# Optimization of Inf-Convolution Regularized Nonconvex Composite Problems

**Emanuel Laude** **Tao Wu** **Daniel Cremers**
Department of Informatics, Technical University of Munich, Germany

## Abstract

In this work, we consider nonconvex composite problems that involve inf-convolution with a *Legendre* function, which gives rise to an anisotropic generalization of the proximal mapping and Moreau-envelope. In a convex setting such problems can be solved via alternating minimization of a splitting formulation, where the consensus constraint is penalized with a Legendre function. In contrast, for nonconvex models it is in general unclear that this approach yields stationary points to the infimal convolution problem. To this end we analytically investigate local regularity properties of the Moreau-envelope function under prox-regularity, which allows us to establish the equivalence between stationary points of the splitting model and the original inf-convolution model. We apply our theory to characterize stationary points of the penalty objective, which is minimized by the *elastic averaging SGD* (EASGD) method for distributed training. Numerically, we demonstrate the practical relevance of the proposed approach on the important task of distributed training of deep neural networks.

## 1 Introduction

In this work, we are interested in optimizing nonconvex composite models which involve *infimal convolutions* with *Legendre* functions:

$$\underset{u\in\mathbb{R}^n}{\text{minimize}} \quad e_\lambda^\phi f(Au) + g(u). \tag{1}$$

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89. 

Here both $f : \mathbb{R}^m \to \overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty, -\infty\}$ and $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ are extended real-valued, proper[1] and lower semi-continuous (lsc) functions which are possibly nonconvex and nonsmooth, and $A \in \mathbb{R}^{m\times n}$ is a coupling matrix. Let $\operatorname{dom} f := \{z \in \mathbb{R}^m : f(z) < \infty\}$ denote the domain of $f$. By $e_\lambda^\phi f$ we denote the infimal convolution of a function $f$ with some Legendre function $\phi : \mathbb{R}^m \to \overline{\mathbb{R}}$; see Definition 3.1. The infimal convolution of $f$ with a potential $\phi$ and scaling parameter $\lambda$ is defined as

$$e_\lambda^\phi f(v) = \inf_{z\in\mathbb{R}^m} f(z) + \frac{1}{\lambda}\phi(v - z). \tag{2}$$

We shall refer to $e_\lambda^\phi f$ as the $\phi$-envelope of $f$, and the corresponding arg min map

$$P_\lambda^\phi f(v) = \operatorname*{arg\,min}_{z\in\mathbb{R}^m} f(z) + \frac{1}{\lambda}\phi(v - z), \tag{3}$$

as the $\phi$-proximal mapping of $f$ at $v$.

Note that for $\phi = \frac{1}{2}\|\cdot\|^2$ they specialize to the classical Moreau-envelope and proximal mapping [16, 17].

Under suitable assumptions that guarantee that the inf is attained when finite, $e_\lambda^\phi f$ yields a regularized variant of $f$ in the sense that the epigraph of $e_\lambda^\phi f$ is obtained via the Minkowski sum of the epigraphs of the individual functions $f$ and $\phi$ [22, Exercise 1.28]. For convex proper lsc $f$ and Lipschitz differentiable[2] $\phi$, $e_\lambda^\phi f$ is a Lipschitz differentiable approximation to $f$. In contrast, when $f$ is nonconvex and nonsmooth, $e_\lambda^\phi f$ remains nonsmooth and nonconvex in general which renders the optimization of (1) challenging.

Inf-convolution models are well grounded in machine learning and signal processing. A variety of convex and nonconvex loss functions and regularizers can be written as an infimal convolution. There, the potential $\phi$ is chosen in accordance with the underlying noise prior,

[1] A function $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ is called proper if $f(\bar{z}) < \infty$ for some $\bar{z} \in \mathbb{R}^m$ and $f(z) > -\infty$ for all $z \in \mathbb{R}^m$.

[2] A function is called Lipschitz differentiable if it is differentiable and its gradient is Lipschitz continuous.

e.g., quadratic for Gaussian. In addition, benefits of inf-convolution regularization (with quadratic $\phi$) have been observed empirically in neural network training in recent works [27, 8].

### 1.1 Motivation

To compute a stationary point of (1), we resort to a splitting model

$$\underset{u\in\mathbb{R}^n,\ z\in\mathbb{R}^m}{\text{minimize}}\ F(u,z) := f(z) + \frac{1}{\lambda}\phi(Au - z) + g(u), \tag{4}$$

where the violation of the constraint $Au = z$ is penalized with $\phi$. From Equation (2) it can be seen that model (4) is equivalent to model (1) in terms of global optima but in general not in terms of stationary points[3]. The splitting formulation (4) is amenable to alternating optimization, which, under mild assumptions, converges subsequentially to a stationarity point $(\bar{u}, \bar{z})$ of $F$, satisfying the following conditions (assuming $\operatorname{dom}\phi$ is open): $A\bar{u} - \bar{z} \in \operatorname{dom}\phi$, $\bar{u} \in \operatorname{dom} g$, $\bar{z} \in \operatorname{dom} f$, and

$$0 \in \partial g(\bar{u}) + \tfrac{1}{\lambda}A^\top\nabla\phi(A\bar{u} - \bar{z}), \tag{5a}$$
$$0 \in \partial f(\bar{z}) - \tfrac{1}{\lambda}\nabla\phi(A\bar{u} - \bar{z}). \tag{5b}$$

Here $\partial$ denotes the (limiting) subdifferential of a function, cf. Definition 4.1.

Meanwhile, in order for $\bar{u}$ to qualify as a stationary point of the original problem (1) it must satisfy

$$0 \in \partial(e_\lambda^\phi f \circ A + g)(\bar{u}). \tag{6}$$

When $e_\lambda^\phi f$ is smooth around $A\bar{u}$ and $\bar{u} \in \operatorname{dom} g$, the stationarity condition (6) simplifies to

$$0 \in A^\top\nabla e_\lambda^\phi f(A\bar{u}) + \partial g(\bar{u}), \tag{7}$$

via [22, Exercise 8.8 (c)].

It is important to realize that conditions (5a)–(5b) do *not* imply (6) or (7) in general when $f$ is nonconvex. This stands in stark contrast to the convex setting, where the stationarity condition (7) (for quadratic $\phi = \frac{1}{2}\|\cdot\|^2$) can be guaranteed via the well known gradient formula for the Moreau-envelope [22, Theorem 2.26]:

$$\nabla e_\lambda^{\|\cdot\|^2/2} f(v) = \frac{1}{\lambda}(v - P_\lambda^{\|\cdot\|^2/2} f(v)). \tag{8}$$

To this end, note that (5b) resembles the necessary (and in the convex setting also sufficient) optimality condition of the $\phi$-proximal mapping $\bar{z} = P_\lambda^\phi(A\bar{u})$. For this reason, a main focus of this work is to derive sufficient conditions that guarantee the *translation of stationarity* in the more general nonconvex setting for (nonquadratic) $\phi$, which ultimately boils down to the following implication:

$$\text{(5b)} \ \Rightarrow\ \tfrac{1}{\lambda}\nabla\phi(A\bar{u} - \bar{z}) = \nabla e_\lambda^\phi f(A\bar{u}). \tag{9}$$

In this sense, the implication in (9) is a generalization of the gradient formula (8) for the the $\phi$-envelope.

### 1.2 Contributions

Our contributions are summarized as follows:

- We consider an anisotropic generalization of the proximal mapping and Moreau-envelope [14, 26, 9] induced by a Legendre function in the nonconvex setting. More precisely we establish local regularity properties of the envelope function and proximal mapping under prox-regularity, including a generalization of the well known gradient formula for the Moreau-envelope. The translation of stationarity is a consequence of this theory.
- We apply our theory to characterize stationary points of the model that is minimized by the elastic averaging SGD (EASGD) [27] method for distributed training with anisotropic (i.e., nonquadratic) penalty functions. There, our theory can be invoked to obtain a robust measure of stationarity via the gradient of the Moreau-envelope.
- Numerically, we apply our algorithm to distributed training of deep neural networks and showcase merits of anisotropic inf-convolution potentials over standard quadratic in this context.

## 2 Related Work

Proximal mappings and Moreau-envelopes date back to the seminal papers of Moreau [16, 17].

[14, 26, 9] consider an anisotropic generalization of the proximal mapping which is obtained by replacing the quadratic penalty with a Legendre function and study its properties in a convex setting. [9] relates the anisotropic proximal mapping to the Bregman proximal mapping (introduced in [7, 24] and investigated in [3]) via a generalization of Moreau's decomposition [16, 17], which holds for convex functions. The Bregman prox and the anisotropic prox are different generalizations of the classical prox with complementary properties.

In the convex setting the Moreau-envelope has strong regularity properties such as Lipschitz differentiability. In the nonconvex setting the Moreau-envelope is

[3] A stationary point $(\bar{u}, \bar{z})$ of (4) is a point that satisfies the necessary first order optimality condition $0 \in \partial F(\bar{u}, \bar{z})$, cf. Fermat's rule generalized [22, Theorem 10.1]. When $(\bar{u}, \bar{z}) \in \operatorname{dom} F$ is feasible and $\phi$ is continuously differentiable on $\operatorname{dom}\phi$ open, $0 \in \partial F(\bar{u}, \bar{z})$ is implied by (5a)–(5b) via [22, Exercise 8.8 (c) and Proposition 10.5].

nonsmooth in general. In [19, 22] the concept of prox-regular functions is introduced, which allows the authors to (locally) recover some of the properties known from the convex setting: These include the (local) single-valuedness of the prox and (local) Lipschitz differentiability of the envelope.

In [15, 18], prox-regularity is used to establish a local convergence result for alternating and averaged projections methods. Similar to this work but specialized to quadratic $\phi$, in [12], prox-regularity is utilized to show a translation of stationarity for the model (1) which is computationally resolved by multiblock ADMM.

More recently, in [10] the gradient of the Moreau-envelope has been used as a stationarity measure in stochastic optimization methods.

Inf-convolution regularization has recently been utilized in neural network training [27, 8]: In [27] the authors have considered the consensus training of deep neural networks by optimizing a relaxed consensus model of the form (4) with quadratic $\phi$. Similar algorithms were later connected to partial differential equations [8].

## 3 Anisotropic Proximal Mapping

In this section, we introduce the notion of Legendre functions that gives rise to an anisotropic generalization of the proximal mapping and Moreau-envelope investigated in the convex setting by [14, 26, 9]. We establish a sufficient condition for the well-definedness of the anisotropic prox (denoted as $\phi$-prox) and envelope (denoted as $\phi$-envelope) in the nonconvex setting, based on a generalized notion of prox-boundedness [22, Definition 1.23].

A proper convex lsc Legendre function is defined below according to [21, Section 26]. Here $\partial\phi$ reduces to the classical convex subdifferential.

**Definition 3.1** (Legendre function)**.** *The proper convex lsc function $\phi : \mathbb{R}^m \to \overline{\mathbb{R}}$ is*

(i) essentially smooth*, if the interior of the domain of $\phi$ is nonempty, i.e.* $\operatorname{int}\operatorname{dom}\phi \neq \emptyset$*, and $\phi$ is differentiable on* $\operatorname{int}\operatorname{dom}\phi$ *such that* $\|\nabla\phi(w^\nu)\| \to \infty$ *whenever* $w^\nu \to w \in \operatorname{bdry}\operatorname{dom}\phi$*;*

(ii) essentially strictly convex*, if $\phi$ is strictly convex on every convex subset of* $\operatorname{dom}\partial\phi := \{w \in \mathbb{R}^m : \partial\phi(w) \neq \emptyset\}$*;*

(iii) Legendre*, if $\phi$ is both essentially smooth and essentially strictly convex.*

Let $\phi^*$ denote the convex conjugate of $\phi$. Then, Legendre functions have the following essential properties:

**Lemma 3.2.** *Let $\phi : \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper lsc convex and Legendre. Then $\phi$ has the following properties:*

(i) $\operatorname{dom}\partial\phi = \operatorname{int}\operatorname{dom}\phi$*, [21, Theorem 26.1].*

(ii) $\nabla\phi : \operatorname{int}\operatorname{dom}\phi \to \operatorname{int}\operatorname{dom}\phi^*$ *is bijective with inverse* $\nabla\phi^* : \operatorname{int}\operatorname{dom}\phi^* \to \operatorname{int}\operatorname{dom}\phi$ *with both $\nabla\phi$ and $\nabla\phi^*$ continuous on* $\operatorname{int}\operatorname{dom}\phi$ *resp.* $\operatorname{int}\operatorname{dom}\phi^*$*, [21, Theorem 26.5].*

Overall we will make the following (additional) assumptions on $\phi$:

(A1) $\phi : \mathbb{R}^m \to \overline{\mathbb{R}}$ is proper lsc convex and Legendre.

(A2) $\operatorname{dom}\phi$ is open.

(A3) $\phi$ is twice continuously differentiable on $\operatorname{int}\operatorname{dom}\phi$ with positive definite Hessian, i.e., $\nabla^2\phi(w) \succ 0$ for any $w \in \operatorname{int}\operatorname{dom}\phi$.

(A4) $\phi$ is *super-coercive*, i.e., $\|\phi(w)\|/\|w\| \to \infty$ whenever $\|w\| \to \infty$.

(A5) $\phi(0) = 0$ and $\nabla\phi(0) = 0$.

(A2) ensures that $\phi(w^\nu) \to \infty$, whenever $w^\nu \to w \in \operatorname{bdry}\operatorname{dom}\phi$. In particular this allows us to utilize alternating gradient descent steps as updates in our algorithm, cf. Section 5. (A3) implies that $\phi$ is "locally" strongly convex and Lipschitz differentiable in the sense of [4, Proposition 2.10]: For any compact and convex $K \subset \operatorname{dom}\partial\phi$, there are constants $\mu, \gamma > 0$ such that for any $w_1, w_2 \in K$:

$$\phi(w_1) \geq \phi(w_2) + \langle \nabla\phi(w_2), w_1 - w_2\rangle + \frac{\mu}{2}\|w_1 - w_2\|^2,$$
$$\|\nabla\phi(w_1) - \nabla\phi(w_2)\| \leq \gamma\|w_1 - w_2\|.$$

Such functions are known under the term *very strictly convex* [4, Definition 2.8] which lie "strictly between the class of strongly convex and the class of strictly convex functions", [4, Remark 2.9]. (A4) is required later on in Section 4 to show the translation of stationarity. (A5) is technically not required in our theory. However, it naturally leads to a smoothing which under prox-regularity preserves stationarity (see Corollary 5.1).

Examples for such $\phi$ include the scaled quadratic $\phi(w) = w^\top Q w$ (with matrix $Q$ symmetric positive definite) or a log-barrier function $\phi(w) = -\log(1-\|w\|^2)$. Further examples are provided in Section 6.

It is important to realize that for nonconvex $f$ the well-definedness of the $\phi$-proximal mapping and envelope requires additional assumptions: More precisely, we shall guarantee that $e_\lambda^\phi f$ is proper and $P_\lambda^\phi f(v) \neq \emptyset$

for any $v \in \operatorname{dom} e_\lambda^\phi f$. In addition, this condition allows us to extract a continuity property for the $\phi$-proximal mapping and envelope which is extensively needed later on to prove the desired translation of stationarity in Section 4:

**Definition 3.3** ($\phi$-prox-boundedness)**.** *We say $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ is $\phi$-prox-bounded if there exists $\lambda > 0$ such that for any $\bar{v} \in \mathbb{R}^m$ there exists $\epsilon > 0$ and a constant $\beta > -\infty$ such that*

$$e_\lambda^\phi f(v) \geq \beta \tag{10}$$

*for any $v$ with $\|v - \bar{v}\| \leq \epsilon$. The supremum of the set of all such $\lambda$ is the threshold $\lambda_f$ of the $\phi$-prox-boundedness.*

When $f$ is bounded from below it is $\phi$-prox-bounded with threshold $\lambda_f = \infty$. Notably, in the classical case (when $\phi$ is quadratic) the definition can be made minimalistic, cf. [22, Definition 1.23]: It suffices to assume the existence of some $\bar{v} \in \mathbb{R}^m$ so that $e_\lambda^\phi f(\bar{v}) > -\infty$.

Overall we summarize below the properties of $\phi$-prox and envelope under $\phi$-prox-boundedness that shall be used along our course in the next section.

**Lemma 3.4.** *Let $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ be proper lsc and $\phi$-prox-bounded with threshold $\lambda_f > 0$. Then for any $\lambda \in (0, \lambda_f)$, $P_\lambda^\phi f$ and $e_\lambda^\phi f$ have the following properties:*

(i) *$P_\lambda^\phi f(v) \neq \emptyset$ is compact for all $v \in \operatorname{dom} e_\lambda^\phi f = \operatorname{dom} f + \operatorname{dom} \phi$, whereas $P_\lambda^\phi f(v) = \emptyset$ for $v \notin \operatorname{dom} e_\lambda^\phi f$.*

(ii) *The $\phi$-envelope $e_\lambda^\phi f$ is continuous relative to $\operatorname{dom} e_\lambda^\phi f$.*

(iii) *For any sequence $v^\nu \to \bar{v}$ contained in $\operatorname{dom} e_\lambda^\phi f$ and $z^\nu \in P_\lambda^\phi f(v^\nu)$ we have $\{z^\nu\}_{\nu \in \mathbb{N}}$ is bounded and all its cluster points $\bar{z}$ lie in $P_\lambda^\phi f(\bar{v})$.*

## 4 Translation of Stationarity

In this section we prove the translation of stationarity, see (9), under prox-regularity: We emphasize that it is a major concern of the splitting approach in the nonconvex setting to justify whether solving the splitting model (4) guarantees solving the original model (1), both in terms of stationarity.

To this end we recall the definitions of the regular and the limiting subdifferential according to [22, Definition 8.3].

**Definition 4.1** (subdifferential)**.** *Let $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ and $\bar{z} \in \operatorname{dom} f$ be given. For $y \in \mathbb{R}^m$, we say*

(i) *$y$ is a regular subgradient of $f$ at $\bar{z}$, written $y \in \widehat{\partial} f(\bar{z})$, if*

$$\liminf_{\substack{z \to \bar{z} \\ z \neq \bar{z}}} \frac{f(z) - f(\bar{z}) - \langle y, z - \bar{z} \rangle}{\|z - \bar{z}\|} \geq 0.$$

*We refer to the set $\widehat{\partial} f(\bar{z})$ as the regular subdifferential of $f$ at $\bar{z}$.*

(ii) *$y$ is a (limiting) subgradient of $f$ at $y$, written $y \in \partial f(\bar{z})$, if there exist $z^\nu \to \bar{z}$ with $f(z^\nu) \to f(\bar{z})$ and $y^\nu \in \widehat{\partial} f(z^\nu)$ with $y^\nu \to y$. We refer to the set $\partial f(\bar{z})$ as the (limiting) subdifferential of $f$ at $\bar{z}$.*

We remark that for $f$ convex, both the regular and the limiting subdifferential coincide with the classical convex subdifferential, [22, Proposition 8.12].

Next, we define prox-regularity of functions, according to [22, Definition 13.27]:

**Definition 4.2** (prox-regularity of functions)**.** *Assume $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ is lsc and finite at $\bar{z} \in \mathbb{R}^m$. We say $f$ is prox-regular at $\bar{z}$ for $\bar{y} \in \partial f(\bar{z})$ if there exist $\epsilon > 0$ and $r \geq 0$ such that for all $\|z' - \bar{z}\| < \epsilon$*

$$f(z') \geq f(z) + \langle y, z' - z \rangle - \frac{r}{2}\|z' - z\|^2, \tag{11}$$

*whenever $\|z - \bar{z}\| < \epsilon$, $f(z) - f(\bar{z}) < \epsilon$, $y \in \partial f(z)$, $\|y - \bar{y}\| < \epsilon$. When this holds for all $\bar{y} \in \partial f(\bar{z})$, $f$ is said to be prox-regular at $\bar{z}$.*

Prox-regularity is a local property in nature. Examples for (everywhere) prox-regular functions include: (i) proper, (weakly) convex, lsc functions; (ii) $\mathcal{C}^2$-functions; and (iii) indicator functions of $\mathcal{C}^2$-manifolds [19, 22]. For further examples we refer to [19, 22].

Based on prox-regularity and $\phi$-prox-boundedness, we now extend [22, Proposition 13.37] to $P_\lambda^\phi f$ and $e_\lambda^\phi f$ in our context (see Theorem 4.3) and eventually derive the translation of stationarity as desired (see Corollary 4.4). As a key ingredient in the proof of Theorem 4.3 we may invoke the generalized implicit function theorem [20, Theorem 2.1] [11, Theorem 2B.5] to assert $P_\lambda^\phi f$ is locally a single-valued, Lipschitz map.

**Theorem 4.3.** *Let $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ proper lsc and $\phi$-prox-bounded with threshold $\lambda_f$. Let $\bar{v} \in \bar{z} + \operatorname{dom} \phi$. Then for any $\lambda \in (0, \lambda_f)$ sufficiently small and $f$ finite and prox-regular at $\bar{z}$ for $\bar{y} \in \partial f(\bar{z})$ with*

$$\bar{y} = \frac{1}{\lambda} \nabla \phi(\bar{v} - \bar{z})$$

*the following statements hold true:*

(i) $P_\lambda^\phi f$ *is a singled-valued, Lipschitz map near* $\bar{v}$ *such that* $\bar{z} = P_\lambda^\phi f(\bar{v})$ *and*

$$P_\lambda^\phi f(v) = (I + \nabla\phi^* \circ \lambda T)^{-1}(v), \tag{12}$$

*where* $T$ *is the* $f$*-attentive* $\epsilon$*-localization of* $\partial f$ *near* $(\bar{z}, \bar{y})$*, i.e. the set-valued mapping* $T : \mathbb{R}^m \rightrightarrows \mathbb{R}^m$ *defined by* $T(z) := \{y \in \partial f(z) : \|y - \bar{y}\| < \epsilon\}$ *if* $\|z - \bar{z}\| < \epsilon$ *and* $f(z) < f(\bar{z}) + \epsilon$*, and* $T(z) := \emptyset$ *otherwise.*

(ii) $e_\lambda^\phi f$ *is Lipschitz differentiable around* $\bar{v}$ *with*

$$\nabla e_\lambda^\phi f(v) = \frac{1}{\lambda}\nabla\phi(v - z). \tag{13}$$

*Proof.* We provide a short proof sketch for part (i). For a detailed proof of part (i) and part (ii) we refer to the supplements.

(i) Using the definition of prox-regularity, the assumptions in the theorem and the continuity property of the prox from Lemma 3.4 (which holds under $\phi$-prox-boundedness) it can be proven that for some $\lambda \in (0, \lambda_f)$ sufficiently small for any $\xi$ sufficiently near 0 we have $\xi \in T(z) - \frac{1}{\lambda}\nabla\phi(\bar{v} - z)$, for some $z$ near $\bar{z}$. Furthermore, using the definition of prox-regularity and Assumption (A3), it can be shown that $T - \frac{1}{\lambda}\nabla\phi(\bar{v} - \cdot)$ is strongly monotone.

This implies that $\xi \mapsto \left(T - \frac{1}{\lambda}\nabla\phi(\bar{v} - \cdot)\right)^{-1}(\xi)$ is a single-valued, Lipschitz map in a neighborhood of 0 such that $\left(T - \frac{1}{\lambda}\nabla\phi(\bar{v} - \cdot)\right)^{-1}(0) = \bar{z}$. Invoking the generalized implicit function theorem [11, Theorem 2B.7], we assert that $v \mapsto P_\lambda^\phi f(v) = \left(T - \frac{1}{\lambda}\nabla\phi(v - \cdot)\right)^{-1}(0)$ is a single-valued, Lipschitz map in a neighborhood of $\bar{v}$ such that $\bar{z} = P_\lambda^\phi f(\bar{v})$. □

As an immediate consequence of the above theorem, the implication in (9) holds true and the translation of stationarity is attained under prox-regularity.

**Corollary 4.4** (translation of stationarity)**.** *Let* $(\bar{u}, \bar{z})$ *be a stationary point for the splitting model* (4) *satisfying* $A\bar{u} - \bar{z} \in \operatorname{dom}\phi$*,* $\bar{u} \in \operatorname{dom} g$*,* $\bar{z} \in \operatorname{dom} f$ *and conditions* (5a)–(5b)*. Let* $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ *be* $\phi$*-prox-bounded with threshold* $\lambda_f$ *and prox-regular at* $\bar{z}$ *for* $\bar{y} := \frac{1}{\lambda}\nabla\phi(A\bar{u} - \bar{z})$*. Then, for* $\lambda \in (0, \lambda_f)$ *sufficiently small, the stationarity condition* (7) *is fulfilled, i.e.,* $\bar{u}$ *is a stationary point for the inf-convolution model* (1)*.*

*Proof.* Invoking Theorem 4.3 with $(\bar{z}, \frac{1}{\lambda}\nabla\phi(A\bar{u} - \bar{z}))$ and $\lambda \in (0, \lambda_f)$ sufficiently small, we obtain $A^\top\nabla e_\lambda^\phi f(A\bar{u}) = \frac{1}{\lambda}A^\top\nabla\phi(A\bar{u} - \bar{z})$. In combination with (5a), this yields $0 \in A^\top\nabla e_\lambda^\phi f(A\bar{u}) + \partial g(\bar{u})$. Since $e_\lambda^\phi f$ is continuously differentiable around $A\bar{u}$ and $\bar{u} \in \operatorname{dom} g$, this implies (7) due to [22, Exercise 8.8 (c)]. □

# 5 Application to Distributed Training

In this section, we present a stochastic alternating minimization scheme (Algorithm 1) tailored to distributed empirical risk minimization in (15). As an interesting special case, we consider scenarios where all workers have access to the entire training set and the method specializes to elastic averaging SGD (EASGD) [27].

## 5.1 Inexact Alternating Minimization

For the optimization of the splitting problem (4) one typically resorts to alternating minimization where the variables $u$ and $z$ are updated as:

$$u^{t+1} \in \operatorname*{arg\,min}_{u \in \mathbb{R}^n} \; g(u) + \frac{1}{\lambda}\phi(Au - z^t), \tag{14a}$$

$$z^{t+1} \in \operatorname*{arg\,min}_{z \in \mathbb{R}^m} \; \frac{1}{\lambda}\phi(Au^{t+1} - z) + f(z). \tag{14b}$$

Such a scheme is known as the Gauss-Seidel or block coordinate descent method and has been investigated in a general setting by, e.g., [2, 5, 25].

In our case we regard alternating minimization as a generic scheme where the subproblems (in particular the $z$-update) may be solved approximately (e.g., by replacing the function with a surrogate) as long as convergence to a stationary point of the splitting problem can be guaranteed. Then, the translation of stationarity from Corollary 4.4 applies under prox-regularity.

For instance, the vanilla Gauss-Seidel method can be extended with a proximal regularization as in [1]. When $\phi$ is Lipschitz differentiable, the coupling term $\phi(Au - z)$ can be replaced by a proximal linearization in $z$ (resp. $u$) at $z^t$ (resp. $u^t$), so that the updates become proximal gradient steps on $F$ as in proximal alternating linearized minimization (PALM) [6].

Alternating minimization can also incorporate stochastic gradient updates, as described in the next subsection.

## 5.2 Stochastic Alternating Minimization for Distributed Training

In distributed learning, a set of $M$ workers collaborates on the training of a model parameterized by consensus weights $u$. To this end, with access to a prescribed subset (indexed by $\mathcal{I}_j$) of the full training set (indexed by $\mathcal{I}$), each individual worker trains its local copy $z_j$ of the model parameters $u$ under a (relaxed) consensus constraint $u = z_j$. As a consequence, all workers can update their copies in parallel.

In terms of the model (4), this is formulated as follows.

For the remainder of this section let $f : \mathbb{R}^{nM} \to \mathbb{R}$ with

$$f(z) = \sum_{j=1}^{M} f_j(z_j)$$

be a separable sum of (regularized) continuously differentiable empirical risks $f_j : \mathbb{R}^n \to \mathbb{R}$, each assigned to worker $j$. More specifically each $f_j$ is defined as

$$f_j(z_j) = \frac{1}{|\mathcal{I}_j|} \sum_{i \in \mathcal{I}_j} \ell(h_i, H(x_i; z_j)) + R(z_j),$$

for training pairs $(x_i, h_i)_{i \in \mathcal{I}_j}$, some prediction function $H(\cdot; z)$ parameterized by weights $z$, a loss function $\ell(\cdot, \cdot)$ and a regularizer $R$. Furthermore, let $g := 0$ and $A := [I, \cdots, I]^\top \in \mathbb{R}^{(nM)\times n}$, and set the Legendre penalty as

$$\phi(w) := \sum_{j=1}^{M} \widehat{\phi}(w_j)$$

which is separable over the copies $w_j$. Overall model (4) then reads:

$$\min_{u, (z_j)_{j=1}^M \in \mathbb{R}^n} F(u, z) = \sum_{j=1}^{M} f_j(z_j) + \frac{1}{\lambda} \widehat{\phi}(u - z_j), \quad (15)$$

where $\widehat{\phi}(u - z_j)$ loosely enforces the consensus constraint $u = z_j$.

The stochastic instance of the alternating minimization scheme is formulated in Algorithm 1 below. Notably, Algorithm 1 specializes to EASGD [27] in the isotropic case, i.e. when $\phi$ is quadratic and all workers have access to the entire training set $\mathcal{I}$, i.e. $\mathcal{I}_j = \mathcal{I}$.

**Algorithm 1** Stochastic Alternating Linearized Minimization

1: **for all** $t = 1, 2, \ldots$ **do**
2: Choose proper step sizes $\sigma_t, \tau > 0$.
3: $u^{t+1} = u^t - \tau A^\top \nabla \phi(Au^t - z^t)$.
4: Draw a sample of the random variable $\xi^t$.
5: Compute $\Delta(z^t; \xi^t)$ as an unbiased estimate of the gradient of $f + \frac{1}{\lambda}\phi(Au^{t+1} - \cdot)$ at $z^t$.
6: $z^{t+1} = z^t - \sigma_t \Delta(z^t; \xi^t)$.
7: **end for**

To obtain an unbiased estimate of the gradient $\nabla_z F(u^{t+1}, z^t)$ of $F(u^{t+1}, \cdot) = f + \frac{1}{\lambda}\phi(Au^{t+1} - \cdot)$ at $z^t$, worker $j$ samples a uniformly random minibatch $\mathcal{B}_j^t$ from $\mathcal{I}_j$, and computes the standard stochastic gradient,

$$\delta_j^t = \frac{1}{|\mathcal{B}_j^t|} \sum_{i \in \mathcal{B}_j^t} \nabla(\ell(h_i, H(x_i; \cdot)))(z_j^t). \quad (16)$$

Then we may define $\Delta(z^t; \xi^t) := (\Delta_j(z_j^t; \xi_j^t))_{j=1}^M$ for

$$\Delta_j(z_j^t; \xi_j^t) = \delta_j^t + \nabla R(z_j^t) - \frac{1}{\lambda} \nabla \phi(u^t - z_j^t). \quad (17)$$

The $u$-update in Algorithm 1 combines the current model parameters $z_j^t$ of the workers into a consensus model $u^{t+1}$. Notably, for the isotropic case and the particular choice $\tau = 1/M$, the consensus update reduces to the arithmetic mean of the copies $z_j^t$.

As a main difference to EASGD, for general $\phi$ the $u$-update can be regarded as a more general form of averaging. For instance, for the non-admissible choice $\phi = \|\cdot\|_1$, it is well known that minimization of $F$ w.r.t. $u$ yields the (componentwise) median.

A convergence proof for the stochastic method is beyond the scope of the paper. Instead we focus on the characterization of the stationary points $(\bar{u}, \bar{z})$, that the algorithm attempts to find.

Invoking the translation of stationarity reveals that the solution $\bar{u}$ is stationary w.r.t. the sum of $\phi$-envelopes $\sum_{j=1}^M e_\lambda^{\widehat{\phi}} f_j$. If furthermore all workers can sample from the entire training set $\mathcal{I}$, i.e., $\mathcal{I}_j = \mathcal{I}$ and therefore all $f_j = \hat{f}$ are equal, perfect consensus $\bar{u} = \bar{z}_j$ holds at stationary points for some finite penalty parameter $1/\lambda > 0$, which translates to a stationary point of the unregularized problem $\min_u \hat{f}(u)$ satisfying $0 \in \partial \hat{f}(\bar{u})$. This is not trivial as the workers may follow different paths to different stationary points if not coupled tightly (due to stochasticity). In addition, our theory shows that the $\phi$-envelope is Lipschitz differentiable at $\bar{u}$ and $\nabla e_\lambda^{\phi} \hat{f}(\bar{u}) = 0$, implying that whenever $u^\nu \to \bar{u}$ it holds that $\|\nabla e_\lambda^{\phi} \hat{f}(u^\nu)\| \to 0$. The gradient norm of the $\phi$-envelope may thus serve as a measure of stationarity, more robust compared to $\operatorname{dist}(0, \partial \hat{f}(u^\nu)) \to 0$, see [10]. All of these properties are formally stated in the following corollary.

**Corollary 5.1.** *Let $(\bar{u}, \bar{z})$ be a stationary point for the splitting model (15) satisfying $\bar{u} - \bar{z}_j \in \operatorname{dom} \widehat{\phi}$ and conditions (5a)–(5b). If $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ is prox-regular at $\bar{z}$ for $\bar{y} = \frac{1}{\lambda}\nabla\phi(A\bar{u} - \bar{z})$ and $\phi$-prox-bounded with threshold $\lambda_f$, then for $\lambda \in (0, \lambda_f)$ sufficiently small, $\bar{u}$ is a stationary point of $\sum_{j=1}^M e_\lambda^{\widehat{\phi}} f_j$, i.e., $\sum_{j=1}^M e_\lambda^{\widehat{\phi}} f_j$ is Lipschitz differentiable around $\bar{u}$ and*

$$\sum_{j=1}^{M} \nabla e_\lambda^{\widehat{\phi}} f_j(\bar{u}) = 0.$$

*If in addition $\mathcal{I}_j = \mathcal{I}$ for all $1 \le j \le M$ and therefore $f_j = \hat{f}$ the stationarity condition reduces to*

$$0 = \nabla e_\lambda^{\widehat{\phi}} \hat{f}(\bar{u}),$$

*and it holds that, $\bar{u} = \bar{z}_j$ for all $j$ and $0 \in \partial \hat{f}(\bar{u})$.*

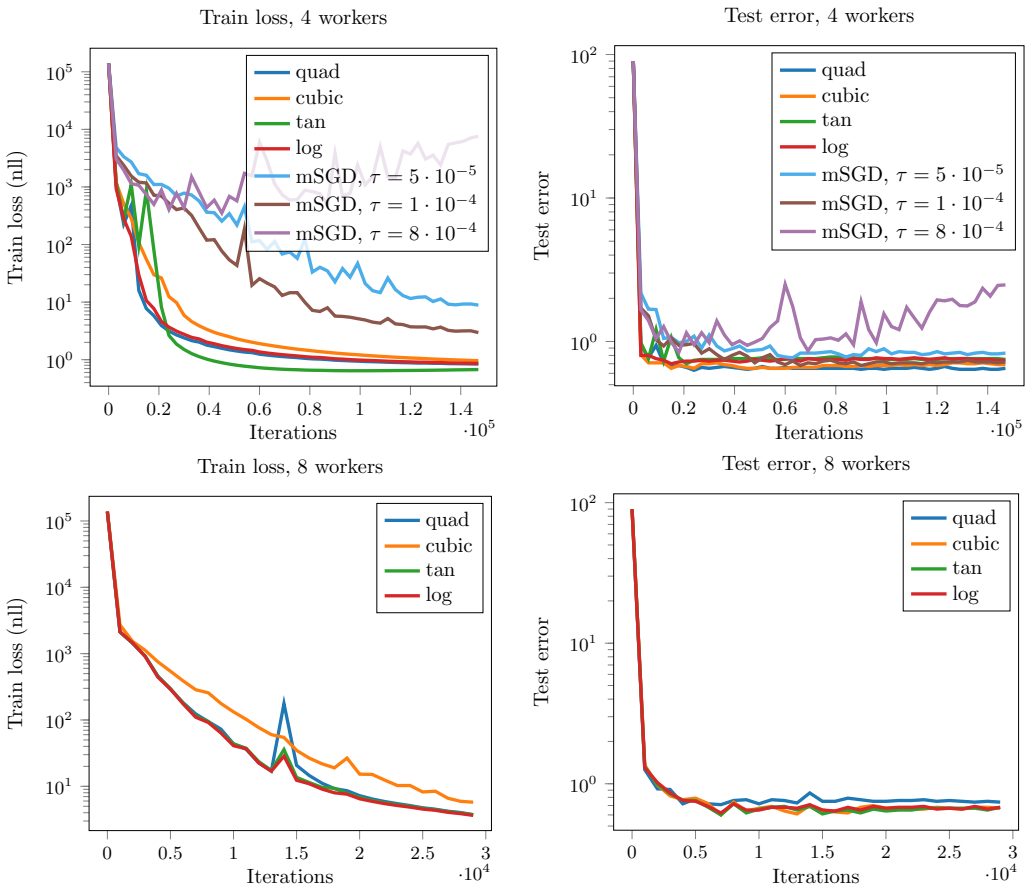


Figure 1: Convergence plots for stochastic distributed training with Algorithm 1 and classical Nesterov momentum SGD on MNIST for 4 workers (upper row) and 8 workers (lower row). In the 4 workers setting each worker completes 50 epochs. In the 8 worker setting each worker completes 10 epochs. In both cases our method is run with a learning rate of $\tau = 0.005$, much higher than the highest one possible with SGD (for $\tau = 8 \cdot 10^{-4}$ mSGD already becomes unstable). For all algorithms the batch-size is 20.

*Proof.* The first part of the corollary follows directly from Corollary 4.4 and the special structure of $A, f, g$ and $\phi$. For the second part we invoke Theorem 4.3 (i) and obtain that for some $\lambda > 0$, sufficiently small $P_\lambda^{\widehat{\phi}} \hat{f}$ is single-valued at $\bar{u}$ and that $\bar{z}_j = P_\lambda^{\widehat{\phi}} \hat{f}(\bar{u})$, showing that all $\bar{z}_j$'s are equal. From (5a) it follows that

$$0 = A^\top \nabla \phi(A\bar{u} - \bar{z}) = M \nabla \widehat{\phi}(\bar{u} - P_\lambda^{\widehat{\phi}} \hat{f}(\bar{u})).$$

By Assumptions (A1), (A5) and Lemma 3.2 we have $\nabla \widehat{\phi}(w) = 0$ if and only if $w = 0$, and therefore $\bar{u} = \bar{z}_j$ and $\bar{y}_j = 0$. From (5b) we know $0 \in \partial \hat{f}(\bar{z}_j) = \partial \hat{f}(\bar{u})$. □

## 6 Numerical Experiments

In the experiments we consider the stochastic distributed training of a deep neural network with Algorithm 1 where all workers have access to the entire training set, i.e. $\mathcal{I}_j = \mathcal{I}$ for all $j$ in terms of the model (15). More precisely, we report comparisons of different choices of potentials. We manually set $\widehat{\phi}$ as

$$\widehat{\phi}(w_j) := \sum_{l=1}^{L} \widehat{\widehat{\phi}} \left( \frac{w_{jl}}{\eta} \right), \tag{18}$$

where $\eta$ is a scaling parameter (in addition to $\lambda$), and $l$ indexes the (learnable) layers. The different choices of $\widehat{\widehat{\phi}}$ are summarized in Table 2. Since both the log- and tan-potentials have bounded domains, we incorporate a line search in our algorithm to ensure that the iterates stay feasible. Note that cubic does not satisfy Assumption (A3) as its Hessian is 0 at 0. Yet we include it in our numerical evaluation.

We apply a variant of our Algorithm 1 with constant step size, that in addition incorporates Nesterov momentum [23] in the SGD updates of the $z_j$. Note that for $\phi := \frac{1}{2}\|\cdot\|^2$ this algorithm specializes to the synchronous EASGD, resp. mEASGD [27].

Table 1: Comparison of different potentials on the stochastic consensus training of a deep neural network on MNIST. Results after 30,000 iterations, so that each of the 4 workers completes 10 epochs with batch size 20. For each potential we report the best values in performance over all configurations. Our evaluation suggests that quadratic is not the best smoothing potential in general.

| $\phi$ | Objective | Train Loss (nll) | Train Error | Test Loss (nll) | Test Error |
|---|---|---|---|---|---|
| quad, [27] | 2.49 | 2.47 | 0.00 % | 294.12 | 0.65 % |
| cubic | 7.44 | 6.87 | 0.00 % | **231.42** | **0.56** % |
| tan | **0.90** | **0.83** | 0.00 % | 300.84 | 0.62 % |
| tan-sep | 0.91 | 0.87 | 0.00 % | 306.91 | 0.64 % |
| log | 2.36 | 2.35 | 0.00 % | 299.60 | 0.64 % |
| log-sep | 2.46 | 2.44 | 0.00 % | 299.82 | 0.67 % |

Table 2: Choices for $\widehat{\widehat{\phi}}$. To ensure that cubic satisfies Assumption (A3) a small quadratic may be added.

| acronyms | $\widehat{\widehat{\phi}}$ | $\operatorname{dom}\widehat{\widehat{\phi}}$ |
|---|---|---|
| quad | $\|\cdot\|^2$ | $\mathbb{R}^m$ |
| cubic | $\|\cdot\|_3^3$ | $\mathbb{R}^m$ |
| tan | $\tan(\|\cdot\|^2)$ | $\{w : \|w\| < \sqrt{\frac{\pi}{2}}\}$ |
| tan-sep | $\sum_i \tan((\cdot)_i^2)$ | $\{w : \|w_i| < \sqrt{\frac{\pi}{2}}\}$ |
| log | $-\log(1-\|\cdot\|^2)$ | $\{w : \|w\| < 1\}$ |
| log-sep | $\sum_i -\log(1-(\cdot)_i^2)$ | $\{w : \|w_i| < 1\}$ |

We report results of the training of a classifier on the MNIST dataset resorting to the standard LeNet-5 CNN architecture [13] given as

$$\mathrm{Conv}_{20,5,1} \to \mathrm{ReLU} \to \mathrm{Pool}_{2,2} \to \mathrm{Conv}_{50,5,1} \to \mathrm{ReLU} \to \mathrm{Pool}_{2,2} \to \mathrm{FC} \to \mathrm{Softmax}$$

In Table 1 we compare different potentials $\widehat{\widehat{\phi}}$. To this end we perform a grid search over different learning rates $\sigma = \tau \in \{0.001, 0.005\}$, momentum parameters $\kappa \in \{0.9, 0.99\}$, and different scalings of the potentials $\lambda \in \{0.1, 0.05, 0.025, 0.01, 0.005, 0.0025\}$, $\eta \in \{0.5, 1, 2\}$. We set the number of workers to 4, the batch size to 20 and the regularization parameter $\nu = 10^{-4}$. We run the algorithm for 30,000 iterations, so that each worker completes 10 epochs. For each potential we report the best performances over all scalings and configurations in Table 1.

In terms of training loss, the tan- and log-potentials seem slightly superior, while the non-separable variants yield slightly better performance than the separable ones. The cubic potential performs worst in terms of training loss. All potentials consistently yield 0% training error after 10 epochs (for each worker).

Notably, we observe that the cubic potential performs better, in terms of low test error, than all other potentials. This is even true for a whole range of scalings.

In Figure 1, we show convergence plots for a representative configuration of our Algorithm 1 for 4 and 8 workers respectively. In addition, we show a comparison to plain SGD with Nesterov momentum (mSGD). The different EASGD variants "see" 4 to 8 times more training examples than mSGD within one iteration. As a result the distributed method is stable at much higher learning rates and requires fewer iterations than mSGD to achieve low objective value.

## 7 Conclusion

In this work we have considered an anisotropic generalization of the proximal mapping and Moreau-envelope. We derived a gradient formula for the Moreau-envelope in the nonconvex setting based on prox-regularity. This allows us to equivalently (in terms of stationary points) reformulate the problem as a splitting problem which is amenable to (stochastic) alternating minimization. As an application of our theory we characterize stationary points of the penalty objective that is optimized by the elastic averaging SGD method. There, our theory can be used to obtain a robust measure of stationarity. Through numerical validations we demonstrated the relevance of our theory and algorithm on the important task of consensus training of deep neural networks.

### Acknowledgement

The work was supported by the DFG Research Grant "Splitting Methods for 3D Reconstruction and SLAM".

## References


[1] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality. *Mathematics of Operations Research*, 35:438–457, 2010.

[2] A. Auslender. *Optimisation: Méthodes Numériques*. Masson, 1976.

[3] H. H. Bauschke, J. M. Borwein, and P. L. Combettes. Bregman monotone optimization algorithms. *SIAM Journal on Control and Optimization*, 42(2):596–636, 2003.

[4] H. H. Bauschke and A. S. Lewis. Dykstras algorithm with Bregman projections: A convergence proof. *Optimization*, 48(4):409–427, 2000.

[5] D. P. Bertsekas and J. N. Tsitsiklis. *Parallel and Distributed Computation: Numerical Methods*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1989.

[6] J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1-2):459–494, 2014.

[7] Y. Censor and S. A. Zenios. Proximal minimization algorithm with D-functions. *Journal of Optimization Theory and Applications*, 73(3):451–464, 1992.

[8] P. Chaudhari, A. Oberman, S. Osher, S. Soatto, and G. Carlier. Deep relaxation: partial differential equations for optimizing deep neural networks. *Research in the Mathematical Sciences*, 5(3):30, 2018.

[9] P. L. Combettes and N. N. Reyes. Moreau's decomposition in Banach spaces. *Mathematical Programming*, 139(1-2):103–114, 2013.

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

[11] A. L. Dontchev and R. T. Rockafellar. *Implicit Functions and Solution Mappings: A View From Variational Analysis*. Springer, New York, 2009.

[12] E. Laude, T. Wu, and D. Cremers. A nonconvex proximal splitting algorithm under Moreau-Yosida regularization. In *Artificial Intelligence and Statistics (AISTATS)*, 2018.

[13] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[14] C. Lescarret. Applications "prox" dans un espace de Banach. *Comptes Rendus de l'Académie des Sciences de Paris Série A*, 265:676–678, 1967.

[15] A. S. Lewis, D. R. Luke, and J. Malick. Local linear convergence for alternating and averaged nonconvex projections. *Foundations of Computational Mathematics*, 9(4):485–513, 2009.

[16] J.-J. Moreau. Fonctions convexes duales et points proximaux dans un espace Hilbertien. *Comptes Rendus de l'Académie des Sciences de Paris Série A*, 255:2897–2899, 1962.

[17] J.-J. Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société Mathématique de France*, 93(2):273–299, 1965.

[18] P. Ochs. Local convergence of the heavy-ball method and iPiano for non-convex optimization. *Journal of Optimization Theory and Applications*, 177:153–180, 2018.

[19] R. A. Poliquin and R. T. Rockafellar. Prox-regular functions in variational analysis. *Trans. Amer. Math. Soc.*, 348:1805–1838, 1996.

[20] S. M. Robinson. Strongly regular generalized equations. *Mathematics of Operations Research*, 5:43–62, 1980.

[21] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, New Jersey, 1970.

[22] R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*. Springer, New York, 1998.

[23] I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *International Conference on Machine Learning (ICML)*, pages 1139–1147, 2013.

[24] M. Teboulle. Entropic proximal mappings with applications to nonlinear programming. *Mathematics of Operations Research*, 17(3):670–690, 1992.

[25] P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494, 2001.

[26] D. Wexler. Prox-mappings associated with a pair of Legendre conjugate functions. *Revue française d'automatique informatique recherche opérationnelle*, 7(R2):39–65, 1973.

[27] S. Zhang, A. Choromanska, and Y. LeCun. Deep learning with elastic averaging SGD. In *Advances in Neural Information Processing Systems (NIPS)*, pages 685–693, 2015.