A Nonconvex Proximal Splitting Algorithm under Moreau-Yosida Regularization

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

Abstract

We tackle composite optimization problems whose objectives are (highly) nonconvex and nonsmooth. Classical nonconvex proximal splitting algorithms, such as nonconvex ADMM, suffer from a lack of convergence for such a problem class. In this work, we consider a Moreau-Yosida regularized variant of the original model and propose a novel multi-block primal-dual algorithm on the resulting lifted problem. We provide a complete convergence analysis of our algorithm, and identify respective optimality qualifications under which stationarity of the regularized problem is retrieved at convergence. Numerically, we demonstrate the relevance of our model and the efficiency of our algorithm on robust regression as well as joint variable selection and transductive learning.

1 Introduction

Many relevant problems in statistics and machine learning take the form as follows:

$$\begin{aligned} \underset{u\in\mathbb{R}^n,\ v\in\mathbb{R}^m}{\text{minimize}} \quad & f(v)+g(u) \\ \text{subject to} \quad & Au=v, \end{aligned} \tag{1}$$

see [4, 13, 12] for an overview. Here $A \in \mathbb{R}^{m\times n}$ is a matrix. Both extended real-valued functions, f and g, are assumed to be proper, lower semi-continuous (lsc), and in general *nonconvex*. The objective is *separable* in the block variables u and v that comply with a *consensus constraint*. Such a formulation is amenable to distributed solution methods such as (accelerated) proximal gradient methods [17, 19] and the alternating direction method of multipliers (ADMM) [11, 10, 9, 8, 13, 15]. Algorithmically, these proximal splitting algorithms typically update the block variables in a Gauss-Seidel fashion, and exploit the blessing that the *proximal mapping* of f and/or g can be efficiently computed. The convergence (or even applicability) is, however, only guaranteed under rather restrictive conditions. At least one of the functions between f and g is required to be Lipschitz differentiable, cf. [13, 15, 12, 24]. In some sense, convexity is traded off against smoothness in a convergent algorithm. When both f and g are nonsmooth and nonconvex, the convergence of classical proximal splitting algorithms is lost.

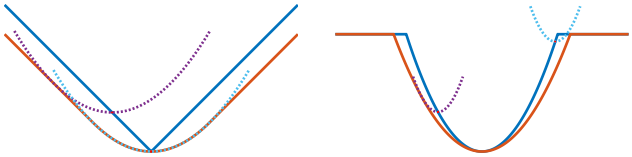

Figure 1: Illustration of the Moreau envelope (Red curves). Left: Convex case with absolute value function (blue curve). The Moreau envelope is convex and Lipschitz differentiable. Right: Nonconvex case with truncated quadratic function (blue curve). The Moreau envelope is again a truncated quadratic that is nonconvex and nonsmooth. The dashed curves correspond to the quadratic functions with slope $1/\lambda$ and illustrate the geometrical construction of the Moreau envelope via the Minkowski sum of the epigraphs, cf. [20, Exercise 1.28].

Bearing this in mind, we introduce the Moreau-Yosida regularized problem:

$$\begin{aligned} \underset{u\in\mathbb{R}^n,\ v\in\mathbb{R}^m}{\text{minimize}} \quad & e_\lambda f(v)+g(u) \\ \text{subject to} \quad & Au=v, \end{aligned} \tag{2}$$

where $e_\lambda f$ is the Moreau-Yosida envelope of f with parameter $\lambda > 0$ defined according to [20].

Definition 1 (Moreau envelope and proximal mapping)**.** *For a proper, lsc function $f : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ and parameter $\lambda > 0$, the Moreau envelope function*

$e_\lambda f$ and the proximal mapping $P_\lambda f$ are defined by

$$e_\lambda f(v) := \inf_{w\in\mathbb{R}^m} f(w) + \frac{1}{2\lambda}\|w-v\|^2, \tag{3}$$

$$P_\lambda f(v) := \operatorname*{arg\,min}_{w\in\mathbb{R}^m} f(w) + \frac{1}{2\lambda}\|w-v\|^2. \tag{4}$$

To demonstrate the relevance of (2) in machine learning, we briefly review some general properties of the Moreau envelope. For convex f, the Moreau envelope $e_\lambda f$ yields a convex and smooth lower approximation to f [20, Theorem 2.26]. A well-known example is the Huber loss, defined as

$$\ell_\lambda(x) := \begin{cases} \frac{1}{2\lambda}x^2 & \text{if } |x| \le \lambda \\ |x| - \frac{\lambda}{2} & \text{otherwise,} \end{cases} \tag{5}$$

which can be written in terms of the Moreau envelope $e_\lambda|\cdot|$ of the absolute value function $|\cdot|$. The Moreau envelope also has a geometrical interpretation. Its epigraph $\operatorname{epi} e_\lambda f := \{(p,q)\in\mathbb{R}^{m+1} : e_\lambda f(p) \le q\}$ is the Minkowski sum of the epigraph of f and the epigraph of the quadratic function $\frac{1}{2\lambda}\|\cdot\|^2$ [20, Exercise 1.28]. This is depicted in Figure 1. In the limit case where $\lambda \to 0$, the Moreau envelope converges pointwisely to the original function f [20, Theorem 1.25]. Overall, this reveals that for nonconvex, nonsmooth f, the Moreau envelope $e_\lambda f$ remains nonsmooth and nonconvex in general which renders its optimization challenging.

Practically, the Moreau envelope allows to express robust nonconvex loss functions which often lead to superior practical performance in regression and classification tasks [6, 7, 21, 22]. For instance, the truncated quadratic loss can be expressed as the Moreau envelope of the ℓ_0-norm, the Moreau envelope of the ramp loss yields the truncated Huberized hinge loss. Model (2) is relevant not only for supervised learning but also for transductive learning and clustering. For instance, the Moreau envelope of the "symmetric hinge loss", written $\min_{\theta\in\{-1,1\}}(1-(\cdot)\theta)_+$, used in transductive SVM models [23, 14, 6] yields the nonconvex nonsmooth symmetric "Huberized hinge loss".

In this work, we propose a novel multiblock primal-dual scheme for solving the regularized Problem (2). Our contributions are summarized as follows:

- We devise a novel algorithm, a multiblock primal-dual scheme, in Section 2.1 and prove its subsequential convergence to a critical point of the *lifted* representation of Problem (2) in Section 3.
- We draw interesting connections of our primal-dual algorithm to existing, fully primal Gauss-Seidel minimization of a quadratic penalty function; see Section 2.2.
- We prove in Section 3, that for Lipschitz f and $\lambda \to 0^+$, we consistently solve the unregularized problem (1), where $e_\lambda f$ is replaced by f, to stationarity. The violation of the linear constraint in (1) is quantified in terms of λ.
- For piecewise smooth, piecewise convex functions of the form $\min_{i\in\mathcal{I}} e_\lambda f_i$ (as a subfamily of nonconvex, nonsmooth functions), we identify an optimality qualification related to *active set*, under which a critical point of the lifted problem translates to a critical point of (2); see Section 4.2.
- We experimentally validate our proposed algorithm on robust regression as well as joint feature selection and transductive learning in Section 5. In comparison with classical methods such as ADMM, our method consistently performs favorably in terms of lower objective values and vanishing optimality gaps.

2 A Multiblock Primal-Dual Algorithm

2.1 Derivation

For smooth $e_\lambda f$, the convergence of ADMM in the nonconvex setting is shown via a monotonic decrease of the augmented Lagrangian [13], which serves as a Lyapunov function. As a key step in the proof of [13], the ascent of the augmented Lagrangian, caused by the dual update, is dominated by a sufficient descent in the primal block, that is updated last.

In order to recover convergence for nonsmooth $e_\lambda f$, we employ a lifting of the problem which yields a third primal block in the optimization. We introduce new variables z, w along with a linear constraint

$$z + w = v, \tag{6}$$

and integrate the Moreau-Yosida regularization into the lifted problem:

$$\begin{aligned} \underset{u\in\mathbb{R}^n,\ w,z\in\mathbb{R}^m}{\text{minimize}} \quad & f(z) + \frac{1}{2\lambda}\|w\|^2 + g(u) \\ \text{subject to} \quad & Au - z - w = 0. \end{aligned} \tag{7}$$

To this lifted problem we apply the following novel multiblock primal-dual scheme, where the block w, updated last, corresponds to the smooth function $\frac{1}{2\lambda}\|\cdot\|^2$, that realizes the Moreau-Yosida regularization. Let the augmented Lagrangian of the problem be defined as

$$\begin{aligned} \mathcal{M}_\rho(u,z,w,y) = f(z) + g(u) + \frac{1}{2\lambda}\|w\|^2 \\ + \langle y, Au - z - w\rangle + \frac{\rho}{2}\|Au - z - w\|^2, \end{aligned} \tag{8}$$

for $\rho > 0$. Also let $M := \frac{1}{\sigma} I - \rho A^\top A$ positive definite for $\sigma\rho\|A\|^2 < 1$ and $\|\cdot\|_M := \sqrt{\langle \cdot, M \cdot \rangle}$. Then our scheme is formulated as

$$\begin{aligned} u^{t+1} &= \arg\min_u \mathcal{M}_\rho(u, z^t, w^t, y^t) + \tfrac{1}{2}\|u - u^t\|_M^2, \\ z^{t+1} &= \arg\min_z \mathcal{M}_\rho(u^{t+1}, z, w^t, y^t), \\ w^{t+1} &= \arg\min_w \mathcal{M}_\rho(u^{t+1}, z^{t+1}, w, y^t), \\ y^{t+1} &= y^t + \rho(Au^{t+1} - z^{t+1} - w^{t+1}). \end{aligned} \tag{9}$$

Note that the update of u involves a proximal term $\frac{1}{2}\|u-u^t\|_M^2$, which makes the algorithm different from classical multiblock ADMM. Such a design is motivated from [5] and makes the subproblem for the u-update computationally tractable, as long as the proximal mapping of g is simple. Interestingly, once rephrasing the update of u in terms of a proximal mapping as $u^{t+1} = P_\sigma g(u^t - \sigma A^\top(y^t + \rho(Au^t - z^t - \lambda y^t)))$, one can interpret it as a proximal gradient descent step on the augmented Lagrangian.

The optimality condition for the last primal block update $0 = {}^1\!/\!_\lambda w^{t+1} - y^t - \rho(Au^{t+1} - z^{t+1} - w^{t+1})$ matches the dual update and shows that w is equal to the dual variable up to scaling, i.e. $\lambda y^{t+1} = w^{t+1}$. Therefore, the variable w can be eliminated from the algorithm, and we arrive at a compact formulation of the proposed multiblock primal-dual scheme for Moreau-Yosida regularized problems, Algorithm 1.

Algorithm 1 (multiblock primal-dual scheme)**.** *Choose ρ, σ so that $\rho\lambda > 1$ and $\sigma\rho\|A\|^2 < 1$. For $t = 1, 2, \ldots$ do*

$$\begin{aligned} u^{t+1} &= P_\sigma g\left(u^t - \sigma A^\top(y^t + \rho(Au^t - z^t - \lambda y^t))\right), \\ z^{t+1} &= P_{1/\rho} f\left(Au^{t+1} + ({}^1\!/\!_\rho - \lambda) y^t\right), \\ y^{t+1} &= \tfrac{1}{1+\rho\lambda}(y^t + \rho(Au^{t+1} - z^{t+1})). \end{aligned}$$

Note that the lifted problem is equivalent to Problem (2) in terms of global minimizers but in general not in terms of critical points. Yet, we show in Section 4 that under mild assumptions, e.g. for piecewise convex piecewise smooth $e_\lambda f$, the limit points produced by this algorithm translate to the critical points of the original Problem 2 by reversing the substitution (6). As a side remark, with $\lambda = 0$ we recover a proximal variant of ADMM, similar to [5, 13, 15], applied to the unregularized problem (1). In the remainder of this paper, we refer to this variant as *proximal ADMM*.

2.2 Primal Optimization as a Special Case

Finally, we draw a connection between Algorithm 1 and existing, fully primal alternating minimization schemes applied to the quadratic penalty of (1):

$$Q(u, z) = f(z) + g(u) + \frac{1}{2\lambda}\|Au - z\|^2. \tag{10}$$

For a non-admissible choice of the step size $\rho = {}^1\!/\!_\lambda$, the Lagrange multiplier y^t in Algorithm 1 becomes obsolete and we recover fully primal Gauss-Seidel minimization of (10) over u, z, Algorithm 2.

Algorithm 2 (proximal penalty method)**.** *Choose σ so that $\sigma\|A\|^2 < \lambda$. For $t = 1, 2, \ldots$ do*

$$\begin{aligned} u^{t+1} &= P_\sigma g\left(u^t - {}^\sigma\!/\!_\lambda A^\top(Au^t - z^t)\right), \\ z^{t+1} &= P_\lambda f\left(Au^{t+1}\right). \end{aligned}$$

The above Algorithm appears similar in form to *proximal alternating linearized minimization* (PALM) [3] applied to (10), where $H(u, z) := \frac{1}{2\lambda}\|Au - z\|^2$ is interpreted as the differentiable coupling term. To update z, PALM invokes a second proximal gradient descent step on (10) with step size $\tau < \lambda$, i.e. $z^{t+1} = P_\tau f(z^t + \frac{\tau}{\lambda}(Au^{t+1} - z^t))$. With the non-admissible step size $\tau = \lambda$, we recast Algorithm 2 from PALM.

3 Convergence Analysis

Our convergence proof borrows arguments from [13], where the convergence of ADMM was shown via a monotonic decrease of the augmented Lagrangian. In our case a Lyapunov function that monotonically decreases over the iterations is obtained by eliminating the variable w from the augmented Lagrangian (9):

$$\begin{aligned} \mathcal{Q}_\rho(u, z, y) = f(z) - \frac{\lambda}{2}\|y\|^2 + g(u) \\ + \langle Au - z, y\rangle + \frac{\rho}{2}\|Au - z - \lambda y\|^2. \end{aligned} \tag{11}$$

The following Lemma is the central part of our convergence proof, as it guarantees convergence of $\mathcal{Q}_\rho(u^t, z^t, y^t)$. In contrast to [13], our algorithm has three primal blocks and invokes a proximal gradient descent step on $\mathcal{Q}_\rho(\cdot, z^{t+1}, y^{t+1})$ to update u. This keeps all block variable updates computationally tractable, as long as the proximal mappings of f and g are simple.

Lemma 1. *Let $\lambda > 0$ be fixed. Choose $\rho > 0$ sufficiently large so that $\lambda\rho > 1$, and then $\sigma > 0$ sufficiently small so that $\sigma\rho\|A\|^2 < 1$. Assume that $e_\lambda f(Au) + g(u) > -\infty$ for any $u \in \mathbb{R}^n$. Then we have*

1. *$\mathcal{Q}_\rho(u^{t+1}, z^{t+1}, y^{t+1})$ is an upper bound of the quadratic penalty (10) at the iterates (u^{t+1}, z^{t+1}), i.e.*

$$Q(u^{t+1}, z^{t+1}) \le \mathcal{Q}_\rho(u^{t+1}, z^{t+1}, y^{t+1}). \tag{12}$$

2. *The sequence* $\{\mathcal{Q}_\rho(u^t, z^t, y^t)\}_{t\in\mathbb{N}}$ *is bounded from below.*

3. *A sufficient decrease over* $\mathcal{Q}_\rho$ *for each iteration is guaranteed:*

$$\begin{aligned}-\infty < \mathcal{Q}_\rho(u^{t+1}, z^{t+1}, y^{t+1}) - \mathcal{Q}_\rho(u^t, z^t, y^t)\\ \leq \left(\tfrac{\rho\|A\|^2}{2} - \tfrac{1}{2\sigma}\right)\|u^{t+1} - u^t\|^2\\ + \left(\tfrac{1}{\rho} - \tfrac{\rho\lambda^2+\lambda}{2}\right)\|y^{t+1} - y^t\|^2.\end{aligned} \tag{13}$$

Our aim is to prove subsequential convergence of the iterates $\{(u^t, z^t, y^t)\}_{t\in\mathbb{N}}$ to a critical point of (7). To guarantee the existence of such a subsequence, the iterates have to be bounded. To ensure, that the conditions of a critical point are met at limit points, the distance of consecutive iterates has to vanish in the limit. This is guaranteed in the following Lemma.

Lemma 2. *Assume that the assumptions in Lemma 1 hold and let the quadratic penalty* (10) *be coercive. Then, the iterates* $\{(u^t, z^t, y^t)\}_{t\in\mathbb{N}}$ *are bounded, feasibility is achieved in the limit*

$$\|Au^t - z^t - \lambda y^t\| \to 0, \tag{14}$$

and the difference of two consecutive iterates goes to zero:

$$\|u^{t+1} - u^t\| \to 0, \tag{15}$$
$$\|z^{t+1} - z^t\| \to 0, \tag{16}$$
$$\|y^{t+1} - y^t\| \to 0. \tag{17}$$

It remains to show that limit points of the algorithm correspond to critical points of the lifted Problem (7). To this end we define the generalized subdifferential [20, Definition 8.3]:

Definition 2 (subgradients)**.** *Consider a function* $f : \mathbb{R}^m \to \mathbb{R}\cup\{\infty\}$ *and a point* $\bar{v}$ *with* $f(\bar{v})$ *finite. For a vector* $y \in \mathbb{R}^n$*, one says that*

1. y *is a regular subgradient of* f *at* $\bar{v}$*, written* $y \in \hat{\partial} f(\bar{v})$*, if*

$$\liminf_{\substack{v\to\bar{v}\\ v\neq\bar{v}}} \frac{f(v) - f(\bar{v}) - \langle y, v - \bar{v}\rangle}{\|v - \bar{v}\|} \geq 0. \tag{18}$$

2. y *is a (general) subgradient of* f *at* $\bar{v}$*, written* $y \in \partial f(\bar{v})$*, if there are sequences* $v^t \to \bar{v}$ *with* $f(v^t) \to f(\bar{v})$ *and* $y^t \in \hat{\partial} f(v^t)$ *with* $y^t \to y$*.*

In this concluding theorem we guarantee subsequent convergence to critical points. Having proven that the difference of consecutive iterates vanishes in the limit, the optimality follows directly from the optimality conditions of the subproblem.

Theorem 1. *Let* (u^*, z^*, y^*) *be a limit point of the sequence* $\{(u^t, z^t, y^t)\}_{t\in\mathbb{N}}$ *produced by Algorithm 1. Then,* (u^*, z^*, y^*) *gives rise to a critical point of the Problem* (7)*, i.e.*

$$0 \in \partial f(z^*) - y^*, \tag{19}$$
$$0 \in \partial g(u^*) + A^\top y^*, \tag{20}$$
$$Au^* - z^* - \lambda y^* = 0, \tag{21}$$

with $w^* := \lambda y^*$*.*

Proof. Let $\{t_j\}_{j\in\mathbb{N}} \subset \{t\}_{t\in\mathbb{N}}$ be the subindices such that $\lim_{j\to\infty}(u^{t_j}, z^{t_j}, y^{t_j}) = (u^*, z^*, y^*)$. The optimality condition for the u-update is given as

$$\begin{aligned}0 \in \partial g(u^{t_j+1}) + \frac{1}{\sigma}(u^{t_j+1} - u^{t_j}) + A^\top y^{t_j}\\ +\rho A^\top(Au^{t_j} - z^{t_j} - \lambda y^{t_j}).\end{aligned}$$

The optimality condition for the z-update is given as

$$0 \in \partial f(z^{t_j+1}) - y^{t_j} - \rho(Au^{t_j+1} - z^{t_j+1} - \lambda y^{t_j}).$$

The y-update gives

$$y^{t_j+1} = \tfrac{1}{1+\rho\lambda}(y^{t_j} + \rho(Au^{t_j+1} - z^{t_j+1})).$$

Letting $j \to \infty$ and by Lemma 2 and the closedness of ∂f, this yields (19), (20) and (21) with $w^* = \lambda y^*$. □

We conclude this section with the following observation. The above result reveals, that we obtain a solution to the original Problem (1), up to a violation of the linear constraint that is absorbed by λy^*. Let f be Lipschitz continuous with modulus L over $\mathrm{dom}(f)$ (which contains z^*). Then we can a-priori specify a bound on the violation of the linear constraint, i.e.

$$\|\lambda y^*\| \leq \lambda L. \tag{22}$$

Furthermore, this implies that in the limit case $\lambda \to 0^+$ we recover the optimality conditions of the unregluarized Problem (1).

4 Optimality Qualifications

In this section our aim is to understand when a critical point of the lifted problem translates to a critical point of the Moreau-Yosida regularized Problem (2) by reversing the substitution (6). We identify optimality qualifications, in case of (local) prox-regularity and piecewise convexity respectively, under which such a translation holds true.

4.1 Prox-regular Functions

It is easy to see, that for convex f reversing the substitution yields a critical point. We show, that this assumption can be relaxed to f being prox-regular at a limit point of the iterates. Prox-regular functions behave locally like convex functions in the sense, that the Moreau envelope is locally Lipschitz differentiable and the proximal mapping is single-valued in a small neighborhood of the input argument. To this end we define prox-bounded and prox-regular functions according to [20].

Definition 3 (prox-boundedness)**.** *A function $f : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ is prox-bounded if there exists $\lambda > 0$ such that $e_\lambda f(p) > -\infty$ for some $p \in \mathbb{R}^m$.*

Definition 4 (prox-regularity)**.** *A function $f : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ is prox-regular at $\bar{z}$ for $\bar{y}$ if f is finite and locally lsc at $\bar{z}$ with $\bar{y} \in \partial f(\bar{z})$, and there exist $\epsilon > 0$ and $\rho \geq 0$ such that*

$$f(z') \geq f(z) + \langle y, z' - z\rangle - \frac{\rho}{2}\|z' - z\|^2, \tag{23}$$

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

Theorem 2. *Let (u^*, z^*, y^*) be a limit point of the sequence $\{(u^t, z^t, y^t)\}_{t\in\mathbb{N}}$ produced by Algorithm 1. Let f be prox-regular at z^* with y^* and also prox-bounded. Set $p^* := z^* + \lambda y^*$. Then, for $\lambda > 0$ sufficiently small, (u^*, p^*, y^*) corresponds to a critical point of the regularized Problem (2):*

$$0 = \nabla e_\lambda f(p^*) - y^*, \tag{24}$$

$$0 \in \partial g(u^*) + A^\top y^*, \tag{25}$$

$$Au^* - p^* = 0. \tag{26}$$

Proof. Conditions (25) and (26) follow directly from Theorem 1. From Theorem 1 we also know $0 \in \partial f(z^*) - y^*$, or equivalently $0 \in \partial(f(\cdot) - \langle \cdot, y^*\rangle)(z^*)$. Since $f(\cdot) - \langle \cdot, y^*\rangle$ is prox-regular at z^* with $\bar{y} = 0$ (cf. [20, Exercise 13.35]) and prox-bounded, we can apply [20, Proposition 13.37] and obtain that there exists $\lambda > 0$ sufficiently small, so that $\nabla e_\lambda(f(\cdot) - \langle \cdot, y^*\rangle)$ is differentiable at z^* with $0 = \nabla e_\lambda(f(\cdot) - \langle \cdot, y^*\rangle)(z^*)$. A straightforward calculation shows that $e_\lambda(f(\cdot) - \langle \cdot, y^*\rangle)(z) = e_\lambda f(z + \lambda y^*) - \frac{1}{2\lambda}\|\lambda y^*\|^2 - \frac{1}{\lambda}\langle z, \lambda y^*\rangle$. Further differentiating with respect to z on both sides yields (24) with $p^* = z^* + \lambda y^*$. □

As prox-regularity is a local property, it comes along with tools [20], particularly well suited to prove local convergence results [18]. However, in our case, we are rather interested in global subsequent convergence. In that sense the above theorem comes with a caveat: There may be a cyclic dependency between the step size $1/\rho < \lambda$ of the algorithm and the choice of λ. A smaller λ requires the choice of the step size $1/\rho$ to be smaller, which may alter the limit point. This makes it impossible to a priori specify a feasible pair of parameter λ and step size, as long as there exists no uniform λ and f is prox-regular everywhere. If f is prox-regular everywhere with uniform λ, one also speaks of semi-convex [1, 16] functions.

4.2 Piecewise Convex Functions

In this subsection, we consider $f : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ being a piecewise convex function with finitely many pieces. Such a function can be expressed in terms of a pointwise minimum over convex functions including indicator functions:

$$f(z) := \min_{i\in\mathcal{I}} f_i(z), \tag{27}$$

with each $f_i : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ convex, proper, lsc and a finite index set $\mathcal{I}$. Its Moreau envelope $e_\lambda f$ is given in terms of a pointwise minimum over the Moreau envelopes of the individual functions $\{f_i\}$, i.e.

$$e_\lambda f(z) = \min_{i\in\mathcal{I}} e_\lambda f_i(z), \tag{28}$$

and therefore remains nonsmooth. Due to [20, Theorem 2.26], each individual piece $e_\lambda f_i$ is continuously differentiable, as f_i is convex, proper, and lsc. Practically, this class is relevant in statistics and machine learning. For examples we refer to Sections 1 and 5. Even though f is prox-regular almost everywhere, Theorem 2 cannot be applied conveniently for predefined λ which rather depends on the (a-priori unknown) limit point. Hence, our goal is to prove a more explicit optimality qualification for piecewise convex f (with predefined λ) related to active set, to overcome the limitation of Theorem 2:

To this end we define the set of active indexes at z as

Definition 5 (active set)**.** *Let $z \in \mathbb{R}^m$ with $f(z)$ finite. The active index set $\mathcal{A}_f(z)$ of f at z is defined as*

$$\mathcal{A}_f(z) = \{i \in \mathcal{I} : f(z) = f_i(z)\}. \tag{29}$$

We will show, that the qualification condition given as:

$$\mathcal{A}_f(z^*) \subset \mathcal{A}_{e_\lambda f}(z^* + \lambda y^*), \tag{30}$$

guarantees optimality of the limit points. In essence, the condition requests that, after translation $z^* + \lambda y^*$, the same piece remains active with respect to its Moreau envelope. We remark that the active-set qualification condition (30) is indeed observed numerically

in our experiments, both along the iterates (after sufficiently many iterations) and at the limit point.

In the following, we characterize the subdifferential in terms of the convex subdifferentials of the pieces. In Lemma 3 we show that an inclusion holds in a general setting. In Lemma 4 we show, that this inclusion holds with equality for the Moreau envelopes if the hypograph of $e_\lambda f$ satisfies the *linear independence constraint qualification* (LICQ).

Lemma 3. *Let $z \in \mathbb{R}^m$ with $f(z)$ finite. Then the following inclusion holds:*

$$\partial f(z) \subset \bigcup_{i \in \mathcal{A}_f(z)} \partial f_i(z). \tag{31}$$

Proof. Let $p \in \partial f(z)$. By definition, we have $z^t \to z$ with $f(z^t) \to f(z)$ and $p^t \in \hat{\partial} f(z^t)$ with $p^t \to p$. Since the active set is never empty we may choose $i^t \in \mathcal{A}_f(z^t)$. Since $\{i^t\}_{t\in\mathbb{N}}$ is discrete and $f(z^t) \to f(z)$, there exists a subsequence $\{t_j\}_{j\in\mathbb{N}} \subset \{t\}_{t\in\mathbb{N}}$ so that $i^{t_j} = i^*$ for all $j \in \mathbb{N}$ with some constant $i^* \in \mathcal{A}_f(z)$. Since $p^{t_j} \in \hat{\partial} f(z^{t_j})$ and $f_{i^*}(z^{t_j}) = f(z^{t_j})$ and since $f_{i^*}(\cdot) \geq f(\cdot)$, we have

$$\begin{aligned} 0 &\leq \liminf_{z' \to z^{t_j}} \frac{f(z') - f(z^{t_j}) - \langle p^{t_j}, z' - z^{t_j}\rangle}{\|z' - z^{t_j}\|} \\ &\leq \liminf_{z' \to z^{t_j}} \frac{f_{i^*}(z') - f_{i^*}(z^{t_j}) - \langle p^{t_j}, z' - z^{t_j}\rangle}{\|z' - z^{t_j}\|}, \end{aligned}$$

which implies that $p^{t_j} \in \hat{\partial} f_{i^*}(z^{t_j})$ for all j. This shows that $p \in \partial f_{i^*}(z)$. □

Lemma 4. *Let $p \in \mathbb{R}^m$. Assume that the set of active normals at p, defined as*

$$\mathcal{U}_p := \left\{ (\nabla e_\lambda f_i(p), -1)^\top : i \in \mathcal{A}_{e_\lambda f}(p) \right\}, \tag{32}$$

is linearly independent. Then (31) *holds with equality for the Moreau envelope $e_\lambda f$:*

$$\partial e_\lambda f(p) = \left\{ \nabla e_\lambda f_i(p) : i \in \mathcal{A}_{e_\lambda f}(p) \right\}. \tag{33}$$

Proof. Let $p \in \mathbb{R}^m$. Note that each $e_\lambda f_i$ is continuously differentiable, as f_i is convex, proper lsc (cf. [20, Theorem 2.26]). To show the desired result we construct for any $i \in \mathcal{A}_{e_\lambda f}(p)$ a direction $v_i \in \mathbb{R}^m$ such that for $\tau > 0$ sufficiently small, i is the only active piece at $p + \tau v_i$, i.e. $\mathcal{A}_{e_\lambda f}(p + \tau v_i) = \{i\}$. Then, cleary $e_\lambda f$ is differentiable at $p + \tau v_i$ with $\nabla e_\lambda f(p + \tau v_i) = \nabla e_\lambda f_i(p + \tau v_i)$, which proves that $\nabla e_\lambda f_i(p) \in \partial e_\lambda f(p)$.

To this end let $l := |\mathcal{A}_{e_\lambda f}(p)|$ and we define the matrix $U \in \mathbb{R}^{(m+1)\times l}$ such that $U_{\cdot,i} = (\nabla e_\lambda f_i(p), -1)^\top$ for $i = 1, \ldots, l$. For each i, let $v_i \in \mathbb{R}^m$, $\beta_i \in \mathbb{R}$, $\alpha_i \in \mathbb{R}^l$ such that

$$(v_i, \beta_i)^\top = U\alpha_i. \tag{34}$$

Now choose the α_i as follows. Let $\gamma_i \in \mathbb{R}^l$ be the i-th unit vector. Since by assumption U has full column-rank, the following linear system

$$U^\top U \alpha_i = -\gamma_i, \tag{35}$$

has a unique solution $\alpha_i = (U^\top U)^{-1}\gamma_i$. This implies that $-1 = \langle \nabla e_\lambda f_i(p), v_i\rangle - \beta_i < \langle \nabla e_\lambda f_j(p), v_i\rangle - \beta_i = 0$, for all $j \in \mathcal{A}_{e_\lambda f}(p) \setminus \{i\}$. For $\tau > 0$ sufficiently small this means that

$$\langle \nabla e_\lambda f_i(p), v_i\rangle + \frac{o(\tau)}{\tau} < \langle \nabla e_\lambda f_j(p), v_i\rangle + \frac{o(\tau)}{\tau} \tag{36}$$

and hence

$$e_\lambda f_i(p + \tau v_i) < e_\lambda f_j(p + \tau v_i), \tag{37}$$

for all $j \in \mathcal{A}_{e_\lambda f}(p) \setminus \{i\}$. Thus, we verify $\mathcal{A}_{e_\lambda f}(p + \tau v_i) = \{i\}$ as desired, and the conclusion follows. □

By characterizing the hypograph $\operatorname{hypo} e_\lambda f := \{(p, q) : q \leq e_\lambda f(p)\}$ of f in terms of nonlinear constraints $g_i(p, q) := q - e_\lambda f_i(p) \leq 0$ for all $i \in \mathcal{I}$, the assumption in Lemma 4 is equivalent to the LICQ applied to $\operatorname{hypo} e_\lambda f$ at the point $(p, e_\lambda f(p))$.

We now conclude this section with two theorems, which guarantee the stationarity of the limit points for Problem (2) under the proposed qualification conditions. Theorem 3 applies to Algorithm 1, and Theorem 4 to Algorithm 2.

Theorem 3. *Let (u^*, z^*, y^*) be a limit point of the sequence $\{(u^t, z^t, y^t)\}_{t\in\mathbb{N}}$ produced by Algorithm 1. Let $p^* = z^* + \lambda y^*$ for $\lambda > 0$ and let $\mathcal{U}_{p^*}$ be linearly independent. If the qualification condition* (30) *holds at (u^*, z^*, y^*), then*

$$0 \in \partial e_\lambda f(p^*) - y^*, \tag{38}$$

and (u^*, p^*, y^*) corresponds to a critical point of the regularized Problem (2)*.*

Proof. Conditions (25) and (26) follow as a direct consequence of Theorem 1.

From Theorem 1 we know that $0 \in \partial f(z^*) - y^*$. From Lemma 3 we know there exists $i^* \in \mathcal{A}_f(z^*)$ so that: $y^* \in \partial f_{i^*}(z^*)$. By the definition of p^* we have $p^* \in z^* + \lambda \partial f_{i^*}(z^*)$. Interpreting the inclusion as the optimality condition of the proximal mapping of the convex, proper, lsc function f_{i^*}, yields that $z^* = P_\lambda f_{i^*}(p^*)$. By [20, Theorem 2.26] we have that for any $\lambda > 0$ it holds that

$$\frac{1}{\lambda}(p^* - P_\lambda f_{i^*}(p^*)) = \nabla e_\lambda f_{i^*}(p^*). \tag{39}$$

Rearranging the terms shows that $0 = \nabla e_\lambda f_{i^*}(p^*) - y^*$. By assumption $i^* \in \mathcal{A}_{e_\lambda f}(p^*)$ and since $\mathcal{U}_{p^*}$ is linearly independent we can apply Lemma 4 and obtain $\nabla e_\lambda f_{i^*}(p^*) \in \partial e_\lambda f(p^*)$. This concludes the proof. □

The qualification condition above comprises the LICQ of the hypograph, which is satisfied everywhere in many practically relevant cases, including robust loss functions or the symmetric hinge loss.

Finally, we consider the case, when $\lambda\rho = 1$, i.e. when our method specializes to Algorithm 2. We show, that in case the iterates of the algorithm satisfy $\|z^{t+1} - z^t\| \to 0$, it solves the regularized problem, without assuming (30). As Algorithm 2 produces no Lagrange multiplier, we set up the multiplier as $y^t := 1/\lambda(Au^t - z^t)$.

Theorem 4. *Let (u^*, z^*) be a limit point of the sequence $\{(u^t, z^t)\}_{t\in\mathbb{N}}$ produced by Algorithm 2. Define $y^* := 1/\lambda(Au^* - z^*)$ and $p^* := Au^*$. Let $\|z^{t+1} - z^t\| \to 0$ and let $\mathcal{U}_{p^*}$ be linearly independent. Then (u^*, p^*, y^*) is a critical point of the Problem (2), i.e. the conditions (38), (25) and (26) hold.*

For the proof we exploit, that the proximal mapping $z^{t+1} = P_\lambda f(Au^{t+1})$ is taken with the same step size as the Moreau envelope $e_\lambda f(p^{t+1})$, and that $p^{t+1} := Au^{t+1}$, which makes (30) obsolete. This comes at the cost that $\|z^{t+1} - z^t\| \to 0$ cannot be guaranteed by Lemma 2, however, this turned out to be valid in all conducted experiments. A complete proof is provided in the Appendix A.3.

5 Numerical Experiments

In this section, we compare our algorithms to existing methods, namely proximal ADMM [5, 13, 15], vanilla ADMM [8] and PALM [3], on the task of robust linear regression and joint variable selection and transductive learning with linear classifiers. We show that our method consistently behaves favorably in terms of lower objective value and vanishing optimality gap. In view of (38), (25) and (26) the optimality gap is defined as

$$\begin{aligned}\text{gap} := \operatorname{dist}^2(0, \partial e_\lambda f(p^*) - y^*) \\ + \operatorname{dist}^2(0, \partial g(u^*) + A^\top y^*) + \|Au^* - p^*\|^2.\end{aligned} \tag{40}$$

Due to Lemma 4 and the relation of $\nabla e_\lambda f_i$ and $P_\lambda f_i$ [20, Theorem 2.26], computing an optimality gap is convenient.

5.1 Robust Linear Regression

In linear regression one is interested in reconstructing a signal $u \in \mathbb{R}^n$ from noisy measurements $b \in \mathbb{R}^m$. The forward model takes the form of

$$b = Au + \epsilon, \tag{41}$$

where $A \in \mathbb{R}^{m\times n}$ describes the linear sampling and ϵ is a disturbance term.

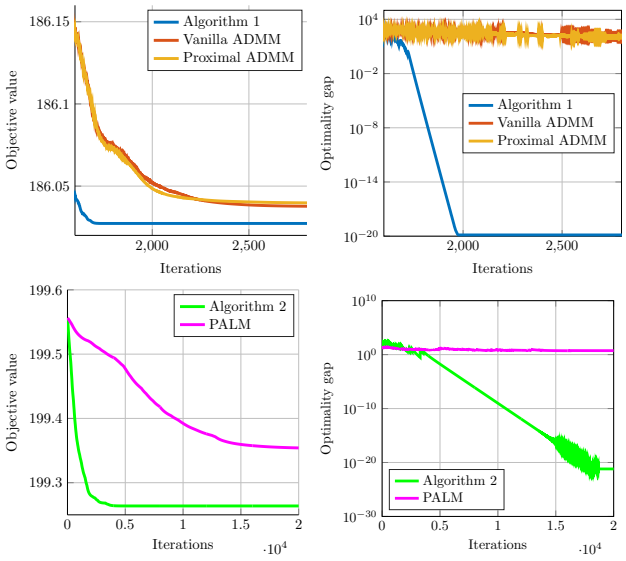

Figure 2: Performance comparison on robust regression task. Upper row: Comparison of primal-dual type methods (proximal ADMM [5, 13, 15], vanilla ADMM [8] and Algorithm 1). Lower row: Comparison of primal methods (PALM [3] and Algorithm 2). In can be seen that our Algorithms converge to critical points of (2) with low objective value. Both ADMM and PALM fail to converge to a critical point of the regularized problem.

Instead of plain least squares, we use the truncated quadratic loss $\ell(x) = \min\{\nu, \frac{1}{2\lambda}x^2\}$ [21, 22], which is more robust against outliers. In order to match the form (2), we express truncated quadratic as the Moreau envelope of the sum of ℓ_0 terms:

$$f(v) := \nu \sum_{i=1}^{m} \|v_i\|_0. \tag{42}$$

We benchmark proximal ADMM [5, 13, 15], vanilla ADMM [13], PALM [3] and Algorithms 1 and 2 on synthetic data. The entries of $A \in \mathbb{R}^{20000\times 10}$ and $b \in \mathbb{R}^{20000}$ are i.i.d. and normal distributed. We further degrade b with additive Gaussian and high impulsive noise by adding a large constant to 60 % of the entries. We manually choose $\lambda = 0.05$ and $\nu = 0.01$. For both ADMM and our proposed Algorithm 1, a warmup phase is launched, where the step size ρ is initialized with a small value and then grows exponentially over the iterations up to a value slightly bigger than $1/\lambda = 20$. In practice, this setup often leads to lower objective value. To give ADMM a better chance to converge, we increase ρ even further up to $8000 \gg 20$. Yet, as shown in Figure 2, both ADMM and PALM fail to converge to a critical point of (2). In contrast, our Algorithm finds critical points with lower objective value.

Table 1: Comparison of proximal ADMM [5, 13, 15], and Algorithm 1 on the task of sparse transductive learning. The results are consistent with the previous experiment. Note, that for the fully supervised case, due to the convexity of $e_\lambda f$, our algorithm produces the same result as proximal ADMM.

	proximal ADMM					Algorithm 1				
% labeled	Test Error	Objective	Sparsity	Iterations	Gap	Test Error	Objective	Sparsity	Iterations	Gap
0.05	99.50 %	3653.58	99.40 %	20000	$4.7 \cdot 10^{5}$	1.61 %	**2029.35**	99.60 %	9427	$\mathbf{1.6 \cdot 10^{-10}}$
0.36	67.50 %	2895.97	99.20 %	20000	$2.2 \cdot 10^{5}$	1.61 %	**2029.35**	99.60 %	9321	$\mathbf{1.6 \cdot 10^{-10}}$
1.19	1.61 %	2035.54	99.40 %	20000	$4.2 \cdot 10^{2}$	1.72 %	**2030.90**	99.40 %	15882	$\mathbf{1.3 \cdot 10^{-9}}$
2.38	1.61 %	2061.15	99.60 %	20000	$3.1 \cdot 10^{3}$	1.67 %	**2031.20**	99.40 %	16066	$\mathbf{1.1 \cdot 10^{-9}}$
11.90	1.61 %	2035.32	99.00 %	20000	$7.5 \cdot 10^{-1}$	1.56 %	**2033.12**	98.80 %	15749	$\mathbf{1.4 \cdot 10^{-9}}$
100.00	1.61 %	**2090.07**	99.60 %	678	$\mathbf{1.3 \cdot 10^{-13}}$	1.61 %	**2090.07**	99.60 %	678	$\mathbf{1.3 \cdot 10^{-13}}$

5.2 Block-sparse Multiclass Huber-TSVM

To further demonstrate flexibility of our model, we consider the problem of joint variable selection [2] and transductive learning [23] with linear classifiers. We use a Huberized one-vs.-rest *transductive SVM* (TSVM) model [23, 14, 6] and a nonconvex, block-sparsity-promoting regularizer on the classifier. The overall task is to jointly infer the labels of the $m-l$ unlabeled examples, estimate the linear one-vs.-rest classifiers u, and select the features.

We set up the individual components of model (2) as follows. Let m be the number of training examples, among which $l \leq m$ examples are labeled and the rest is unlabeled. Let $X \in \mathbb{R}^{m \times d}$ be the feature matrix and $u \in \mathbb{R}^{d \times c}$ the classifier matrix. We introduce linear constraints $Xu = v \in R^{m \times c}$, so that each row $v_{i,\cdot} = X_{i,\cdot}u \in \mathbb{R}^{1 \times c}$ contains the linear classifier scores associated with the i-th training example. Then, each entry of $v_{i,\cdot}$ corresponds to a class $1 \leq j \leq c$. We model the term f as the sum of two terms: the first corresponds to the $m-l$ unlabeled training examples; the second term corresponds to the the labeled examples. More explicitly, the data term reads

$$f(v) := \sum_{i=1}^{m-l} \min_{1 \leq \theta_i \leq c} \ell(v_i; \theta_i) + \sum_{i=m-l+1}^{m} \ell(v_i; \theta_i), \quad (43)$$

where $\theta_i \in \{1, \ldots, c\}$ is fixed for $i > m-l$. The individual loss terms $\ell(\cdot; \theta_i)$ are given as $\ell(v_i; \theta_i) := \sum_{j=1}^{c} (1 - v_{ij}(2[\theta_i = j] - 1))_+$, where $(1-x)_+$ denotes the hinge loss, and $[\cdot]$ the Iverson bracket.

For feature selection, we promote block-sparsity on the classifier matrix u by the $\ell_{2,0}$-norm regularization defined as $\|u\|_{2,0} := \sum_{j=1}^{d} \|(\sum_{i=1}^{c} |u_{ji}|^2)^{1/2}\|_0$. The block-sparsity enforces a consensus feature selection among all classifiers. We also include a squared Frobenius norm, $\beta\|u\|_F^2$, to control the margin in the SVM model and ensure the coercivity of the model. Altogether, the regularization term g is set up as

$$g(u) = \alpha\|u\|_{2,0} + \beta\|u\|_F^2. \quad (44)$$

In Table 1 we benchmark proximal ADMM vs. Algorithm 1 on synthetic data (4200 examples, 3 classes) that is not linearly separable and degraded by 500 additional feature components containing noise. We manually choose $\alpha = 0.2$, $\beta = 10$, $\lambda = 0.05$. In view of Lemma 2 we stop the algorithm, when the difference of two consecutive iterates is below a certain threshold or the maximum number of 20000 iterations is reached. The results are consistent with the previous example: ADMM does not converge to a critical point within the maximum number of iterations, except for the fully supervised case $l = m$. In that case, both, Algorithm 1 and proximal ADMM converge and yield the same result, because of the convexity of f and the smoothness of $e_\lambda f$ for $l = m$.

6 Conclusion

In this work we have tackled highly nonconvex Moreau-Yosida regularized composite problems, where both terms in the objective are nonsmooth. Classical proximal splitting algorithms such as ADMM fail to converge on this problem class. To overcome this limitation, we devise a novel primal-dual proximal splitting algorithm, that intrinsically regularizes the behavior of the dual variable. For piecewise convex functions we devise mild qualification conditions, that guarantee convergence to a critical point of the regularized problem. We validated our method on the optimization of challenging highly nonconvex machine learning objectives. For future work we will address a randomized variant of our algorithm suited to distributed computation in large-scale machine learning.

A Proofs

A.1 Proof of Lemma 1

Proof. To show the lower boundedness of $\mathcal{Q}_\rho(u^{t+1}, z^{t+1}, y^{t+1})$ we rewrite

$$\begin{aligned}\mathcal{Q}_\rho(u^{t+1}, z^{t+1}, y^{t+1}) &= f(z^{t+1}) + g(u^{t+1}) \\ &+ \frac{\rho}{2}\|Au^{t+1} - z^{t+1} - \lambda y^{t+1}\|^2 \\ &+ \frac{1}{2\lambda}\|Au^{t+1} - z^{t+1}\|^2 - \frac{\lambda}{2}\|y^{t+1}\|^2 \\ &+ \langle Au^{t+1} - z^{t+1}, y^{t+1}\rangle \\ &- \frac{1}{2\lambda}\|Au^{t+1} - z^{t+1}\|^2 \\ = f(z^{t+1}) &+ g(u^{t+1}) \\ &+ \frac{\rho}{2}\|Au^{t+1} - z^{t+1} - \lambda y^{t+1}\|^2 \\ &+ \frac{1}{2\lambda}\|Au^{t+1} - z^{t+1}\|^2 \\ &- \frac{1}{2\lambda}\|Au^{t+1} - z^{t+1} - \lambda y^{t+1}\|^2\end{aligned}$$

Since $\rho > \frac{1}{\lambda}$ we can further bound $\mathcal{Q}_\rho(u^{t+1}, z^{t+1}, y^{t+1})$ from below by the quadratic penalty

$$\begin{aligned}\mathcal{Q}_\rho(u^{t+1}, z^{t+1}, y^{t+1}) &\geq f(z^{t+1}) + g(u^{t+1}) \\ &+ \frac{1}{2\lambda}\|Au^{t+1} - z^{t+1}\|^2 \\ &\geq g(u^{t+1}) + \inf_{z\in\mathbb{R}^m} f(z) + \frac{1}{2\lambda}\|z - Au^{t+1}\|^2 \\ &= e_\lambda f(Au^{t+1}) + g(u^{t+1}) > -\infty\end{aligned}$$

To show the sufficient decrease of $\mathcal{Q}_\rho(u^{t+1}, z^{t+1}, y^{t+1})$ we split the difference above at the iterations $t+1$ and t into the terms

$$\begin{aligned}&\mathcal{Q}_\rho(u^{t+1}, z^{t+1}, y^{t+1}) - \mathcal{Q}_\rho(u^t, z^t, y^t) \\ &= \mathcal{Q}_\rho(u^{t+1}, z^t, y^t) - \mathcal{Q}_\rho(u^t, z^t, y^t) \\ &\quad + \mathcal{Q}_\rho(u^{t+1}, z^{t+1}, y^t) - \mathcal{Q}_\rho(u^{t+1}, z^t, y^t) \\ &\quad + \mathcal{Q}_\rho(u^{t+1}, z^{t+1}, y^{t+1}) - \mathcal{Q}_\rho(u^{t+1}, z^{t+1}, y^t)\end{aligned}$$

and bound each term separately. We find an estimate for $\mathcal{Q}_\rho(u^{t+1}, z^t, y^t) - \mathcal{Q}_\rho(u^t, z^t, y^t)$. Let

$$\mathcal{C}_\rho(u, z, y) := \langle Au, y\rangle + \frac{\rho}{2}\|Au - z - \lambda y\|^2,$$

denote the differentiable part of (11), so that

$$\begin{aligned}\mathcal{Q}_\rho(u^{t+1}, z^t, y^t) - \mathcal{Q}_\rho(u^t, z^t, y^t) &= g(u^{t+1}) \\ &+ \mathcal{C}_\rho(u^{t+1}, z^t, y^t) - g(u^t) - \mathcal{C}_\rho(u^t, z^t, y^t).\end{aligned}$$

By the definition of u^{t+1} as the solution of a proximal operator (it minimizes g plus a quadratic globally), we have the estimate

$$\begin{aligned}&\sigma g(u^{t+1}) + \frac{1}{2}\|u^{t+1} - (u^t - \sigma\nabla_u\mathcal{C}_\rho(u^t, z^t, y^t))\|^2 \\ &\quad\leq \sigma g(u^t) + \frac{1}{2}\|\sigma\nabla_u\mathcal{C}_\rho(u^t, z^t, y^t)\|^2\end{aligned} \tag{45}$$

Expanding the square norm and simplifying yields:

$$\begin{aligned}&g(u^{t+1}) + \frac{1}{2\sigma}\|u^{t+1} - u^t\|_2^2 \\ &\quad + \langle\nabla_u\mathcal{C}_\rho(u^t, z^t, y^t), u^{t+1} - u^t\rangle \leq g(u^t)\end{aligned} \tag{46}$$

Since $\nabla_u\mathcal{C}_{\boldsymbol{\rho}}$ is $\rho\|A\|^2$-smooth, we have:

$$\begin{aligned}&\mathcal{C}_\rho(u^{t+1}, z^t, y^t) - \mathcal{C}_\rho(u^t, z^t, y^t) - \frac{\rho\|A\|^2}{2}\|u^{t+1} - u^t\|_2^2 \\ &\quad\leq \langle\nabla_u\mathcal{C}_\rho(u^t, z^t, y^t), u^{t+1} - u^t\rangle\end{aligned} \tag{47}$$

This yields the estimate

$$\begin{aligned}&\mathcal{Q}_\rho(u^{t+1}, z^t, y^t) - \mathcal{Q}_\rho(u^t, z^t, y^t) \\ &\quad\leq \left(\frac{\rho\|A\|^2}{2} - \frac{1}{2\sigma}\right)\|u^{t+1} - u^t\|_2^2,\end{aligned} \tag{48}$$

which leads to a sufficient descent whenever $\sigma\rho\|A\|^2 < 1$.

The optimality for the z-update guarantees

$$\mathcal{Q}_\rho(u^{t+1}, z^{t+1}, y^t) - \mathcal{Q}_\rho(u^{t+1}, z^t, y^t) \leq 0 \tag{49}$$

Finally we bound the term

$$\begin{aligned}\mathcal{Q}_\rho(u^{t+1}, z^{t+1}, y^{t+1}) - \mathcal{Q}_\rho(u^{t+1}, z^{t+1}, y^t) &= -\frac{\lambda}{2}\|y^{t+1}\|^2 \\ &+ \frac{\lambda}{2}\|y^t\|^2 + \langle Au^{t+1} - z^{t+1}, y^{t+1} - y^t\rangle \\ &+ \frac{\rho}{2}\|Au^{t+1} - z^{t+1} - \lambda y^{t+1}\|^2 \\ &- \frac{\rho}{2}\|Au^{t+1} - z^{t+1} - \lambda y^t\|^2.\end{aligned} \tag{50}$$

Since $\frac{1}{\rho}(y^{t+1} - y^t) + \lambda y^{t+1} = Au^{t+1} - z^{t+1}$, we can rewrite

$$\begin{aligned}&-\frac{\lambda}{2}\|y^{t+1}\|^2 + \frac{\lambda}{2}\|y^t\|^2 + \langle Au^{t+1} - z^{t+1}, y^{t+1} - y^t\rangle \\ &= -\frac{\lambda}{2}\|y^{t+1}\|^2 + \frac{\lambda}{2}\|y^t\|^2 + \frac{1}{\rho}\|y^{t+1} - y^t\|^2 + \lambda\|y^{t+1}\|^2 \\ &\quad - \lambda\langle y^{t+1}, y^t\rangle \\ &= \frac{\lambda}{2}\|y^{t+1}\|^2 - \lambda\langle y^{t+1}, y^t\rangle + \frac{\lambda}{2}\|y^t\|^2 + \frac{1}{\rho}\|y^{t+1} - y^t\|^2 \\ &= \left(\frac{1}{\rho} + \frac{\lambda}{2}\right)\|y^{t+1} - y^t\|^2.\end{aligned}$$

We apply the identity $\|a + c\|^2 - \|b + c\|^2 = -\|b - a\|^2 + 2\langle a + c, a - b\rangle$ with $a := -\lambda y^{t+1}$, $b := -\lambda y^t$ and

$c = Au^{t+1} - z^{t+1}$ and obtain

$$\begin{aligned}
&\frac{\rho}{2}\|Au^{t+1} - z^{t+1} - \lambda y^{t+1}\|^2 - \frac{\rho}{2}\|Au^{t+1} - z^{t+1} - \lambda y^t\|^2 \\
&= -\frac{\rho\lambda^2}{2}\|y^{t+1} - y^t\|^2 \\
&\quad - \lambda\rho\langle Au^{t+1} - z^{t+1} - \lambda y^{t+1}, y^{t+1} - y^t\rangle \\
&= -\frac{\rho\lambda^2 + 2\lambda}{2}\|y^{t+1} - y^t\|^2.
\end{aligned}$$

Overall we have:

$$\begin{aligned}
&\mathcal{Q}_\rho(u^{t+1}, z^{t+1}, y^{t+1}) - \mathcal{Q}_\rho(u^{t+1}, z^{t+1}, y^t) \\
&\quad = \left(\frac{1}{\rho} - \frac{\rho\lambda^2 + \lambda}{2}\right)\|y^{t+1} - y^t\|^2.
\end{aligned} \tag{51}$$

Merging all together we obtain the desired result. □

A.2 Proof of Lemma 2

Proof. Since $\mathcal{Q}_\rho(u^t, z^t, y^t)$ monotonically decreases, by Lemma 1, it is upper bounded. Since $\mathcal{Q}_\rho(u^t, z^t, y^t)$ is an upper bound for the quadratic penalty $Q(u^t, z^t)$, and since Q is coercive by assumption, u^t and z^t are bounded. We sum the estimate (13) form Lemma 1 from $t = 1$ to T and obtain due to the lower boundedness of the iterates $\mathcal{Q}_\rho(u^t, z^t, y^t)$:

$$\begin{aligned}
-\infty < \mathcal{Q}_\rho(u^{T+1}, z^{T+1}, y^{T+1}) - \mathcal{Q}_\rho(u^1, z^1, y^1) \\
\leq -\sum_{t=1}^{T}\left(\frac{\rho\|A\|^2}{2} - \frac{1}{2\sigma}\right)\|u^{t+1} - u^t\|^2 \\
-\sum_{t=1}^{T}\left(\frac{1}{\rho} - \frac{\rho\lambda^2 + \lambda}{2}\right)\|y^{t+1} - y^t\|^2.
\end{aligned}$$

Letting $T \to \infty$ yields that $\|u^{t+1} - u^t\| \to 0$ and $\|y^{t+1} - y^t\| \to 0$ for sufficiently large ρ. From $\frac{1}{\rho}(y^{t+1} - y^t) = Au^{t+1} - z^{t+1} - \lambda y^{t+1}$ we have that,

$$\begin{aligned}
0 &\leq \|z^t - z^{t+1}\| \\
&= \| - z^{t+1} + z^t + A(u^{t+1} - u^t) - A(u^{t+1} - u^t) \\
&\qquad + \lambda y^{t+1} - \lambda y^t - \lambda y^{t+1} + \lambda y^t\| \\
&\leq \frac{1}{\rho}\|y^{t+1} - y^t\| + \|A\|\|u^{t+1} - u^t\| \\
&\qquad + \lambda\|y^{t+1} - y^t\| \to 0
\end{aligned}$$

and $\|Au^t - z^t - \lambda y^t\| \to 0$. This also shows, that y^t is bounded. □

A.3 Proof of Theorem 4

Proof. Let (u^*, z^*) be a limit point of the sequence $\{(u^t, z^t)\}_{t\in\mathbb{N}}$. Let $\{t_j\}_{j\in\mathbb{N}} \subset \{t\}_{t\in\mathbb{N}}$ be the corresponding subsequence. Since $z^{t_j} = P_{\lambda f}(Au^{t_j})$, there is $i^{t_j} \in \mathcal{A}_{e_\lambda f}(p^{t_j})$ so that $z^{t_j} = P_{\lambda f_{i^{t_j}}}(Au^{t_j})$. This follows directly from the definition of the proximal mapping. Wlog. $i^* = i^{t_j}$ constant. From the optimality condition of the proximal mapping it follows that

$$0 \in \partial f_{i^*}(z^{t_j+1}) - \frac{1}{\lambda}(Au^{t_j+1} - z^{t_j+1}). \tag{52}$$

We take the limit $j \to \infty$. By Lemma 2 and since $\|z^{t+1} - z^t\| \to 0$ this yields

$$0 \in \partial f_{i^*}(z^*) - \frac{1}{\lambda}(Au^* - z^*). \tag{53}$$

By the definition of $p^* := z^* + \lambda y^*$ we have $p^* \in z^* + \lambda\partial f_{i^*}(z^*)$. Interpreting the inclusion as the optimality condition of the proximal mapping of the convex, proper, lsc function f_{i^*}, proves that $z^* = P_\lambda f_{i^*}(p^*)$. By [20, Theorem 2.26] we have, that for any $\lambda > 0$ it holds that

$$\frac{1}{\lambda}(p^* - P_\lambda f_{i^*}(p^*)) = \nabla e_\lambda f_{i^*}(p^*). \tag{54}$$

Rearranging the terms shows that $0 = \nabla e_\lambda f_{i^*}(p^*) - y^*$. Since $e_\lambda f_i$ is continuous also $e_\lambda f$ is continuous as it is a pointwise minimum over continuous functions. Therefore, $i^* \in \mathcal{A}_{e_\lambda f}(p^*)$ and since $\mathcal{U}_{p^*}$ is linearly independent we can apply Lemma 4 and obtain $\nabla e_\lambda f_{i^*}(p^*) \in \partial e_\lambda f(p^*)$. Conditions (25) and (26) hold as a direct consequence of Theorem 1. This concludes the proof. □

References

[1] Marco Artina, Massimo Fornasier, and Francesco Solombrino. Linearly constrained nonsmooth and nonconvex minimization. *SIAM Journal on Optimization*, 23(3):1904–1937, 2013.

[2] Jinbo Bi, Kristin Bennett, Mark Embrechts, Curt Breneman, and Minghu Song. Dimensionality reduction via sparse support vector machines. *Journal of Machine Learning Research*, 3:1229–1243, 2003.

[3] Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1-2):459–494, 2014.

[4] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3:1–122, 2010.

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

[6] Ronan Collobert, Fabian Sinz, Jason Weston, and Léon Bottou. Large scale transductive svms. *Journal of Machine Learning Research*, 7:1687–1712, 2006.

[7] Ronan Collobert, Fabian H. Sinz, Jason Weston, and Léon Bottou. Trading convexity for scalability. In *Proceedings of the 23rd International Conference on Machine Learning (ICML 2006)*, pages 201–208, 2006.

[8] Jonathan Eckstein and Dimitri P. Bertsekas. On the douglas-rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55:293–318, 1992.

[9] Daniel Gabay. Applications of the method of multipliers to variational inequalities. *Studies in Mathematics and Its Applications*, 15:299–331, 1983.

[10] Daniel Gabay and Bertrand Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2(1):17–40, 1976.

[11] Roland Glowinski and A Marroco. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires. *Revue française d'automatique, informatique, recherche opérationnelle. Analyse numérique*, 9(2):41–76, 1975.

[12] D. Hajinezhad, M. Hong, T. Zhao, and Z. Wang. NESTT: A nonconvex primal-dual splitting method for distributed and stochastic optimization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3215–3223, 2016.

[13] Mingyi Hong, Zhi-Quan Luo, and Meisam Razaviyayn. Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. *SIAM Journal on Optimization*, 26(1):337–364, 2016.

[14] Thorsten Joachims. Transductive inference for text classification using support vector machines. In *Proceedings of the 16th International Conference on Machine Learning (ICML 1999)*, pages 200–209, 1999.

[15] Guoyin Li and Ting Kei Pong. Global convergence of splitting methods for nonconvex composite optimization. *SIAM Journal on Optimization*, 25(4):2434–2460, 2015.

[16] T. Möllenhoff, E. Strekalovskiy, M. Möller, and D. Cremers. The primal-dual hybrid gradient method for semiconvex splittings. *SIAM Journal on Imaging Sciences*, 8(2):827–857, 2015.

[17] Yu Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

[18] Peter Ochs. Local convergence of the heavy-ball method and ipiano for non-convex optimization. *arXiv preprint arXiv:1606.09070*, 2016.

[19] Peter Ochs, Yunjin Chen, Thomas Brox, and Thomas Pock. ipiano: Inertial proximal algorithm for nonconvex optimization. *SIAM Journal on Imaging Sciences*, 7(2):1388–1419, 2014.

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

[21] Yiyuan She and Art B Owen. Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association*, 106(494):626–639, 2011.

[22] Liu T and Jiang H. Minimizing sum of truncated convex functions and its applications. *Journal of Computational and Graphical Statistics*, 2017.

[23] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[24] Y. Wang, W. Yin, and J. Zeng. Global convergence of ADMM in nonconvex nonsmooth optimization. *arXiv:1511.06324v5*, 2017.