SIAM J. Optim. 15 (2004), 229–251.

[3] A. Auslender, M. Teboulle, Interior projection-like methods for monotone variational inequalities, Math. Program. Ser. A 104 (2005), 39–68.

[4] P. L. Lions, B. Mercier, Splitting algorithms for the sum of two nonlinear operators, SIAM J. Numer. Anal. 16 (1979), 964–979.

[5] G. B. Passty, Ergodic convergence to a zero of the sum of monotone operators in Hilbert space, J. Math. Anal. Appl. 72 (1979), 383–390.

[6] D. Gabay, Applications of the Method of Multipliers to Variational Inequalities, In: M. Fortin and R. Glowinski, (ed.) Studies in Mathematics and Its Applications, vol. 15, pp. 299–331, Elsevier, Amsterdam, 1983.

[7] R. Glowinski, P. Le Tallec, Augmented Lagrangian and operator splitting methods in nonlinear mechanics, Society for Industrial Mathematics, Philadelphia, 1989.

[8] S. Sabach, M. Teboulle, Lagrangian methods for composite optimization, In: R. Kimmel, X.C. Tai, (ed.) Handbook of Numerical Analysis, vol. 20, pp. 401–436, Elsevier, Amsterdam, 2019.

[9] Y. Nesterov, Smooth minimization of non-smooth functions, Math. Program. Ser. A 103) (2005), 127–152.

[10] A. Beck, M. Teboulle, Smoothing and first order methods: a unified framework, SIAM J. Optim. 22 (2012), 557–580.

[11] A. Chambolle, T. Pock, An introduction to continuous optimization for imaging, Acta Numer. 25 (2016), 161–319.

[12] Y. Drori, S. Sabach, M. Teboulle, A simple algorithm for a class of nonsmooth convex–concave saddle-point problems, Oper. Res. Lett. 43 (2015), 209–214.

[13] J. J. Moreau, Proximité et dualité dans un espace hilbertien, Bulletin de la Société mathématique de France, 93 (1965), 273–299.

[14] A. Auslender, M. Teboulle, Interior gradient and proximal methods for convex and conic optimization, SIAM J. Optim. 16 (2006), 697–725.

[15] R. Rockafellar, Convex Analysis, Princeton Univ. Press, Princeton, New Jersey, 1970.

[16] R. Rockafellar, J. Wets, Variational Analysis, Springer, New York, 2004.

[17] A. Auslender, M. Teboulle, Asymptotic Cones and Functions in Optimization and Variational Inequalities, Springer Monographs in Mathematics, Springer, New York, 2003.

[18] A. Nemirovsky, D. Yudin, Problem complexity and method efficiency in optimization, A Wiley-Interscience Publication, John Wiley & Sons Inc., New York, 1983. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics.
[19] L. M. Bregman, The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming, USSR Comput. Math. Math. Phys. 7 (1967), 200–217.
[20] Y. Censor, S. A. Zenios, Proximal minimization algorithm withd-functions, J. Optim. Theory Appl. 73 (1992), 451–464.
[21] M. Teboulle, Entropic proximal mappings with applications to nonlinear programming, Math. Oper. Res. 17 (1992), 670–690.
[22] G. Chen, M. Teboulle, Convergence analysis of a proximal-like minimization algorithm using bregman functions, SIAM J. Optim. 3 (1993), 538–543.
[23] J. Eckstein, Nonlinear proximal point algorithms using bregman functions, with applications to convex programming, Math. Oper. Res. 18 (1993), 202–226.
[24] M. Teboulle, A simplified view of first order methods for optimization, Math. Program. 170 (2018), 67–96.
[25] P. P. B. Eggermont, Multiplicative iterative algorithms for convex programming, Linear Algebra Appl. 130 (1990), 25–42.
[26] M. Teboulle, Convergence of proximal-like algorithms, SIAM J. Optim. 7 (1997), 1069–1083.
[27] S. Gopal, Y. Yang, Distributed training of large-scale logistic models, International Conference on Machine Learning, PMLR, pp. 289–297, 2013.
[28] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc. Ser. B. Stat. Methodol. 67 (2005), 301–320.
[29] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, K. Knight, Sparsity and smoothness via the fused lasso, J. R. Stat. Soc. Ser. B. Stat. Methodol. 67 (2005), 91–108.