Minimizing Nonconvex Non-Separable Functions

Yaoliang Yu **Xun Zheng** **Micol Marchetti-Bowick** **Eric P. Xing**
Machine Learning Department, Carnegie Mellon University

Abstract

Regularization has played a key role in deriving sensible estimators in high dimensional statistical inference. A substantial amount of recent works has argued for nonconvex regularizers in favor of their superior theoretical properties and excellent practical performances. In a different but analogous vein, nonconvex loss functions are promoted because of their robustness against "outliers". However, these nonconvex formulations are computationally more challenging, especially in the presence of nonsmoothness and non-separability. To address this issue, we propose a new proximal gradient meta-algorithm by rigorously extending the *proximal average* to the nonconvex setting. We formally prove its nice convergence properties, and illustrate its effectiveness on two applications: multi-task graph-guided fused lasso and robust support vector machines. Experiments demonstrate that our method compares favorably against other alternatives.

1 Introduction

Regularization has played a major role in recent development of statistical machine learning algorithms and applications. Many regularizers, with their unique properties, have been designed. In particular, convex regularizers have been prevalent due to their computational convenience. However, the potential superiority of nonconvex regularizers has long been recognized and pursued [1–5]. Empirically, nonconvex regularizers often yield better results than their convex counterparts [6–8]. On the flip side, nonconvex regularizers are computationally more challenging, but there has been steady progress [6, 9–12]. For instance, [9, 10, 12] are among the first to apply the proximal gradient to nonconvex regularizers; [11] extended the coordinate descent to the nonconvex and nonsmooth setting; [6, 8] employed the convex-concave procedure; and [7] applied the alternating direction method of multipliers; etc. These existing works have greatly expanded our tool sets for coping with nonconvexity, generating remarkable successes but also suffering some limitations: a). Only apply to special scenarios [9, 11]; b). No convergence result [7] or merely the weak "convergence" in terms of function values [10–12]; c). Slow convergence due to successive linearization [6, 8]; d). Incapable of handling non-separability [9–12]. In this work we propose a meta-algorithm that enjoys stronger convergence guarantees and works in broader settings.

We are interested in the general setting where the regularizer (or the loss function) is nonconvex, nonsmooth, and non-separable. For instance, the overlapping group pursuit [8] advocated a nonconvex regularizer for each *overlapping* group and achieved better estimates. The same idea can be extended to the graph-guided fused lasso [13], see Example 1 below. However, the resulting optimization is now highly nontrivial, rendering many of the existing algorithms inapplicable, hence deserving a serious investigation.

We borrow the *proximal averaging* idea of the recent work [14] and significantly extend it to the nonconvex setting, by making the following contributions:

- Rigorously addressing the multi-valuedness and non-uniqueness of the proximal map. This difficulty does not occur for convex functions but is common for nonconvex ones. It is the key to deal with *non-separable* functions where most existing works (such as [9–12, 15]) do not apply.
- Re-establishing, sometimes with essential modifications, the many key properties of the proximal average, including a complete characterization on the real line. For instance, we show in Example 3 that there are infinitely many functions that all lead to the same hard-thresholding rule, thus shedding new lights on both the statistical and algorithmic aspects of nonconvex regularizers. These theoretical developments are com-

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.

pletely new and provide a solid ground for some ongoing work.

- Proving the convergence of the whole sequence produced by our algorithm, which is even new for the convex case. This contribution is particularly important in applications (such as biostatistics) where variable selection, if not the sole purpose, is as desirable as achieving a small prediction error.
- Experimentally validating the proposed algorithm on applications where nonconvexity and non-separability really makes a difference. The algorithm can be easily parallelized so even better efficiency can be anticipated.

To further demonstrate its flexibility, we also apply our approach to the robust support vector machines (RSVM) by swapping the role of loss and regularizer. Here, we encounter a second motivation for nonconvex functions: As shown in [16, 17], any convex loss cannot be robust against adversarial outliers. Accordingly, RSVM replaces the convex hinge loss with the nonconvex truncated hinge loss [18–20]. Through experiments we show that our algorithm is much more efficient than previous approaches such as alternating [18] and the convex-concave procedure [19–21].

We formally state our problem in Section 2. Section 3 contains all technical results that are essential for the convergence proof in Section 4. Experiments on both multi-task GFlasso and robust SVM are conducted in Section 5, and we conclude in Section 6.

2 Problem Formulation

We are interested in solving the minimization problem:

$$\min_{\mathbf{w}\in\mathbb{R}^p} \ \ell(\mathbf{w}) + \bar{f}(\mathbf{w}), \text{ where } \bar{f}(\mathbf{w}) = \sum_{k=1}^{K} \alpha_k f_k(\mathbf{w}), \quad (1)$$

and the scalars, $\alpha_k \geq 0, \sum_k \alpha_k = 1$, are fixed constants throughout the paper. It is clear that many statistical machine learning algorithms can be cast under our general formulation (1). For instance, take ℓ as the least squares loss and f_1 as the 1-norm (with $K = 1$) we recover lasso [22]. We can also swap the role of the "loss" ℓ and the "regularizer" $\bar{f}$. For instance, let f_k be the hinge loss for the k-th training data and ℓ be the squared 2-norm, we recover the support vector machines [23]. These special cases are convex problems, and have been extensively studied in the past. Instead, we will focus on the more general setting where both functions ℓ and $\{f_k\}$ are *nonconvex* and *non-smooth*. The motivation to have nonconvex functions can be diverse, and will be illustrated in Example 1 and Example 2 below. We emphasize that the function $\bar{f}$ is *non-separable*, in the sense that its components $\{f_k\}$ have overlapping argument $\mathbf{w}$.

In general, (1) can be very challenging to solve, even when we lower our expectation to the convergence to some critical point. Fortunately, practical problems usually come with some structure that we can (and should) exploit. In particular, we make the following assumption:

Assumption 1 *The function ℓ has L-Lipschitz continuous gradient $\nabla\ell$, each f_k is M_k-Lipschitz continuous (all w.r.t. the Euclidean norm $\|\cdot\|$), and the proximal map $\mathsf{P}^{\mu}_{f_k}$ can be computed "easily" for any $\mu > 0$.*

Recall that a mapping $g : \mathbb{R}^p \to \mathbb{R}^d$ is M-Lipschitz continuous for some $M > 0$ if for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$, $\|g(\mathbf{x}) - g(\mathbf{y})\| \leq M\|\mathbf{x} - \mathbf{y}\|$. The proximal map P^{μ}_{f} for any function f and parameter $\mu > 0$ is defined as:

$$\mathsf{P}^{\mu}_{f}(\mathbf{w}) = \operatorname{argmin}_{\mathbf{z}\in\mathbb{R}^p} \tfrac{1}{2\mu}\|\mathbf{z} - \mathbf{w}\|^2 + f(\mathbf{z}). \quad (2)$$

When f is the indicator function of some set C, $\mathsf{P}^{\mu}_{f}(\mathbf{w})$ simply returns the closest point in C to $\mathbf{w}$, namely the familiar Euclidean projection. If f is the 1-norm, then P^{μ}_{f} becomes the well-known soft-shrinkage operator that is widely used in sparse methods, e.g. lasso. The parameter $\mu > 0$ in the definition (2) plays the role of step size in the algorithms we will develop, and needs to be set properly. Assumption 1 requires the proximal map to be "easily" computable, meaning roughly that its complexity should be on par with that of computing the gradient of function ℓ. This avoids the proximal map to become the bottleneck if we use a gradient-type algorithm. As we will see, Assumption 1 is quite reasonable in a number of applications.

In the nonconvex setting, we need to pay extra care to even some "obvious" properties of the proximal map (such as non-emptiness and non-uniqueness). Such technicalities, albeit important, will be postponed until Section 3. For the purpose of explaining our main idea let us pretend momentarily that P^{μ}_{f} is a well-defined "function", i.e., $\mathsf{P}^{\mu}_{f}(\mathbf{w})$ is some "closest point" to $\mathbf{w}$, measured by the function f. With this "simplification" we can now demonstrate how Assumption 1 is naturally satisfied in some important applications:

Example 1 (GFlasso, [13]) *The graph-guided fused lasso exploits some graph structure to improve feature selection. Given some* a priori *graph whose nodes correspond to the feature variables, [13] used the regularizer $\tilde{f}_{ij}(\mathbf{w}) = |w_i - w_j|$ for every edge $(i, j) \in E$, to encourage connected nodes to be selected jointly. However, as pointed out in [13], this regularizer brings a large bias as it also requires connected nodes to have similar weights, which is likely not true in general. Here we reduce the bias by proposing the nonconvex regularizer: $f_{ij} = \min\{\tilde{f}_{ij}, \tau\}$, i.e., we cap the regularizer at the threshold τ. This will allow some weights to differ significantly without getting heavily penalized. In this example ℓ is the least squares loss.*

Algorithm 1 PA-PG.

1: Initialize $\mathbf{w}_0$, μ.
2: **for** $t = 1, 2, \ldots$ **do**
3: $\quad \mathbf{z}_t = \mathbf{w}_{t-1} - \mu \nabla \ell(\mathbf{w}_{t-1})$,
4: $\quad \mathbf{w}_t \in \overline{\sum_k \alpha_k \cdot \hat{\mathsf{P}}^{\mu}_{f_k}}(\mathbf{z}_t)$.
5: **end for**

Algorithm 2 PA-APG.

1: Initialize $\mathbf{w}_0 = \mathbf{u}_1$, μ, $\eta_1 = 1$.
2: **for** $t = 1, 2, \ldots$ **do**
3: $\quad \mathbf{z}_t = \mathbf{u}_t - \mu \nabla \ell(\mathbf{u}_t)$,
4: $\quad \mathbf{w}_t \in \overline{\sum_k \alpha_k \cdot \hat{\mathsf{P}}^{\mu}_{f_k}}(\mathbf{z}_t)$,
5: $\quad \eta_{t+1} = \frac{1+\sqrt{1+4\eta_t^2}}{2}$,
6: $\quad \mathbf{u}_{t+1} = \mathbf{w}_t + \frac{\eta_t - 1}{\eta_{t+1}}(\mathbf{w}_t - \mathbf{w}_{t-1})$.
7: **end for**

It is easily verified that each f_{ij} is $\sqrt{2}$-Lipschitz continuous w.r.t. the Euclidean norm. Moreover, the proximal map $\mathsf{P}^{\mu}_{f_{ij}}$ can be computed in closed-form (see Appendix J for the detailed derivation): For $s \in \{i, j\}$, trivially $[\mathsf{P}^{\mu}_{f_{ij}}(\mathbf{w})]_s = w_s$, while for $\{s, t\} = \{i, j\}$,

$$[\mathsf{P}^{\mu}_{f_{ij}}(\mathbf{w})]_s = w_s - \operatorname{sign}(w_s - w_t) \min\{\mu\eta, \tfrac{|w_i - w_j|}{2}\},$$

$$\eta = \begin{cases} 0, & |w_i - w_j| \geq 2\sqrt{\mu\tau} + \big((\sqrt{\tau} - \sqrt{\mu})_+\big)^2 \\ 1, & |w_i - w_j| \leq 2\sqrt{\mu\tau} + \big((\sqrt{\tau} - \sqrt{\mu})_+\big)^2 \end{cases}.$$

Roughly speaking, $\eta = 0$ iff $|w_i - w_j|$ is large, and consequently $[\mathsf{P}^{\mu}_{f_{ij}}(\mathbf{w})]_s = w_s$, i.e., the algorithm gives up "fusing" w_i and w_j, which can be beneficial in reducing the estimation bias when w_i and w_j are truly different.

The next example swaps the role of the loss ℓ and the regularizer f_k, demonstrating the flexibility of the general formulation (1).

Example 2 (Robust SVM, [18–20]) *Support vector machine (SVM) is one of the most popular algorithms for binary classification. However, it is known not to be robust against outliers [16–18, 20]. In fact, [16] constructed an example on which all algorithms based on convex losses fail. Instead, [18, 20] proposed the (nonconvex) truncated hinge loss as a robust alternative: $f_i(\mathbf{w}) = \min\{\tau, (1 - y_i \mathbf{x}_i^\top \mathbf{w})_+\}$. It is easy to verify that f_i is $\|\mathbf{x}_i\|$-Lipschitz continuous w.r.t. the Euclidean norm, and its proximal map can be computed as (see Appendix K for the detailed derivation):*

$$\mathsf{P}^{\mu}_{f_i}(\mathbf{w}) = \mathbf{w} + \left[\tfrac{1 - y_i \mathbf{w}^\top \mathbf{x}_i}{\mathbf{x}_i^\top \mathbf{x}_i}\right]_0^{\mu\eta} \cdot y_i \mathbf{x}_i, \tag{3}$$

where $[\cdot]_0^{\mu}$ denotes the projection onto the interval $[0, \mu]$, and the parameter $\eta \in \{0, 1\}$ is explicitly given in (33) in the appendix. Roughly speaking, $\eta = 0$ iff the margin $y_i \mathbf{x}_i^\top \mathbf{w}$ is small, i.e., the pair $(\mathbf{x}_i, y_i)$ is likely to be an outlier, in which case $\mathsf{P}^{\mu}_{f_i}(\mathbf{w}) = \mathbf{w}$, i.e. the algorithm "refuses" to update the weight. In this example ℓ is the (multiple of the) squared 2-norm.

The two examples represent two extremes in applications: the former has a nonconvex *non-separable* regularizer while the loss is the simple least squares, and the latter has a nonconvex *non-separable* loss while the regularizer is the simple (squared) 2-norm. Of course it is possible to have other combinations (e.g. the overlapping group lasso [8]), but for illustration purpose we shall contend ourselves with the above examples.

Having demonstrated the relevance of problem (1), we now turn to how to solve it efficiently. The main difficulty, apart from the nonconvexity, is the non-separability of the functions $\{f_k\}$: they all share the same weight $\mathbf{w}$. Accordingly, the coordinate descent algorithm of [11] cannot be efficiently applied. Similarly, the (block) proximal gradient (PG) algorithm, such as those in [9, 10, 12, 15], cannot be directly applied either, because we do not know how to efficiently compute the proximal map $\mathsf{P}^{\mu}_{\bar{f}}$, even when we assume each $\mathsf{P}^{\mu}_{f_k}$ is easy to compute. Other possible algorithms include the alternating strategy [18] and the convex-concave procedure [8, 20]. However, due to successive linearizations, these algorithms can be slow, and normally would need the functions f_k to be smooth.

Our idea is to approximate the proximal map $\mathsf{P}^{\mu}_{\bar{f}}$ using the linearization:

$$\mathsf{P}^{\mu}_{\bar{f}} \quad \approx \quad \sum_k \alpha_k \mathsf{P}^{\mu}_{f_k}. \tag{4}$$

Pretending $\mathsf{P}^{\mu}_{f_k}(\mathbf{w})$ is a single "point", we can plug the right-hand approximation into the PG algorithm. The resulting algorithm (PA-PG), summarized in Algorithm 1, can now be used to solve (1). It is extremely simple: alternating between a standard gradient step w.r.t. the loss ℓ and a proximal step[1] w.r.t. the regularizers $\{f_k\}$. Although the approximation (4) may seem overly naive, linearizing a nonlinear object (at least "locally") is a ubiquitously useful technique (such as the Taylor expansion in calculus). Indeed, for convex functions ℓ and $\{f_k\}$, the recent work [14] gave a formal justification of Algorithm 1 and also the accelerated variation in Algorithm 2. In our later experiments we found Algorithm 2 to be again more effective than Algorithm 1 even for nonconvex functions.

We aim to generalize the nice results of [14] to the current nonconvex setting and demonstrate its effectiveness through experiments. This goal, as clear as it

[1]The big overbar and hat notation will be understood after we present relevant technical results in Section 3.

is, is far from trivial though. The difficulties we face include: a). How should we interpret the sum in (4), considering that for nonconvex f_k, $\mathsf{P}^{\mu}_{f_k}$ may no longer be single-valued? b). Does the right-hand side still correspond to a proximal map of some function? c). Can we formally justify the linearization? d). What guarantee does (4) enjoy if plugged into the proximal gradient algorithm? These questions cannot be answered by existing works, such as [14] which relies entirely on convexity, or [15] which relies entirely on separability (i.e., the functions f_k have non-overlapping arguments). A substantial technical development is needed, which we do in Section 3.

Before going into the technical details, let us point out a few practical advantages of Algorithm 1 and Algorithm 2: 1). As a general meta-algorithm, they can be used in a variety of settings. 2). For *separable* functions, they reduce to the block PG algorithm [15] while for $K = 1$ we recover the algorithms in [9, 10, 12], including the popular FISTA [24, 25] when convexity is present. 3). They can be easily parallelized when K is large (such as the RSVM example). 4). The iterates they generate converge to a critical point (globally optimal if convexity is assumed). These nice properties are not shared by too many algorithms, let alone simultaneously.

3 Technical Results

To justify our new algorithm, we need a few technical tools from variational analysis [26]. We equip $\mathbb{R}^p$ with the usual inner product $\langle\cdot,\cdot\rangle$ and the induced Euclidean norm $\|\cdot\|$. For any *closed*[2] function f (not necessarily convex), its Moreau envelope (with parameter $\mu > 0$) is defined as [26]:

$$\mathsf{e}^{\mu}_{f}(\mathbf{w}) = \inf_{\mathbf{z}} \tfrac{1}{2\mu}\|\mathbf{w}-\mathbf{z}\|^2 + f(\mathbf{z}), \tag{5}$$

and the proximal map is the corresponding minimizer(s), see (2). One can roughly think the envelope function e^{μ}_{f} as a "regularized" version of f. For instance, if f is the 1-norm, then e^{μ}_{f} is the celebrated Huber's function in robust statistics [17]. Since we have stepped out of the convex domain, many "obvious" properties, such as well-definedness, smoothness, and uniqueness, can no longer be taken for granted. Fortunately, many appealing properties retain, possibly under an alternative interpretation.

It can be shown that P^{μ}_{f} is nonempty-valued iff f majorizes some quadratic function and μ is small [26]. Here, for simplicity, we assume throughout that f is bounded from below so that we need not restrict μ. Thus, $\mathsf{P}^{\mu}_{f} : \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ is nonempty-, compact-, possibly nonconvex- and multi-valued. In fact, Lemma 1 in Appendix C showed that the proximal map $\mathsf{P}^{\mu}_{f}(\mathbf{w})$ is single valued iff the envelope function e^{μ}_{f} is differentiable at $\mathbf{w}$. Both are trivially true for convex functions, but they may fail for general nonconvex functions, particularly functions that are "capped" at a certain value—our running examples in this paper. When $\mu \downarrow 0$, $\mathsf{e}^{\mu}_{f}(\mathbf{w}) \uparrow f(\mathbf{w})$ for all $\mathbf{w}$ [26]. As mentioned before, the proximal map is the key component of the proximal gradient algorithm.

In the nonsmooth and nonconvex setting, the usual gradient or subgradient no longer applies to characterize critical points. Instead, we are forced to "localize". We first define the regular (or Frechét) subdifferential $\hat{\partial} f(\mathbf{w})$ at $\mathbf{w}$, as the collection of vectors $\mathbf{v}$ such that

$$\forall\, \mathbf{z},\ f(\mathbf{z}) \geq f(\mathbf{w}) + \langle \mathbf{z}-\mathbf{w}, \mathbf{v}\rangle + o(\|\mathbf{z}-\mathbf{w}\|),$$

where the little-o term signifies a local neighborhood. Since $\hat{\partial} f$ can be empty even for Lipschitz functions (e.g. $-|\cdot|$ at the origin), we take its "closure" to avoid this degeneracy, arriving at the subdifferential $\partial f(\mathbf{w})$:

$$\{\mathbf{v} : \exists \mathbf{w}_n \to \mathbf{w}, f(\mathbf{w}_n) \to f(\mathbf{w}), \mathbf{v}_n \in \hat{\partial} f(\mathbf{w}_n), \mathbf{v}_n \to \mathbf{v}\}.$$

Clearly, $\hat{\partial} f(\mathbf{w}) \subseteq \partial f(\mathbf{w})$ for all $\mathbf{w}$. Pleasantly, if f is (resp. continuously) differentiable at $\mathbf{w}$, then $\hat{\partial} f(\mathbf{w})$ (resp. $\partial f(\mathbf{w})$) coincides with the usual derivative. From the definition it follows that if $\mathbf{w}$ is a local minimizer, then $0 \in \hat{\partial} f(\mathbf{w}) \subseteq \partial f(\mathbf{w})$, which generalizes the familiar Fermat's rule. We will be interested in finding (asymptotically) some $\mathbf{w}$ so that $0 \in \partial f(\mathbf{w})$, i.e., the critical points of f.

We reassure ourselves some nice properties of the Moreau envelope and the proximal map in Appendix A. In particular, we proved that any Moreau envelope is concave after subtracting the function $\frac{1}{2\mu}\|\cdot\|^2$. The converse, which we prove next, will allow us to "average" functions in a somewhat peculiar but computationally appealing way.

Let SCV_{μ} be the set of *finite-valued* μ-semiconcave functions, that is, functions $f : \mathbb{R}^p \to \mathbb{R}$ such that $f - \frac{1}{2\mu}\|\cdot\|^2$ is concave. For conciseness, denote CPB the class of closed, proper, and bounded from below functions. The next result, whose proof is deferred to Appendix B, significantly extends [14, Proposition 2]:

Proposition 1 *Fix $\mu > 0$ and $f \in \mathsf{CPB}$. Then $f = \mathsf{e}^{\mu}_{g}$ for some function $g \in \mathsf{CPB}$ iff $f \in \mathsf{SCV}_{\mu}$. Moreover, the Moreau envelope map $\mathsf{e}^{\mu} : \mathsf{CPB} \to \mathsf{SCV}_{\mu}$ that sends $f \in \mathsf{CPB}$ to e^{μ}_{f} is increasing, and concave on any convex subset of CPB (under the pointwise order).*

Note that in the nonconvex setting, the Moreau envelope (for any fixed $\mu > 0$) is no longer injective (see

[2] The function $f : \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$ is closed iff its epigraph $\{(\mathbf{x}, t) \in \mathbb{R}^p \times \mathbb{R} : f(\mathbf{x}) \leq t\}$ is a closed set.

Example 3 below). It is clear that SCV_μ is a *convex set*. Therefore we can average μ-semiconcave functions and still be able to find a pre-image under the Moreau envelope map, thanks to Proposition 1. This leads to the generalized notion of the proximal average. Specifically, recall that $\bar{f} = \sum_k \alpha_k f_k$ with each $f_k \in \mathsf{CPB}$, i.e. $\bar{f}$ is the convex combination of the component functions $\{f_k\}$ under the weight $\{\alpha_k\}$. Note that we always assume $\bar{f} \in \mathsf{CPB}$.

Definition 1 (Proximal average) *The proximal average A^μ is any function g such that $\mathsf{e}_g^\mu = \sum_{k=1}^K \alpha_k \mathsf{e}_{f_k}^\mu$. In face of non-uniqueness, we will always pick $\mathsf{A}^\mu = -\mathsf{e}_{\mathsf{M}}^\mu$, where $\mathsf{M} := -\sum_k \alpha_k \mathsf{e}_{f_k}^\mu$.*

The main idea behind this definition is to find some function whose Moreau envelope is simply the average $\sum_k \alpha_k \mathsf{e}_{f_k}^\mu$. Indeed, the existence of such a function follows from the surjectivity of e^μ, which we proved in Proposition 1. However, unlike the convex case, the proximal average for nonconvex functions need *not* be unique, and for concreteness we have picked a convenient representative in Definition 1. We remark that [27] used a slightly different definition for the sake of pursuing smoothness; for us the current definition is more useful. Note that for any function f, the so-called μ-proximal hull $\mathsf{h}_f^\mu := -\mathsf{e}_{(-\mathsf{e}_f^\mu)}^\mu$ has the same Moreau envelope as f but need not coincide with f [26] (while for convex functions f, always $\mathsf{h}_f^\mu = f$). One easily verifies, through the proximal hull, that our particular choice in Definition 1 is indeed legitimate (and will prove convenient later, see Proposition 5).

To facilitate our discussions, let us first prove some results that are interesting in their own rights. For any multi-valued map $\mathsf{P} : \mathbb{R}^p \rightrightarrows \mathbb{R}^p$, we define its closure $\bar{\mathsf{P}} : \mathbb{R}^p \rightrightarrows \mathbb{R}^p$, $\mathbf{w} \mapsto \{\mathbf{z} : \exists \{\mathbf{w}_n, \mathbf{z}_n\} \to (\mathbf{w}, \mathbf{z}), \mathbf{z}_n \in \mathsf{P}(\mathbf{w}_n)\}$, i.e., the graph of the closure $\bar{\mathsf{P}}$ is simply the closure of the graph of P. By "closing" a map we gain some continuity property. Also define $\hat{\mathsf{P}}(\mathbf{w}) = \mathsf{P}(\mathbf{w})$ at points $\mathbf{w}$ where $\mathsf{P}(\mathbf{w})$ is single-valued and empty otherwise.

Definition 2 (Extremal proximal maps) *Define the limiting proximal map $\mathsf{L}_f^\mu = \overline{\hat{\mathsf{P}}_f^\mu}$, and the hull proximal map $\mathsf{H}_f^\mu : \mathbb{R}^p \rightrightarrows \mathbb{R}^p, \mathbf{w} \mapsto \mathrm{conv}(\mathsf{P}_f^\mu(\mathbf{w}))$, where* conv *denotes the convex hull.*

Thanks to item iv) of Proposition 7 (in Appendix A) and [26, Theorem 1.25], we know $\emptyset \neq \mathsf{L}_f^\mu(\mathbf{w}) \subseteq \mathsf{P}_f^\mu(\mathbf{w}) \subseteq \mathsf{H}_f^\mu(\mathbf{w})$ for all $\mathbf{w}$. The inclusion can be strict, as we will see shortly. It may help to keep in mind that L_f^μ and H_f^μ are the smallest and biggest proximal map "compatible" with the function f, respectively. In Appendix C we prove many new structural properties of these different notions of proximal maps.

The next result characterizes exactly when Moreau envelopes coincide (proof in Appendix D).

Proposition 2 *Fix $\mu > 0$. For any $f \in \mathsf{CPB}$, there exist $\mathsf{h}_f^\mu, \ell_f^\mu \in \mathsf{CPB}$ such that for any $g \in \mathsf{CPB}$, $\mathsf{e}_g^\mu = \mathsf{e}_f^\mu + c$ for some constant c iff $\mathsf{h}_f^\mu \leq g - c \leq \ell_f^\mu$ iff $\mathsf{P}_{\ell_f^\mu}^\mu(\mathbf{w}) \subseteq \mathsf{P}_g^\mu(\mathbf{w}) \subseteq \mathsf{H}_f^\mu(\mathbf{w})$ for all $\mathbf{w}$.*

In fact, ℓ_f^μ is the restriction of h_f^μ onto some closed set. Their explicit forms can be found in the proof. It is also true that $\mathsf{P}_{\ell_f^\mu}^\mu = \mathsf{L}_f^\mu$ on the real line. Using Proposition 2 we can easily characterize when the proximal average is unique (essentially our particular choice in Definition 1 plays the role of h_f^μ). It also leads to the following result that completely characterizes proximal maps on the real line (proof in Appendix D):

Proposition 3 *The map $\mathsf{P} : \mathbb{R} \rightrightarrows \mathbb{R}$ is a proximal map iff it is (nonempty) compact-valued, monotone, and has a closed graph. Moreover, there is a unique function (up to addition of a constant) f such that $\mathsf{P}_f = \mathsf{P}$ iff P is also convex-valued.*

Thus, both the SCAD [2] and the MC+ [3] thresholding rules correspond to a *unique* regularization function. In contrast, there are infinitely many different regularizers that all lead to the hard thresholding rule, see Example 3. Importantly, Proposition 3 allows us to directly design the proximal map (thresholding rule), without even the need to refer to the regularizer f!

We now come to the main result for justifying Algorithm 1. Recall that the main property of the proximal average, as seen from its definition, is that its Moreau envelope is the convex combination of the Moreau envelopes of the component functions. We wish to say something similar for its proximal map. Indeed, this is possible after an appropriate modification (proof in Appendix E):

Proposition 4 *For all $\mathbf{w}$, $\emptyset \neq \overline{\sum_k \alpha_k \hat{\mathsf{P}}_{f_k}^\mu}(\mathbf{w}) \subseteq \left[\mathsf{P}_{\mathsf{A}^\mu}^\mu(\mathbf{w}) \bigcap \left(\sum_{k=1}^K \alpha_k \mathsf{P}_{f_k}^\mu(\mathbf{w})\right)\right]$.*

Recall that the middle term is exactly the approximation we employed in Section 2. Unlike the convex case, we have to replace the simpler average $\sum_{k=1}^K \alpha_k \mathsf{P}_{f_k}^\mu$ with the slightly more complicated closure $\overline{\sum_k \alpha_k \hat{\mathsf{P}}_{f_k}^\mu}$, due to the possible multi-valuedness of $\mathsf{P}_{f_k}^\mu$. Indeed, some element in $\sum_{k=1}^K \alpha_k \mathsf{P}_{f_k}^\mu(\mathbf{w})$ may not be in $\mathsf{P}_{\mathsf{A}^\mu}^\mu(\mathbf{w})$, which itself may change if we use a different proximal average. Proposition 4 avoids such pathology by always picking a common element. Moreover, in Section 4 we prove Algorithm 1 converges to a critical point of the proximal average, which itself may not even be unique! This ambiguity is resolved using Proposition 4, which guarantees all realizations

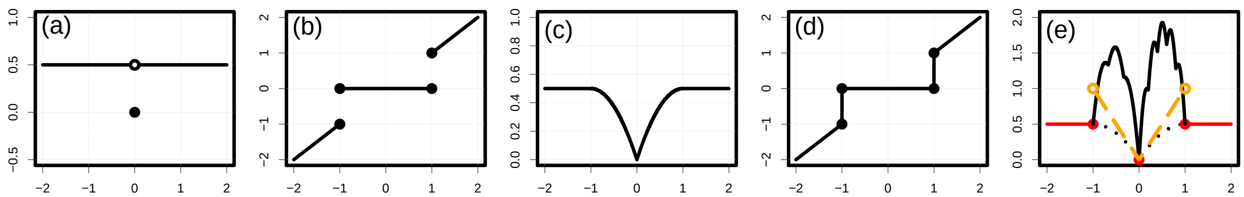

Figure 1: (a): the ℓ_0 function; (b): hard-thresholding operator; (c): proximal hull $\mathsf{h}_{|\cdot|}$; (d): proximal map of $\mathsf{h}_{|\cdot|}$; (e): solid: another (continuous) function that has (b) as proximal map, dashed (+ red solid): function g. See Example 3 for the explanations and formulas. Throughout $\mu = 1, \lambda = 1$.

of the proximal average coincide along the trajectory of Algorithm 1.

On the real line, thanks to Proposition 3, $\overline{\sum_k \alpha_k \hat{\mathsf{P}}^{\mu}_{f_k}} = \sum_k \alpha_k \mathsf{L}^{\mu}_{f_k}$, with the latter being readily available. For our Example 1 and Example 2, the computations also reduce to the real line (see Appendix J and Appendix K, respectively), hence can be easily addressed.

Let us now demonstrate some pathology of the proximal map, using a familiar nonconvex function.

Example 3 *Consider the cardinality function on the real line $|x|_0 = \frac{\lambda^2}{2}\mathbf{1}_{x\neq 0}$. Its proximal map (with $\mu = 1$) is the well-known hard-thresholding operator:*

$$\mathsf{P}_{|\cdot|_0}(x) = \begin{cases} x, & |x| > \lambda \\ \{0, x\}, & |x| = \lambda \\ 0, & |x| < \lambda \end{cases}. \tag{6}$$

In the literature, the above proximal map at $|x| = \lambda$ is usually set to 0. Mathematically, this is not precise and possibly confusing: If $\mathsf{P}_{|\cdot|_0}$ was single-valued at $|x| = \lambda$, then $\mathsf{e}_{|\cdot|_0}(x) = \frac{1}{2}\min\{\lambda^2, x^2\}$ would be continuously differentiable, which is not true. Interestingly, without the tools we have developed so far, [1] noticed that the functions

$$h(x) := \tfrac{1}{2}(\lambda^2 - (|x| - \lambda)^2 \mathbf{1}_{|x|\leq\lambda}) \tag{7}$$

$$g(x) := \lambda|x|\mathbf{1}_{|x|<\lambda} + \tfrac{\lambda^2}{2}\mathbf{1}_{|x|\geq\lambda} \tag{8}$$

also have (6) as their (limiting) proximal map. We verify that the former is exactly the proximal hull $\mathsf{h}_{|\cdot|_0}$. Applying Proposition 2 we know any function $f \geq \mathsf{h}_{|\cdot|_0}$ with equality on the closed set $[-\infty, -\lambda] \cup [\lambda, \infty] \cup \{0\}$ will have (6) as its limiting proximal map. Furthermore, if f is strictly larger than $\mathsf{h}_{|\cdot|_0}$ on $]-\lambda, \lambda[$, such as the function g above, then according to Lemma 7 (in Appendix C) it has the same proximal map as the cardinality function! See Figure 1 for the illustrations.

Statistically, one is often interested in the hard-thresholding rule (6), rather than the cardinality function itself [1, 2, 10]. Figure 1 shows that there are actually infinitely many functions that all yield the same proximal map (6). This observation suggests that we should not base our algorithm on any particular function form but on the proximal map directly (which is less ambiguous). In this sense the proximal gradient algorithm seems to be a well fit. Similar conclusions have been made in [10]. We point out that the lessons we learned from this example extend to most nonconvex regularizers therefore deserve some attention.

To provide a strong convergence guarantee for Algorithm 1, we will (and perhaps should) restrict the (nonconvex and nonsmooth) functions under our consideration, for otherwise they can behave very pathologically. To do so we recall some notions from semi-algebraic geometry [28]. A set $A \subseteq \mathbb{R}^p$ is semi-algebraic if it is the *finite* unions of *finite* intersections of the sets $\{\mathbf{w} \in \mathbb{R}^p : p_0(\mathbf{w}) = 0, p_1(\mathbf{w}) < 0\}$, where p_0, p_1 are polynomials with real coefficients. For instance, hyperplanes, halfspaces, spheres, ellipsoids, the positive semi-definite cone, are all semi-algebraic. The most striking property of semi-algebraic sets is that their intersection with any line is the union of *finitely* many points and open intervals (due to the fact that any polynomial admits only finitely many roots). Thus, for instance, the set of all natural numbers is *not* semi-algebraic. A function $f : \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$ is semi-algebraic iff its graph $\{(\mathbf{w}, f(\mathbf{w})) : \mathbf{w} \in \operatorname{dom} f\}$ is a semi-algebraic set. For instance, all power functions with rational exponent and all polyhedral functions are semi-algebraic. On one hand, semi-algebraic functions are extremely well-structured, allowing one to prove many strong results; on the other hand, they appear very naturally in various applications, such as the ones we consider here: All functions appear in our experiments are semi-algebraic. Note that the exponential function and power functions with irrational exponent are not semi-algebraic. They belong to the more general class of definable functions[3]. For brevity, we omit the relevant definitions here but refer to the nice article [28]. Since all of our results extend to definable functions (w.r.t. some order-minimal structure that contains all semi-algebraic functions), we will use the term definable freely below; for all practical purposes, one can simply think of "definable" as semi-algebraic.

[3]In case one wonders, there do exist non-definable (even convex) functions, which oscillate infinitely often. Such functions are unlikely to be useful in applications though.

Definable functions are closed under all familiar algebraic operations. For instance, the sum, product, and composition of definable functions are definable, respectively. So is the scalar multiple and the inverse. Moreover, the following result is useful to us (proof in Appendix F).

Proposition 5 *The function f is definable iff $\mathsf{e}_f^\mu(\mathbf{w})$, as a function of $(\mathbf{w}, \mu)$, is definable on $\mathbb{R}^p \times \mathbb{R}_{++}$. In particular, the proximal average (cf. Definition 1) of definable functions is definable.*

As we mentioned before, when $\mu \downarrow 0$, $\mathsf{e}_f^\mu \uparrow f$ pointwise [26]. Under the Lipschitz assumption, we can strengthen the convergence to be uniform (Proof in Appendix G):

Proposition 6 *Under Assumption 1 we have $0 \le \bar{f} - \mathsf{A}^\mu \le \bar{f} - \sum_k \alpha_k \mathsf{e}_{f_k}^\mu \le \frac{\mu}{2}\sum_{k=1}^K \alpha_k M_k^2$.*

Similar as the convex case, we see that the proximal average A^μ is a better under-approximation to $\bar{f}$ than the average of Moreau envelopes, i.e. $\sum_k \alpha_k \mathsf{e}_{f_k}^\mu$.

4 Theoretical Justification

Given our development in the previous section, it is now clear that Algorithm 1 aims at solving the approximate problem:

$$\min_{\mathbf{w}} \ \ell(\mathbf{w}) + \mathsf{A}^\mu(\mathbf{w}). \tag{9}$$

The next important pieces are to show a). Algorithm 1 converges for the approximate problem (9); b). The approximate problem (9) is reasonably "close" to the original problem (1). Indeed, for the first piece, we have the following result (proof in Appendix H):

Theorem 1 *Let Assumption 1 hold and the functions ℓ and $\{f_k\}$ be definable. Choose $\mu < 1/L$, then Algorithm 1 converges to a critical point of (9), provided that the iterates are bounded.*

The last assumption is trivially met if, say the objective in (9) has bounded sublevel sets. To appreciate the significance of Theorem 1, let us consider a simple example: Assume say both 1 and -1 are critical points of our problem, then any limit point of the iterate sequence $\{1, -1, 1, -1, \ldots\}$ is indeed critical, but the whole sequence does not converge at all. This behavior can happen for the coordinate descent algorithm [11] or the convex-concave procedures (cccp) [8, 19, 20], but is eliminated for our algorithm, thanks to Theorem 1.

To fulfill our second piece, we need a notion of approximate minimizer. We call $\mathbf{w}$ an ϵ-local minimizer of f if there exists some neighborhood $\mathcal{N}$ of $\mathbf{w}$ such that for all $\mathbf{z} \in \mathcal{N}$, $f(\mathbf{w}) \le f(\mathbf{z}) + \epsilon$. Of course, when $\epsilon = 0$, we retrieve the usual notion of local minimizer. Then we have (proof in Appendix I):

Theorem 2 *Let Assumption 1 hold. Fix the accuracy $\epsilon > 0$ and choose $\mu < \min\{1/L, 2\epsilon/\sum_k \alpha_k M_k^2\}$. If Algorithm 1 converges to an ϵ-local minimizer of (9), $\tilde{\mathbf{w}}$, then $\tilde{\mathbf{w}}$ is also a (2ϵ)-local minimizer of (1). Same is true if $\tilde{\mathbf{w}}$ is in fact an ϵ-global minimizer.*

It is possible to prove that locally Algorithm 1 converges at a rate no slower than sublinear (when all functions are semi-algebraic). Moreover, if we let $\mu \downarrow 0$, we can prove that the iterates converge to a critical point of the original problem (1). In experiments, we found that a relatively small μ already yields satisfying results, therefore we omit the rather technical discussions.

5 Experiments

We evaluate our algorithm on two application domains: truncated GFlasso (Example 1) and robust SVM (Example 2). We demonstrate the benefits of nonconvex formulations by comparing with the convex counterparts, and we verify the effectiveness of our proposed algorithm against alternative optimization methods such as alternating coordinate descent (`alter`) [18] and the convex-concave procedure (`cccp`) [19]. We found that Algorithm 2 is always faster than Algorithm 1 in all our experiments so only the former (denoted as `proxavg`) is included here.

5.1 Multi-task GFlasso (Example 1)

Formally, the multi-task graph-guided fused lasso model is given by:

$$\begin{aligned}\tfrac{1}{2}\|Y - XW\|_F^2 + \lambda \textstyle\sum_{j=1}^q \phi(\mathbf{w}_j) \\ + \gamma \textstyle\sum_{(j,k)\in E} \omega_{jk}\, \psi(\mathbf{w}_j - \mathrm{sign}(\omega_{jk})\mathbf{w}_k),\end{aligned} \tag{10}$$

where $\mathbf{w}_j \in \mathbb{R}^p$ is the j-th column of W. Here ϕ is a regularizer that encourages sparsity among the elements of $\mathbf{w}_j$, and ψ is a regularizer that encourages fusion between the elements of $\mathbf{w}_j$ and $\mathbf{w}_k$ when output variables j and k are connected in the graph. In our experiment we use $\phi(\mathbf{u}) = \|\mathbf{u}\|_1$ and $\psi(\mathbf{u}) = \sum_i \min\{|u_i|, \tau\}$. If $\tau = 0$, we recover the multi-task lasso while if $\tau = \infty$, we recover the GFlasso which is convex. For any $\tau > 0$, the fusion regularizer is non-separable and nonconvex.

We compare different methods (corresponding to different τ's) on a synthetic data in which pairs of correlated output variables, $\mathbf{y}_j$ and $\mathbf{y}_k$, have similar weights, $\mathbf{w}_j$ and $\mathbf{w}_k$. We choose a block correlation graph for concreteness and generate the data as follows. First, partition the output variables $\mathbf{y}_1, \ldots, \mathbf{y}_q$ into disjoint

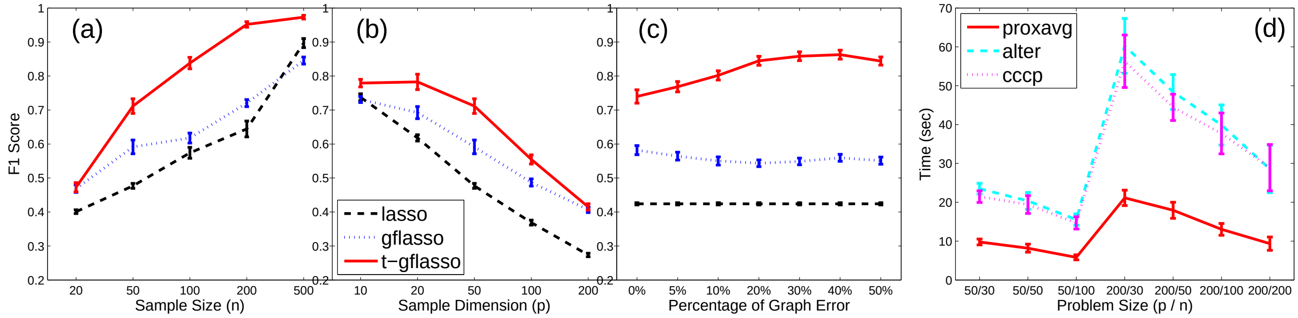

Figure 2: The F1 score under (a): varying sample size (n); (b): varying dimensionality (p); (c): varying graph error. (d): Speed comparison between our algorithm (proxavg), alternating, and cccp.

ccat	`proxavg`	`alter`	`cccp`	svm	`sparse`
training time (sec)	25.34	897.98	1151.26	69.85	110.21
test accuracy (%)	91.12	84.77	87.02	69.96	89.20
detected outlier (%)	9.94	20.12	18.45	-	10.47

Table 1: Results on the CCAT data set.

groups. Next, generate the sparsity pattern for the weight matrix W by assigning the same (randomly chosen) set of active input features to all the output variables in each group. The nonzero entries of W are drawn from Uniform$(0.4, 0.8)$, but all variables in the same group are given the same weight for each feature. Draw $X \sim \mathcal{N}(0, I)$ and $Y \sim \mathcal{N}(XW, \sigma^2 I)$. Finally, generate the correlation graph E over the output variables by thresholding the sample covariance matrix at some value ν. We also test the robustness of the algorithms by randomly changing certain percentage of the edge weights.

A series of experimental results are shown in Figure 2. Regularization parameters λ, γ, and τ (whenever applicable) are selected by optimizing the prediction error on a held-out set. We observe that a). The nonconvex fusion regularizer ($0 < \tau < \infty$) consistently outperforms lasso ($\tau = 0$) and GFlasso ($\tau = \infty$) in terms of both feature selection (F1 score) and prediction error (not shown); b). Our algorithm is several times faster than both `alter` and `cccp`.

5.2 Robust SVM (Example 2)

We conducted experiments on two benchmark data sets. The Long-Servedio [16] dataset is a well-crafted synthetic data for testing robustness against label noise and delicate leverage points. We generate 10,000 training examples (each with 21 features) and randomly flip 10% labels. The other real dataset CCAT from RCV1 [29] contains 23,149 training examples (each with 47,152 features) and 781,265 test examples. Similarly, we randomly flip 10% of the labels in the training set, and scale the corresponding features by 10. We average the results with 10 repetitions and report them in Table 1. For both SVM, `alter` [18], and `cccp` [19], we use the state-of-the-art Liblinear solver [30]. Instead of the (squared) 2-norm regularizer, our algorithm extends easily to the 1-norm regularizer hence we also include `sparse` to further demonstrate the flexibility. In contrast, Liblinear cannot deal with the 1-norm regularizer.

We confirmed that SVM fails miserably on the Long-Servedio dataset (achieving 72.14% prediction error), while all other solvers (aimed for the robust SVM) achieve nearly perfect results and identify the correct amount of outliers. Our algorithm is fastest but the margin is small (due to the small size of the dataset). For the CCAT dataset, our algorithm not only achieves superior prediction accuracy but also much pronounced efficiency. Again, SVM severely suffers from outliers while `alter` and `cccp` are slow due to their sequential nature: multiple calls of the SVM solver can only be executed consecutively. Interestingly, with small sacrifice in accuracy and training time, `sparse`, using 1-norm regularizer in SVM, learns a model with only 4.8% nonzero entries, whereas the models learned by all other methods are at least 10 times denser. This could hugely reduce the test time—a critical requirement in some financial applications.

6 Conclusions

We successfully extended the proximal average proximal gradient algorithm into the nonconvex setting, through a careful examination of the now multi-valued proximal map. We proved that the whole sequence of iterates converges to a critical point. Experimentally, the proposed algorithm has shown much promise, and naive parallelizability makes it even more favorable. We intend to strengthen the convergence guarantee and develop a fully distributed implementation.

References

[1] Anestis Antoniadis. "Wavelets in Statistics: A Review." *Journal of the Italian Statistical Association*, vol. 2 (1997), pp. 97–130 (cit. on pp. 1, 6, 18).

[2] Jianqing Fan and Runze Li. "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties." *Journal of the American Statistical Association*, vol. 96, no. 456 (2001), pp. 1348–1360 (cit. on pp. 1, 5, 6, 16).

[3] Cun-Hui Zhang. "Nearly unbaised variable selection under minimax concave penalty." *Annals of Statistics*, vol. 38, no. 2 (2010), pp. 894–942 (cit. on pp. 1, 5, 16).

[4] Cun-Hui Zhang and Tong Zhang. "A general theory of concave regularization for high-dimensional sparse estimation problems." *Statistical Science*, vol. 27, no. 4 (2012), pp. 576–593 (cit. on p. 1).

[5] Zhaoran Wang, Han Liu, and Tong Zhang. "Optimal computational and statistical rates of convergence for sparse nonconvex learning problems." *The Annals of Statistics*, vol. 42, no. 6 (2014), pp. 2164–2201 (cit. on p. 1).

[6] Gilles Gasso, Alain Rakotomamonjy, and Stéphane Canu. "Recovering Sparse Signals With a Certain Family of Nonconvex Penalties and DC Programming." *IEEE Transactions on Signal Processing*, vol. 57, no. 12 (2009), pp. 4686–4698 (cit. on p. 1).

[7] Rick Chartrand. "Nonconvex splitting for regularized low-rank + sparse decomposition." *IEEE Transactions on Signal Processing*, vol. 60, no. 11 (2012), pp. 5810–5819 (cit. on p. 1).

[8] Yunzhang Zhu, Xiaotong Shen, and Wei Pan. "Simultaneous grouping pursuit and feature selection over an undirected graph." *Journal of the American Statistical Association*, vol. 108, no. 502 (2013), pp. 713–725 (cit. on pp. 1, 3, 7).

[9] Thomas Blumensath and Mike E. Davies. "Iterative thresholding for sparse approximations." *Journal of Fourier Analysis and Applications*, vol. 14 (2008), pp. 629–654 (cit. on pp. 1, 3, 4).

[10] Yiyuan She. "Thresholding-based iterative selection procedures for model selection and shrinkage." *Electronic Journal of Statistics*, vol. 3 (2009), pp. 384–415 (cit. on pp. 1, 3, 4, 6).

[11] Rahul Mazumder, Jerome H. Friedman, and Trevor Hastie. "SparseNet: Coordinate Descent With Nonconvex Penalties." *Journal of the American Statistical Association*, vol. 106, no. 495 (2011), pp. 1125–1138 (cit. on pp. 1, 3, 7).

[12] Pinghua Gong, Changshui Zhang, Zhaosong Lu, Jianhua Z. Huang, and Jieping Ye. "A General Iterative Shrinkage and Thresholding Algorithm for Non-convex Regularized Optimization Problems." In: *ICML*. 2013 (cit. on pp. 1, 3, 4).

[13] Seyoung Kim and Eric P. Xing. "Statistical Estimation of Correlated Genome Associations to a Quantitative Trait Network." *PLoS Genetics*, vol. 5, no. 8 (2009), pp. 1–18 (cit. on pp. 1, 2).

[14] Yaoliang Yu. "Better Approximation and Faster Algorithm Using the Proximal Average." In: *NIPS*. 2013 (cit. on pp. 1, 3, 4, 11, 17, 18, 20).

[15] Jérôme Bolte, Shoham Sabach, and Marc Teboulle. "Proximal alternating linearized minimization for nonconvex and nonsmooth problems." *Mathematical Programming, Series A*, vol. 146 (2014), pp. 459–494 (cit. on pp. 1, 3, 4, 18).

[16] Philip M. Long and Rocco A. Servedio. "Random Classification Noise Defeats All Convex Potential Boosters." *Machine Learning*, vol. 78, no. 3 (2010), pp. 287–304 (cit. on pp. 2, 3, 8).

[17] Yaoliang Yu, Özlem Aslan, and Dale Schuurmans. "A Polynomial-time Form of Robust Regression." In: *NIPS*. 2012 (cit. on pp. 2–4).

[18] Linli Xu, Koby Crammer, and Dale Schuurmans. "Robust support vector machine training via convex outlier ablation." In: *AAAI*. 2006 (cit. on pp. 2, 3, 7, 8).

[19] Ronan Collobert, Fabian Sinz, Jason Weston, and Léon Bottou. "Trading Convexity for Scalability." In: *ICML*. 2006, pp. 201–208 (cit. on pp. 2, 3, 7, 8).

[20] Yichao Wu and Yufeng Liu. "Robust truncated hinge loss support vector machines." *Journal of the American Statistical Association*, vol. 102, no. 479 (2007), pp. 974–983 (cit. on pp. 2, 3, 7).

[21] Yufeng Liu, Xiaotong Shen, and Hani Doss. "Multicategory ψ-learning and support vector machine: computational tools." *Journal of Computational and Graphical Statistics*, vol. 14, no. 1 (2005), pp. 219–236 (cit. on p. 2).

[22] Robert Tibshirani. "Regression Shrinkage and Selection Via the Lasso." *Journal of the Royal Statistical Society, Series B*, vol. 58 (1996), pp. 267–288 (cit. on p. 2).

[23] Corinna Cortes and Vladimir Vapnik. "Support-Vector Networks." *Machine Learning*, vol. 20 (1995), pp. 273–297 (cit. on p. 2).

[24] Amir Beck and Marc Teboulle. "A fast iterative shrinkage-thresholding algorithm for linear inverse problems." *SIAM Journal on Imaging Sciences*, vol. 2, no. 1 (2009), pp. 183–202 (cit. on p. 4).

[25] Yurii Nesterov. "Gradient methods for minimizing composite functions." *Mathematical Programming, Series B*, vol. 140 (2013), pp. 125–161 (cit. on p. 4).

[26] Ralph Tyrell Rockafellar and Roger J-B Wets. *Variational Analysis*. Springer, 1998 (cit. on pp. 4, 5, 7, 10–16, 18).

[27] Waren L. Hare. "A proximal average for nonconvex functions: A proximal stability perspective." *SIAM Journal on Optimization*, vol. 20, no. 2 (2009), pp. 650–666 (cit. on p. 5).

[28] Lou van den Dries and Chris Miller. "Geometric Categories and O-minimal Structures." *Duke Mathematical Journal*, vol. 84, no. 2 (1996), pp. 497–540 (cit. on pp. 6, 17).

[29] David D. Lewis, Yiming Yang, Tony G. Rose, Fan Li, G. Dietterich, and Fan Li. "RCV1: A new benchmark collection for text categorization research." *Journal of Machine Learning Research*, vol. 5 (2004), pp. 361–397 (cit. on p. 8).

[30] Cho-Jui Hsieh, Kai-Wei Chang, Chih-Jen Lin, S. Sathiya Keerthi, and S. Sundararajan. "A Dual Coordinate Descent Method for Large-scale Linear SVM." In: *ICML*. 2008, pp. 408–415 (cit. on p. 8).

Supplement for *Minimizing Nonconvex and Non-separable Functions*

We start with recalling some definitions needed throughout.

For general nonsmooth and nonconvex functions, the usual gradient or subgradient, of course, no longer applies. Fortunately, a suitable theory from variational analysis is available, e.g. [26]. For any closed function f, its regular (or Frechét) subdifferential at $\mathbf{w}$, $\hat{\partial} f(\mathbf{w})$, is the collection of vectors $\mathbf{v}$ such that

$$\forall\ \mathbf{z},\ f(\mathbf{z}) \geq f(\mathbf{w}) + \langle \mathbf{z} - \mathbf{w}, \mathbf{v} \rangle + o(\|\mathbf{z} - \mathbf{w}\|).$$

The last lower order term implies that the regular subdifferential is a local property of the function (as it should be in the absence of convexity). Unfortunately, $\hat{\partial} f$ can be empty at certain points even for Lipschitz continuous functions (e.g. $-|\cdot|$ at the origin). Taking an appropriate closure we can avoid this degeneracy and arrive at the subdifferential ∂f:

$$\mathbf{v} \in \partial f(\mathbf{w}) \iff \exists \mathbf{w}_n \to \mathbf{w}, f(\mathbf{w}_n) \to f(\mathbf{w}), \mathbf{v}_n \in \hat{\partial} f(\mathbf{w}_n), \mathbf{v}_n \to \mathbf{v}.$$

Clearly, $\hat{\partial} f(\mathbf{w}) \subseteq \partial f(\mathbf{w})$ for all $\mathbf{w}$. If f is (resp. continuously) differentiable at $\mathbf{w}$, then $\hat{\partial} f(\mathbf{w})$ (resp. $\partial f(\mathbf{w})$) coincides with the usual derivative. From the definition it follows that if $\mathbf{w}$ is a local minimizer, then $0 \in \hat{\partial} f(\mathbf{w})$ and $0 \in \partial f(\mathbf{w})$, which generalizes the familiar Fermat's rule. In the main text, we are interested in finding some $\mathbf{w}$ so that $0 \in \partial f(\mathbf{w})$, i.e., the critical points of f.

We caution that the subdifferential alone no longer characterizes the function (even in the presence of differentiability) [31], although such pathologies cannot happen for definable functions.

For any, not necessarily convex, function f, its Fenchel conjugate

$$f^*(\mathbf{z}) := \max_{\mathbf{w}} \langle \mathbf{w}, \mathbf{z} \rangle - f(\mathbf{w})$$

is always convex.

A Properties of the Moreau envelope and proximal map

Proposition 7 *Let $\mu, \lambda > 0$, f be a closed, proper, and bounded from below function, then*

i). $(\mathsf{e}_f^\mu)^* = f^* + \frac{\mu}{2} \|\cdot\|^2$;

ii). $\mathsf{e}_f^\mu \leq f$, $\inf_{\mathbf{w}} \mathsf{e}_f^\mu(\mathbf{w}) = \inf_{\mathbf{w}} f(\mathbf{w})$, $\operatorname{argmin}_{\mathbf{w}} \mathsf{e}_f^\mu(\mathbf{w}) = \operatorname{argmin}_{\mathbf{w}} f(\mathbf{w}) \subseteq \{\mathbf{w} : \mathbf{w} \in \mathsf{P}_f^\mu(\mathbf{w})\}$;

iii). $\mathbf{z} \in \mathsf{P}_f^\mu(\mathbf{w}) \implies \mathbf{w} \in \mathbf{z} + \mu \cdot \partial f(\mathbf{z})$, *and* $\partial \mathsf{e}_f^\mu(\mathbf{w}) \subseteq \frac{1}{\mu}(\mathbf{w} - \mathsf{P}_f^\mu(\mathbf{w}))$;

iv). Up to a (Lebesgue) null set, P_f^μ *is single-valued and* $\nabla \mathsf{e}_f^\mu(\mathbf{w}) = \frac{1}{\mu}(\mathbf{w} - \mathsf{P}_f^\mu(\mathbf{w}))$.

v). $\mathsf{e}_{\lambda f}^\mu(\mathbf{w}) = \lambda \mathsf{e}_f^{\lambda\mu}(\mathbf{w})$ *and* $\mathsf{P}_{\lambda f}^\mu(\mathbf{w}) = \mathsf{P}_f^{\lambda\mu}(\mathbf{w}) = \lambda \cdot \mathsf{P}_{f\lambda^{-1}}^\mu(\lambda^{-1}\mathbf{w})$ *for all* $\mathbf{w}$;

vi). $\mathsf{e}_{\mathsf{e}_f^\mu}^\lambda(\mathbf{w}) = \mathsf{e}_f^{\lambda+\mu}(\mathbf{w})$ *and* $\mathsf{P}_{\mathsf{e}_f^\mu}^\lambda(\mathbf{w}) \bigcap \left[\frac{\mu}{\lambda+\mu}\mathbf{w} + \frac{\lambda}{\lambda+\mu}\mathsf{P}_f^{\lambda+\mu}(\mathbf{w})\right] \neq \emptyset$ *for all* $\mathbf{w}$;

vii). $\mu \mathsf{e}_f^\mu + (\mu f + \frac{1}{2}\|\cdot\|^2)^* = \frac{1}{2}\|\cdot\|^2$;

Proof: The first item is the usual duality from infimal convolution to summation, see e.g. [26].

For item ii), setting $\mathbf{z} = \mathbf{w}$ in the definition of the Moreau envelope, we see $\mathsf{e}_f^\mu(\mathbf{w}) \leq f(\mathbf{w})$ for all $\mathbf{w}$. Similarly using the nonnegativity of $\frac{1}{2}\|\cdot\|^2$ we can prove $\inf_{\mathbf{w}} \mathsf{e}_f^\mu(\mathbf{w}) = \inf_{\mathbf{w}} f(\mathbf{w})$ and $\operatorname{argmin}_{\mathbf{w}} \mathsf{e}_f^\mu(\mathbf{w}) = \operatorname{argmin}_{\mathbf{w}} f(\mathbf{w})$. Moreover, if $\mathbf{w}$ is a global minimizer of f, then we claim $\mathbf{w} \in \mathsf{P}_f^\mu(\mathbf{w})$ for otherwise by choosing any $\mathbf{z} \in \mathsf{P}_f^\mu(\mathbf{w})$ we have $\mathsf{e}_f^\mu(\mathbf{z}) < \mathsf{e}_f^\mu(\mathbf{w})$, contradicting the fact that $\mathbf{w}$ is also a global minimizer of e_f^μ.

We come to item iii). If $\mathbf{z} \in \mathsf{P}_f^\mu(\mathbf{w})$, then using the (generalized) Fermat's rule we have $0 \in \frac{1}{\mu}(\mathbf{z} - \mathbf{w}) + \partial f(\mathbf{z})$. The fact that $\partial \mathsf{e}_f^\mu(\mathbf{w}) \subseteq \frac{1}{\mu}(\mathbf{w} - \mathsf{P}_f^\mu(\mathbf{w}))$ follows from the general calculus rule of subdifferentials, see e.g. [26, Theorem 10.13].

For item iv), we notice that any Moreau envelope is the difference of two finite-valued convex functions, which follows from vii) of Proposition 7 (and is proved below). Thanks to the Rademacher theorem (see e.g. [26, Theorem 9.60]), any Moreau envelope is differentiable up to a (Lebesgue) null set. But if e_f^μ is differentiable at $\mathbf{w}$, then $-\nabla \mathsf{e}_f^\mu(\mathbf{w}) = \hat\partial(-\mathsf{e}_f^\mu)(\mathbf{w}) = \frac{1}{\mu}(\mathrm{conv}(\mathsf{P}_f^\mu(\mathbf{w})) - \mathbf{w})$, see e.g. [26, Example 10.32]. Thus $\mathsf{P}_f^\mu(\mathbf{w})$ is a singleton and $\nabla \mathsf{e}_f^\mu(\mathbf{w}) = \frac{1}{\mu}(\mathbf{w} - \mathsf{P}_f^\mu(\mathbf{w}))$ up to a null set.

Item v) is the result of simple algebraic manipulations.

For the first claim in vi), use the definition:

$$
\begin{aligned}
\mathsf{e}_{\mathsf{e}_f^\mu}^\lambda(\mathbf{w}) &= \inf_{\mathbf{z}} \tfrac{1}{2\lambda} \|\mathbf{w} - \mathbf{z}\|^2 + \mathsf{e}_f^\mu(\mathbf{z}) \\
&= \inf_{\mathbf{z}} \inf_{\mathbf{u}} \tfrac{1}{2\lambda} \|\mathbf{w} - \mathbf{z}\|^2 + \tfrac{1}{2\mu} \|\mathbf{z} - \mathbf{u}\|^2 + f(\mathbf{u}).
\end{aligned}
$$

Fix $\mathbf{u}$ and minimize $\mathbf{z}$ we obtain $\mathbf{z} = \frac{\mu\mathbf{w}+\lambda\mathbf{u}}{\mu+\lambda}$. Plug it back in and simplify we verify the claim. The second claim follows from taking the subdifferential on both sides of the first claim:

$$
\partial \mathsf{e}_{\mathsf{e}_f^\mu}^\lambda(\mathbf{w}) = \partial \mathsf{e}_f^{\lambda+\mu}(\mathbf{w}).
$$

Indeed, the second result in item iii) implies that there exists some $\mathbf{z} \in \partial \mathsf{e}_{\mathsf{e}_f^\mu}^\lambda(\mathbf{w}) = \partial \mathsf{e}_f^{\lambda+\mu}(\mathbf{w})$ such that $\mathbf{z} \in \frac{1}{\lambda}(\mathbf{w} - \mathsf{P}_{\mathsf{e}_f^\mu}^\lambda(\mathbf{w}))$ and $\mathbf{z} \in \frac{1}{\lambda+\mu}(\mathbf{w} - \mathsf{P}_f^{\lambda+\mu}(\mathbf{w}))$. Rearranging we obtain the claim.

For the last item iv), simply note that

$$
\mu \mathsf{e}_f^\mu(\mathbf{w}) = \min_{\mathbf{z}} \tfrac{1}{2} \|\mathbf{z} - \mathbf{w}\|^2 + \mu f(\mathbf{z}) = \tfrac{1}{2} \|\mathbf{w}\|^2 - \sup_{\mathbf{z}} \Big\{ \langle \mathbf{w}, \mathbf{z} \rangle - \big[\mu f(\mathbf{z}) + \tfrac{1}{2} \|\mathbf{z}\|^2\big] \Big\},
$$

and resort to the definition of the Fenchel conjugate. ■

The results resemble Proposition 1 of [14], with some equalities replaced by subset containments (which is necessary as demonstrated in Example 3). The last property in ii) shows in particular that any global minimizer of f is a fixed point of the proximal map, hinting that the proximal gradient algorithm might still work in the nonconvex setting. Item iv) reassures that we can still treat the proximal map P_f^μ (up to a small change) as the derivative of the Moreau envelope e_f^μ, except perhaps on a null set. The last item vii) is a more general form of Moreau's identity, from which the continuity of e_f^μ is apparent.

B Proof of Proposition 1

Proof: We prove the first part, which will imply that $\mathsf{e}^\mu : \mathsf{CPB} \to \mathsf{SCV}_\mu$ is onto.

$\Rightarrow$: This is already mentioned in the last item in Proposition 7.

$\Leftarrow$: Let $h = \frac{1}{2\mu} \|\cdot\|^2 - f$ be convex and finite valued. Then the function $j(\mathbf{w}) := h^*(\mu^{-1}\mathbf{w}) - \frac{1}{2\mu} \|\mathbf{w}\|^2$ is clearly closed. Moreover,

$$
\begin{aligned}
\mathsf{e}_j^\mu(\mathbf{w}) &= \inf_{\mathbf{z}} \tfrac{1}{2\mu} \|\mathbf{w} - \mathbf{z}\|^2 + h^*(\mu^{-1}\mathbf{z}) - \tfrac{1}{2\mu} \|\mathbf{z}\|^2 \\
&= \tfrac{1}{2\mu} \|\mathbf{w}\|^2 - \sup_{\mathbf{z}} \big[\langle \mathbf{w}, \mu^{-1}\mathbf{z} \rangle - h^*(\mu^{-1}\mathbf{z}) \big] \\
&= \tfrac{1}{2\mu} \|\mathbf{w}\|^2 - h(\mathbf{w}) = f(\mathbf{w}),
\end{aligned}
$$

due to the convexity of h.

The rest of the proof follows that of [14]. It is clear that $\mathsf{e}^\mu : \mathsf{CPB} \to \mathsf{SCV}_\mu$ is increasing w.r.t. the pointwise order, i.e., $f \geq g \implies \mathsf{e}_f^\mu \geq \mathsf{e}_g^\mu$.

Let $\alpha \in]0,1[$, then

$$\begin{aligned}
\mathsf{e}^{\mu}_{\alpha f+(1-\alpha)g}(\mathbf{w}) &= \inf_{\mathbf{z}} \tfrac{1}{2\mu}\|\mathbf{w}-\mathbf{z}\|^2 + \alpha f(\mathbf{z}) + (1-\alpha)g(\mathbf{z}) \\
&= \inf_{\mathbf{z}} \tfrac{\alpha}{2\mu}\|\mathbf{w}-\mathbf{z}\|^2 + \alpha f(\mathbf{z}) + \tfrac{1-\alpha}{2\mu}\|\mathbf{w}-\mathbf{z}\|^2 + (1-\alpha)g(\mathbf{z}) \\
&\geq \inf_{\mathbf{z}} \tfrac{\alpha}{2\mu}\|\mathbf{w}-\mathbf{z}\|^2 + \alpha f(\mathbf{z}) + \inf_{\mathbf{z}} \tfrac{1-\alpha}{2\mu}\|\mathbf{w}-\mathbf{z}\|^2 + (1-\alpha)g(\mathbf{z}) \\
&= \alpha \mathsf{e}^{\mu}_{f}(\mathbf{w}) + (1-\alpha)\mathsf{e}^{\mu}_{g}(\mathbf{w}),
\end{aligned}$$

verifying the concavity of e^{μ}. ■

C Discussion on different notions of the proximal maps

We document in this section some new results about the different notions of proximal maps defined in the main paper.

Recall that a multi-valued map $\mathsf{P} : \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ is upper semicontinuous and compact valued (usco) iff for all $\mathbf{w} \in \mathbb{R}^p$, $\mathsf{P}(\mathbf{w})$ is nonempty and for any sequence $(\mathbf{w}_n, \mathbf{z}_n)$ with $\mathbf{z}_n \in \mathsf{P}(\mathbf{w}_n)$, $\mathbf{w}_n \to \mathbf{w}$, the sequence $\mathbf{z}_n$ has a cluster point in $\mathsf{P}(\mathbf{w})$. Note that for a usco map P, $\mathsf{P}(\mathbf{w})$ is compact for each $\mathbf{w}$.

Fix $\mu > 0$ and $f \in \mathsf{CPB}$, recall the definition of the map $\hat{\mathsf{P}}^{\mu}_{f}$:

$$\hat{\mathsf{P}}^{\mu}_{f}(\mathbf{w}) = \begin{cases} \mathsf{P}^{\mu}_{f}(\mathbf{w}), & \text{if } \mathsf{P}^{\mu}_{f}(\mathbf{w}) \text{ is single-valued} \\ \emptyset, & \text{otherwise} \end{cases}.$$

We record the following result for convenience:

Lemma 1 $\mathsf{P}^{\mu}_{f}(\mathbf{w})$ *is a singleton iff* e^{μ}_{f} *is differentiable at* $\mathbf{w}$.

Proof: Suppose first that $\mathsf{P}^{\mu}_{f}(\mathbf{w})$ is a singleton, then according to [26, Example 10.32] both $\partial(-\mathsf{e}^{\mu}_{f})(\mathbf{w})$ and $\partial(\mathsf{e}^{\mu}_{f})(\mathbf{w})$ are singletons. Applying [26, Theorem 9.18] we know e^{μ}_{f} is differentiable. Conversely, if e^{μ}_{f} is differentiable at $\mathbf{w}$, so is $-\mathsf{e}^{\mu}_{f}$. Applying again [26, Example 10.32] together with [26, Exercise 8.8] and [26, Corollary 8.11] we know $\mathsf{P}^{\mu}_{f}(\mathbf{w})$ is a singleton. ■

Thus $\hat{\mathsf{P}}^{\mu}_{f}$ is almost everywhere defined. Our next goal is to extend its domain into the whole space by semicontinuity. Succinctly, we take the closure of its graph and obtain the limiting proximal map L^{μ}_{f} (see Definition 2).

Lemma 2 L^{μ}_{f} *is the minimal (in the sense of graph inclusion) usco map that extends* $\hat{\mathsf{P}}^{\mu}_{f}$.

Proof: Clearly for each $\mathbf{w}$, $\mathsf{L}^{\mu}_{f}(\mathbf{w}) \supseteq \hat{\mathsf{P}}^{\mu}_{f}(\mathbf{w})$. Thanks to item iv) of Proposition 7 and [26, Theorem 1.25], we know $\emptyset \neq \mathsf{L}^{\mu}_{f}(\mathbf{w}) \subseteq \mathsf{P}^{\mu}_{f}(\mathbf{w})$ for all $\mathbf{w}$. In order to prove L^{μ}_{f} is usco, take any sequence $(\mathbf{w}_n, \mathbf{z}_n)$ with $\mathbf{z}_n \in \mathsf{L}^{\mu}_{f}(\mathbf{w}_n)$ and $\mathbf{w}_n \xrightarrow{n\to\infty} \mathbf{w}$, and we want to find a cluster point $\mathbf{z}$ of $\{\mathbf{z}_n\}$ that is in $\mathsf{L}^{\mu}_{f}(\mathbf{w})$. Using the definition of L^{μ}_{f} we know there exists $(\mathbf{w}_{n,m}, \mathbf{z}_{n,m}) \xrightarrow{m\to\infty} (\mathbf{w}_n, \mathbf{z}_n), \mathbf{z}_{n,m} = \hat{\mathsf{P}}^{\mu}_{f}(\mathbf{w}_{n,m})$. Since P^{μ}_{f} is locally bounded [26, Theorem 1.25] and $\mathbf{w}_{n,m} \xrightarrow{m\to\infty} \mathbf{w}_n \xrightarrow{n\to\infty} \mathbf{w}$, w.l.o.g. we can assume $\{\mathbf{w}_{n,m}\}$ hence also $\{\mathbf{z}_{n,m}\}$ are bounded. For each n, we choose some m_n such that $|\mathbf{w}_{n,m_n} - \mathbf{w}_n| \leq \frac{1}{n}$ and $|\mathbf{z}_{n,m_n} - \mathbf{z}_n| \leq \frac{1}{n}$. Passing to a subsequence if necessary we can assume the sequence $\{\mathbf{z}_{n,m_n}\}$ is convergent. Clearly $\mathbf{z}_n$ converges to the same limit say $\mathbf{z}$. Thus we have constructed the sequence $(\mathbf{w}_{n,m_n}, \mathbf{z}_{n,m_n})$ with $\mathbf{z}_{n,m_n} = \hat{\mathsf{P}}^{\mu}_{f}(\mathbf{w}_{n,m_n})$ and $\mathbf{w}_{n,m_n} \xrightarrow{n\to\infty} \mathbf{w}$, therefore the limit $\mathbf{z} \in \mathsf{L}^{\mu}_{f}(\mathbf{w})$, completing the proof for the usco of L^{μ}_{f}.

It is clear from the definition that any usco map that extends $\hat{\mathsf{P}}^{\mu}_{f}$ must also extend L^{μ}_{f}, implying the minimality of the latter. ■

Later on we will need some beautiful results about monotone maps hence we recall some definitions here. A multi-valued map $\mathsf{P} : \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ is monotone if for all $\mathbf{w}, \mathbf{z}$ and $\mathbf{u} \in \mathsf{P}(\mathbf{w}), \mathbf{v} \in \mathsf{P}(\mathbf{z})$ we have $\langle \mathbf{w} - \mathbf{z}, \mathbf{u} - \mathbf{v} \rangle \geq 0$.

Monotone maps are generalizations of monotone functions. A maximal monotone map is a monotone map whose graph is not properly included in any other monotone map. According to [26, Proposition 12.19], any proximal map P_f^μ is monotone. Clearly, the restriction $\hat{\mathsf{P}}_f^\mu$ is monotone as well, so is its "closure":

Lemma 3 L_f^μ *is monotone.*

Proof: For any $\mathbf{w}, \mathbf{z}, \mathbf{u} \in \mathsf{L}_f^\mu(\mathbf{w}), \mathbf{v} \in \mathsf{L}_f^\mu(\mathbf{z})$, we find $(\mathbf{w}_n, \mathbf{u}_n) \to (\mathbf{w}, \mathbf{u})$ with $\mathbf{u}_n \in \hat{\mathsf{P}}_f^\mu(\mathbf{w}_n)$ and similarly $(\mathbf{z}_m, \mathbf{v}_m) \to (\mathbf{z}, \mathbf{v})$ with $\mathbf{v}_m \in \hat{\mathsf{P}}_f^\mu(\mathbf{z}_m)$. Thus $0 \leq \langle \mathbf{w}_n - \mathbf{z}_m, \mathbf{u}_n - \mathbf{v}_m \rangle \to \langle \mathbf{w} - \mathbf{z}, \mathbf{u} - \mathbf{v} \rangle$. ■

Lemma 4 $\hat{\partial}(\mathsf{e}_f^\mu)(\mathbf{w}) = \mu^{-1}(\mathbf{w} - \hat{\mathsf{P}}_f^\mu(\mathbf{w})), \partial(\mathsf{e}_f^\mu)(\mathbf{w}) = \mu^{-1}(\mathbf{w} - \mathsf{L}_f^\mu(\mathbf{w}))$.

Proof: Simply combine Lemma 1, [26, Corollary 9.21], and [26, Example 10.32]. ■

The μ-proximal hull of $f \in \mathsf{CPB}$ is defined as $\mathsf{h}_f^\mu := -\mathsf{e}^\mu_{(-\mathsf{e}_f^\mu)}$. As shown in [26, Exercise 1.45], $\mathsf{e}_f^\mu = \mathsf{e}_g^\mu \iff \mathsf{h}_f^\mu = \mathsf{h}_g^\mu$. Moreover, f agrees with h_f^μ on the (closure of the) range of its proximal map $\mathrm{Im}(\mathsf{P}_f^\mu)$. It turns out that the proximal map of the proxmal hull is simply the convex hull:

Lemma 5 $\mathsf{H}_f^\mu = \mathsf{P}^\mu_{\mathsf{h}_f^\mu}$.

Proof: Since P_f^μ is usco and monotone [26, Proposition 12.19], its "convex hull" H_f^μ is also monotone and usco [32, Lemma 7.12]. By [32, Lemma 7.7] we know H_f^μ is maximal monotone. On the other hand, combining [26, Example 11.26] and [26, Proposition 12.19] we also know $\mathsf{P}^\mu_{\mathsf{h}_f^\mu}$ is maximal monotone. Since $\mathsf{e}_f^\mu = \mathsf{e}^\mu_{\mathsf{h}_f^\mu}$ [26, Example 1.44] and f agrees with h_f^μ on $\mathrm{Im}(\mathsf{P}_f^\mu)$, it easily follows that $\mathsf{P}^\mu_{\mathsf{h}_f^\mu}(\mathbf{w}) \supseteq \mathsf{P}_f^\mu(\mathbf{w})$ for all $\mathbf{w}$. By construction $\mathsf{H}_f^\mu(\mathbf{w}) \supseteq \mathsf{P}_f^\mu(\mathbf{w})$ for all $\mathbf{w}$. Therefore we have two maximal monotone maps $\mathsf{P}^\mu_{\mathsf{h}_f^\mu}$ and H_f^μ both extending the monotone map P_f^μ. Applying [32, Theorem 7.13] completes our proof. ■

Note that a similar argument around maximal monotonicity reveals that $\mathrm{conv}(\mathsf{L}_f^\mu(\mathbf{w})) = \mathsf{H}_f^\mu(\mathbf{w})$ for all $\mathbf{w}$.

We can now start to characterize when two functions have the same Moreau envelope.

Lemma 6 *Fix any $\mu > 0$ and $f, g \in \mathsf{CPB}$. Then the following are equivalent:*

(i). $\mathsf{e}_g^\mu = \mathsf{e}_f^\mu + c$ *for some constant c;*

(ii). For all $\mathbf{w}$, $\mathsf{P}_g^\mu(\mathbf{w}) \cap \mathsf{P}_f^\mu(\mathbf{w}) \neq \emptyset$;

(iii). $\mathsf{L}_g^\mu = \mathsf{L}_f^\mu$;

(iv). $\mathsf{H}_g^\mu = \mathsf{H}_f^\mu$.

Proof: (i) $\implies$ (ii): Clearly $\partial \mathsf{e}_f^\mu(\mathbf{w}) = \partial \mathsf{e}_g^\mu(\mathbf{w})$ for all $\mathbf{w}$. For any $\mathbf{w}$, according to item iii) of Proposition 7, we know there exists some $\mathbf{z} \in \mathsf{P}_f^\mu(\mathbf{w})$ such that $\frac{1}{\mu}(\mathbf{w} - \mathbf{z}) \in \partial \mathsf{e}_f^\mu(\mathbf{w})$. Similarly, there exists some $\mathbf{u} \in \mathsf{P}_g^\mu(\mathbf{w})$ such that $\frac{1}{\mu}(\mathbf{w} - \mathbf{z}) = \frac{1}{\mu}(\mathbf{w} - \mathbf{u})$, namely that $\mathbf{z} = \mathbf{u}$. Therefore $\mathsf{P}_f^\mu(\mathbf{w}) \bigcap \mathsf{P}_g^\mu(\mathbf{w}) \neq \emptyset$ for all $\mathbf{w}$.

(ii) $\implies$ (i): Observe that for any $h \in \mathsf{CPB}$ the Moreau envelope e_h^μ, being a difference of two finite-valued convex functions (see e.g. item vii) of Proposition 7), is locally Lipschitz continuous. Thanks to Rademacher's theorem, we thus know e_h^μ is differentiable up to a (Lebesgue) null set. For any $\mathbf{w}$, consider its (open) neighborhood $N_{\mathbf{w}}$ such that the restrictions of both e_f^μ and e_g^μ are Lipschitz continuous. Clearly $\mathsf{e}_g^\mu - \mathsf{e}_f^\mu$ is also differentiable on $N_{\mathbf{w}}$ up to a null set. On the other hand, according to Lemma 1, if e_h^μ is differentiable at $\mathbf{w}$, then $\mathsf{P}_h^\mu(\mathbf{w})$ is a singleton and $\nabla \mathsf{e}_h^\mu(\mathbf{w}) = \frac{1}{\mu}(\mathbf{w} - \mathsf{P}_h^\mu(\mathbf{w}))$. Thus on $N_{\mathbf{w}}$, up to a null set, the derivative of $\mathsf{e}_g^\mu - \mathsf{e}_f^\mu$ not only exists but also vanishes, due to the assumption $\mathsf{P}_g^\mu(\mathbf{w}) \bigcap \mathsf{P}_f^\mu(\mathbf{w}) \neq \emptyset$ for all $\mathbf{w}$. Since a Lipschitz continuous function is absolutely continuous, using the mean integral theorem we know $\mathsf{e}_g^\mu - \mathsf{e}_f^\mu = c_{\mathbf{w}}$ on the neighborhood $N_{\mathbf{w}}$ for some constant $c_{\mathbf{w}}$. Hence we have proved that the continuous function $\mathsf{e}_g^\mu - \mathsf{e}_f^\mu$ is locally constant on the connected set $\mathbb{R}^p$. A simple topological argument shows that $\mathsf{e}_g^\mu - \mathsf{e}_f^\mu$ must be a constant on all of $\mathbb{R}^p$.

(iii) $\implies$ (ii): Clear, since $\mathsf{P}_f^\mu(\mathbf{w}) \supseteq \mathsf{L}_f^\mu(\mathbf{w}) \neq \emptyset$ for all $\mathbf{w}$ and $f \in \mathsf{CPB}$.

(i) $\implies$ (iii): Apply Lemma 4.

(i) $\implies$ (iv): $\mathsf{e}_g^\mu = \mathsf{e}_f^\mu + c$ implies $\mathsf{h}_g^\mu = \mathsf{h}_f^\mu + c$ hence $\mathsf{H}_g^\mu = \mathsf{H}_f^\mu$, thanks to Lemma 5.

(iv) $\implies$ (iii): Using Lemma 5 we have $\mathsf{P}_{\mathsf{h}_g^\mu}^\mu = \mathsf{P}_{\mathsf{h}_f^\mu}^\mu$ hence $\mathsf{L}_{\mathsf{h}_g^\mu}^\mu = \mathsf{L}_{\mathsf{h}_f^\mu}^\mu$. But the already established equivalence (i) $\iff$ (iii) implies $\mathsf{L}_{\mathsf{h}_f^\mu}^\mu = \mathsf{L}_f^\mu$ for any $f \in \mathsf{CPB}$. Thus $\mathsf{L}_g^\mu = \mathsf{L}_f^\mu$. ■

Lemma 7 *Let P_f^μ be the proximal map of some function $f \in \mathsf{CPB}$. Then we can explicitly construct the function*

$$g(\mathbf{w}) = \begin{cases} \mathsf{h}_f^\mu(\mathbf{w}), & \mathbf{w} \in \overline{\mathrm{Im}(\mathsf{P}_f^\mu)} \\ a(\mathbf{w}), & \textit{otherwise} \end{cases} \tag{11}$$

with any $a(\mathbf{w}) \geq \mathsf{h}_f^\mu(\mathbf{w}) + \epsilon_{\mathbf{w}}$ for some $\epsilon_{\mathbf{w}} > 0$, such that $\mathsf{P}_g^\mu = \mathsf{P}_f^\mu$.

Proof: By definition $\mathsf{h}_f^\mu \leq g \leq \ell_f^\mu$, see (15). Applying Proposition 2 we have $\mathsf{e}_g^\mu = \mathsf{e}_f^\mu$. Let $\mathbf{z} \in \mathsf{P}_f^\mu(\mathbf{w})$, then

$$\begin{aligned} \mathsf{e}_f^\mu(\mathbf{w}) &= \tfrac{1}{2\mu}\left\|\mathbf{z} - \mathbf{w}\right\|^2 + f(\mathbf{z}) \\ &= \tfrac{1}{2\mu}\left\|\mathbf{z} - \mathbf{w}\right\|^2 + \mathsf{h}_f^\mu(\mathbf{z}) \\ &= \tfrac{1}{2\mu}\left\|\mathbf{z} - \mathbf{w}\right\|^2 + g(\mathbf{z}) \\ &= \mathsf{e}_g^\mu(\mathbf{w}), \end{aligned}$$

implying $\mathbf{z} \in \mathsf{P}_g^\mu(\mathbf{w})$. Similarly, any $\mathbf{z} \in \mathsf{P}_g^\mu(\mathbf{w}) \cap \overline{\mathrm{Im}(\mathsf{P}_f^\mu)}$ must be in $\mathsf{P}_f^\mu(\mathbf{w})$ as well. If there exists $\mathbf{z} \in \mathsf{P}_g^\mu(\mathbf{w}) \backslash \overline{\mathrm{Im}(\mathsf{P}_f^\mu)}$, then

$$\mathsf{e}_g^\mu(\mathbf{w}) = \tfrac{1}{2\mu}\left\|\mathbf{z} - \mathbf{w}\right\|^2 + a(\mathbf{z}) > \tfrac{1}{2\mu}\left\|\mathbf{z} - \mathbf{w}\right\|^2 + \mathsf{h}_f^\mu(\mathbf{z}) \geq \mathsf{e}_{\mathsf{h}_f^\mu}^\mu(\mathbf{w}) = \mathsf{e}_f^\mu(\mathbf{w}),$$

contradiction. Therefore $\mathsf{P}_f^\mu = \mathsf{P}_g^\mu$.

The explicit construction of g relies on that of the proximal hull h_f^μ. We first take the convex hull of P at each point. This recovers $\mathsf{H}_f^\mu = \mathsf{P}_{\mathsf{h}_f^\mu}^\mu$. Next we recover the proximal hull h_f^μ. It follows from [26] that $\mathsf{h}_f^\mu + \frac{1}{2\mu}\|\cdot\|^2$ is convex, thus

$$\mathsf{e}_{\mathsf{h}_f^\mu}^\mu(\mathbf{w}) = \min_{\mathbf{z}} \tfrac{1}{2\mu}\left\|\mathbf{w} - \mathbf{z}\right\|^2 + \mathsf{h}_f^\mu(\mathbf{z}) \tag{12}$$

$$= \tfrac{1}{2\mu}\|\mathbf{w}\|^2 - \sup_{\mathbf{z}} \langle \mathbf{w}/\mu, \mathbf{z}\rangle - (\mathsf{h}_f^\mu(\mathbf{z}) + \tfrac{1}{2\mu}\|\mathbf{z}\|^2) \tag{13}$$

Thanks to convexity, we have $\mathsf{H}_f^\mu(\mathbf{w}) = \partial(\mathsf{h}_f^\mu + \frac{1}{2\mu}\|\cdot\|^2)^*(\mathbf{w}/\mu)$. Integrating H_f^μ along rays we can recover $(\mathsf{h}_f^\mu + \frac{1}{2\mu}\|\cdot\|^2)^*$. Lastly, taking Fenchel conjugate we have h_f^μ hence g explicitly. ■

Note that the function g in (11) may not be closed, hence can be inconvenient. However, if we choose $a(\mathbf{w}) - \mathsf{h}_f^\mu(\mathbf{w}) \geq \epsilon > 0$ (i.e., ϵ is independent of $\mathbf{w}$), then closedness can be achieved, without harming the conclusion, by taking the lower semicontinuous hull.

Corollary 1 $\mathrm{Im}(\mathsf{P}_f^\mu) = \mathrm{Im}(\mathsf{H}_f^\mu) \iff \overline{\mathrm{Im}(\mathsf{P}_f^\mu)} = \overline{\mathrm{Im}(\mathsf{H}_f^\mu)} \iff \mathsf{P}_f^\mu = \mathsf{H}_f^\mu$.

Proof: Apply Lemma 5 and Lemma 7. ■

On the real line we can completely characterize the proximal map.

Lemma 8 *If* $\mathsf{P} : \mathbb{R} \rightrightarrows \mathbb{R}$ *is maximal monotone, then it is a proximal map.*

Proof: It follows from [26, Exercise 12.26, Theorem 12.25] that P is the subdifferential of some convex function f. Let $\mathsf{h} = f^* - \frac{1}{2}\|\cdot\|^2$. We claim that $\mathsf{P}_{\mathsf{h}} = \mathsf{P}$. Indeed,

$$\mathsf{e}_{\mathsf{h}}(w) = \min_z \tfrac{1}{2}\|w - z\|^2 + \mathsf{h}(z) = \tfrac{1}{2}\|w\|^2 - \sup_z \langle w, z\rangle - f^*(z).$$

Thus $\mathsf{P}_{\mathsf{h}}(w) = \partial f(w) = \mathsf{P}(w)$, thanks to the convexity of f. ■

Proposition 3 $\mathsf{P} : \mathbb{R} \rightrightarrows \mathbb{R}$ *is a proximal map iff it is (nonempty) compact-valued, monotone, and has a closed graph. Moreover,* P *is maximal monotone iff there is a unique function (up to addition of a constant)* f *such that* $\mathsf{P}_f = \mathsf{P}$ *iff* P *is also convex-valued.*

Proof: If P is a proximal map (of some function f), then it is clearly compact-valued, monotone, and has a closed graph, see [26]. Conversely, let $\mathsf{P} : \mathbb{R} \rightrightarrows \mathbb{R}$ be compact-valued, monotone and have closed graph, then its (pointwise) convex hull H is maximal monotone [32, Lemma 7.12, Lemma 7.7]. Applying Lemma 8 we know there exists a function $\mathsf{h} \in \mathsf{CPB}$ such that $\mathsf{P}_{\mathsf{h}} = \mathsf{H}$. We construct the closed function

$$g(w) = \begin{cases} \mathsf{h}(w), & w \in \overline{\mathrm{Im}(\mathsf{P})} \\ \infty, & \text{otherwise} \end{cases}. \tag{14}$$

According to Lemma 7, P is a proximal map iff $\mathsf{P}_g = \mathsf{P}$.

Indeed, let $q \in \mathsf{P}(w) \subseteq \mathsf{P}_{\mathsf{h}}(w)$. Then

$$\mathsf{e}_{\mathsf{h}}(w) = \tfrac{1}{2}(q - w)^2 + \mathsf{h}(q) = \tfrac{1}{2}(q - w)^2 + g(q) \geq \mathsf{e}_g(w) \geq \mathsf{e}_{\mathsf{h}}(w),$$

implying $q \in \mathsf{P}_g(w)$. Therefore $\mathsf{P}_g(w) \supseteq \mathsf{P}(w)$ for all w. For the converse, first let $q \in \mathsf{P}_g(w) \cap \mathrm{Im}(\mathsf{P})$, thus we know $q \in \mathsf{P}(z) \subseteq \mathsf{P}_g(z)$ for some $z \in \mathbb{R}$. If $z \neq w$, then due to monotonicity, we must have $q \in \mathsf{P}_g(u)$ for any $(w \wedge z) \leq u \leq (w \vee z)$. Note that P and P_g agree almost everywhere (since both are single-valued almost everywhere). Using the closedness of the graph of P we know $q \in \mathsf{P}(w)$. Therefore $\mathsf{P}(w) \supseteq \mathsf{P}_g(w) \cap \mathrm{Im}(\mathsf{P})$ for all w. To complete the proof we need to remove the intersection with $\mathrm{Im}(\mathsf{P})$. Let $s = \sup\{\mathrm{Im}(\mathsf{P})\}$ and $i = \inf\{\mathrm{Im}(\mathsf{P})\}$. Let $q \in \mathsf{P}_g(w) \backslash \mathrm{Im}(\mathsf{P})$ (there is nothing to prove if there does not exist such q). We claim that it is impossible to have $i < q < s$. Suppose not, then q is sandwiched in the bounded interval $]a, b[$ with some $a \in \mathsf{P}(u), b \in \mathsf{P}(v)$. By our definition of g in (14), $q \in \overline{\mathrm{Im}(\mathsf{P})}$, thus there exists $\mathsf{P}(w_n) \ni q_n \to q$. W.l.o.g. we can assume $a < q_n < b$. Due to monotonicity of P, we must have $u \leq w_n \leq v$. Therefore we can find a subsequence of $\{w_n\}$ that converges to, say z. Since P has a closed graph, we must have $q \in \mathsf{P}(z) \subseteq \mathrm{Im}(\mathsf{P})$, contradiction. We are left with $q = s$ or $q = i$. Assume the former, which implies $s < \infty$. For any $z \geq w$, using monotonicity and maximality we must have $s \in \mathsf{P}_g(z)$. Since P_g agree with P almost everywhere, we must have again $q \in \mathrm{Im}(\mathsf{P})$. The case $q = i$ is dealt with similarly. In summary, we have proved that actually $\mathsf{P}_g(w) \subseteq \mathrm{Im}(\mathsf{P})$ for all w, hence completes the proof for $\mathsf{P} = \mathsf{P}_g$.

Turning to the second claim, we first note that the maximal monotonicity of P clearly implies its convex-valuedness. The converse follows from [32, Lemma 7.7].

If P is not convex valued at some point w, then the range of P must have a gap around $\mathsf{P}(w)$. We construct the function g in (11) with different $a(w)$. They all have the same proximal map but differ more than just a (global) constant. Conversely, if $\mathsf{P} = \mathsf{P}_f = \mathsf{P}_g$ is maximal monotone, then $\mathsf{Id} + \partial f = \mathsf{Id} + \partial g$ and the functions $f + \frac{1}{2}\|\cdot\|^2$ and $g + \frac{1}{2}\|\cdot\|^2$ are convex [26, Proposition 12.19]. Since convex functions are determined by their subdifferentiable (up to a constant), we have $f = g + c$ for some constant c. ■

Proposition 3 provides guidance on designing different thresholding rules while Lemma 7 enables the construction of the corresponding regularization function. Together they consist of a significant generalization of [33, Proposition 3.2].

Corollary 2 *If* $\mathsf{P} : \mathbb{R} \to \mathbb{R}$ *is increasing and continuous, then there is a unique function* f *(up to the addition of a constant) such that* $\mathsf{P}_f = \mathsf{P}$.

Proof: Simply note that any continuous monotone map is maximal monotone. ■

Thus, both the SCAD [2] and the MC+ [3] thresholding rules correspond to a unique regularization function. In contrast, there are infinitely many different regularizers that all lead to the hard thresholding rule, see Example 3. Note that, unlike the convex case, the proximal map in general need not be non-expansive.

D Proof of Proposition 2

Proof: We define the functions $\mathsf{h}_f^\mu := -\mathsf{e}^\mu_{(-\mathsf{e}_f^\mu)}$ (namely the μ-proximal hull of f) and

$$\ell_f^\mu(\mathbf{w}) = \begin{cases} \mathsf{h}_f^\mu(\mathbf{w}), & \mathbf{w} \in \overline{\mathrm{Im}(\mathsf{L}_f^\mu)} \\ \infty, & \text{otherwise} \end{cases}. \tag{15}$$

Due to the closure operation on $\mathrm{Im}(\mathsf{L}_f^\mu)$, ℓ_f^μ is closed. We note that $f - \mathsf{h}_f^\mu \geq 0$ with equality on $\mathrm{Im}(\mathsf{P}_f^\mu)$, hence also on $\overline{\mathrm{Im}(\mathsf{P}_f^\mu)}$ due to lower semicontinuity.

First assume that $\mathsf{e}_g^\mu = \mathsf{e}_f^\mu + c$ and we prove that $\mathsf{h}_f^\mu \leq g - c \leq \ell_f^\mu$. As shown in [26], $f \geq \mathsf{h}_f^\mu$ for all $f \in \mathsf{CPB}$. Thus $g \geq \mathsf{h}_g^\mu = -\mathsf{e}^\mu_{(-\mathsf{e}_g^\mu)} = -\mathsf{e}^\mu_{(-\mathsf{e}_f^\mu - c)} = -\mathsf{e}^\mu_{(-\mathsf{e}_f^\mu)} + c = \mathsf{h}_f^\mu + c$, which is the first inequality to be proved. For the second inequality, Lemma 6 in Appendix C shows that $\mathsf{L}_f^\mu = \mathsf{L}_g^\mu$, thus g agrees with $\mathsf{h}_g^\mu = \mathsf{h}_f^\mu + c$ on $\overline{\mathrm{Im}(\mathsf{P}_g^\mu)} \supseteq \overline{\mathrm{Im}(\mathsf{L}_g^\mu)} = \overline{\mathrm{Im}(\mathsf{L}_f^\mu)}$. Thus $g - c \leq \ell_f^\mu$.

Next we prove that $\mathsf{e}^\mu_{\ell_f^\mu} = \mathsf{e}_f^\mu$, which will certify the converse for the first two claims. By definition, for all $\mathbf{w}$

$$\begin{aligned} \mathsf{e}^\mu_{\ell_f^\mu}(\mathbf{w}) &= \min_{\mathbf{z} \in \overline{\mathrm{Im}(\mathsf{L}_f^\mu)}} \tfrac{1}{2\mu} \|\mathbf{w} - \mathbf{z}\|^2 + \mathsf{h}_f^\mu(\mathbf{z}) \\ &= \min_{\mathbf{z} \in \overline{\mathrm{Im}(\mathsf{L}_f^\mu)}} \tfrac{1}{2\mu} \|\mathbf{w} - \mathbf{z}\|^2 + f(\mathbf{z}) \\ &\leq \mathsf{e}_f^\mu(\mathbf{w}), \end{aligned}$$

since f agrees with h_f^μ on $\overline{\mathrm{Im}(\mathsf{P}_f^\mu)} \supseteq \overline{\mathrm{Im}(\mathsf{L}_f^\mu)}$ and $\mathsf{L}_f^\mu(\mathbf{w}) \neq \emptyset$. On the other hand, we clearly have $\ell_f^\mu \geq \mathsf{h}_f^\mu$ hence $\mathsf{e}^\mu_{\ell_f^\mu} \geq \mathsf{e}^\mu_{\mathsf{h}_f^\mu} = \mathsf{e}_f^\mu$, completing the proof for $\mathsf{e}^\mu_{\ell_f^\mu} = \mathsf{e}_f^\mu$.

From the equality $\mathsf{e}^\mu_{\ell_f^\mu} = \mathsf{e}_f^\mu$ follows $\mathsf{L}_f^\mu(\mathbf{w}) \subseteq \mathsf{P}^\mu_{\ell_f^\mu}(\mathbf{w})$ for all $\mathbf{w}$. It is then clear from Lemma 6 that $\mathsf{P}^\mu_{\ell_f^\mu}(\mathbf{w}) \subseteq \mathsf{P}_g^\mu(\mathbf{w}) \subseteq \mathsf{H}_f^\mu(\mathbf{w})$ for all $\mathbf{w}$ implies $\mathsf{e}_g^\mu = \mathsf{e}_f^\mu + c$. We prove its converse. Clearly, we have $\mathsf{P}_g^\mu(\mathbf{w}) \subseteq \mathsf{H}_g^\mu(\mathbf{w}) = \mathsf{H}_f^\mu(\mathbf{w})$ for all $\mathbf{w}$, thanks again to Lemma 6. Note that for any $\mathbf{z} \in \overline{\mathrm{Im}(\mathsf{L}_f^\mu)}$ we have from $\mathsf{e}_g^\mu = \mathsf{e}_f^\mu + c$ that $\mathsf{h}_f^\mu(\mathbf{z}) = \mathsf{h}_g^\mu(\mathbf{z}) - c = g(\mathbf{z}) - c$, since g agrees with h_g^μ on $\overline{\mathrm{Im}(\mathsf{P}_g^\mu)} \supseteq \overline{\mathrm{Im}(\mathsf{L}_g^\mu)} = \overline{\mathrm{Im}(\mathsf{L}_f^\mu)}$. Now take any $\mathbf{z} \in \mathsf{P}^\mu_{\ell_f^\mu}(\mathbf{w})$. Clearly $\mathbf{z} \in \overline{\mathrm{Im}(\mathsf{L}_f^\mu)} = \overline{\mathrm{Im}(\mathsf{L}_g^\mu)}$ hence

$$\begin{aligned} \mathsf{e}^\mu_{\ell_f^\mu}(\mathbf{w}) &= \tfrac{1}{2\mu} \|\mathbf{w} - \mathbf{z}\|^2 + \mathsf{h}_f^\mu(\mathbf{z}) = \tfrac{1}{2\mu} \|\mathbf{w} - \mathbf{z}\|^2 + g(\mathbf{z}) - c \\ &= \min_{\mathbf{u} \in \overline{\mathrm{Im}(\mathsf{L}_f^\mu)}} \tfrac{1}{2\mu} \|\mathbf{w} - \mathbf{u}\|^2 + \mathsf{h}_f^\mu(\mathbf{u}) \\ &= \min_{\mathbf{u} \in \overline{\mathrm{Im}(\mathsf{L}_g^\mu)}} \tfrac{1}{2\mu} \|\mathbf{w} - \mathbf{u}\|^2 + g(\mathbf{u}) - c \\ &= \mathsf{e}_g^\mu(\mathbf{w}) - c. \end{aligned}$$

Thus $\mathbf{z} \in \mathsf{P}_g^\mu(\mathbf{w})$, proving $\mathsf{P}^\mu_{\ell_f^\mu}(\mathbf{w}) \subseteq \mathsf{P}_g^\mu(\mathbf{w})$ for all $\mathbf{w}$. ■

E Proof of Proposition 4

Proof: By Lemma 1, on the domain D of the map $\sum_k \alpha_k \hat{\mathsf{P}}_{f_k}$, $\sum_k \alpha_k \mathsf{e}^\mu_{f_k}$ is differentiable, hence any version of the proximal average A^μ must also be differentiable at points in D. Thus on D, $\mathsf{P}^\mu_{\mathsf{A}^\mu} = \sum_k \alpha_k \hat{\mathsf{P}}_{f_k}$. Since both

$\mathsf{P}^{\mu}_{\mathsf{A}^{\mu}}$ and $\sum_k \alpha_k \mathsf{P}^{\mu}_{f_k}$ have closed graphs, the closure $\overline{\sum_k \alpha_k \hat{\mathsf{P}}_{f_k}}$ is in their intersection. Note that $\sum_k \alpha_k \hat{\mathsf{P}}^{\mu}_{f_k}$ is almost everywhere defined, thus its closure is everywhere defined. ■

F Proof of Proposition 5

Proof: The first part of Proposition 5 is a standard exercise in semi-algebraic geometry. Recall that if the set $A \subseteq \mathbb{R}^p \times \mathbb{R}^d$ is definable, then its projection $\{\mathbf{w} \in \mathbb{R}^p : \exists \mathbf{z}, (\mathbf{w}, \mathbf{z}) \in A\}$ is definable too. In the semi-algebraic setting, this is the well-known Tarski-Seidenberg theorem, while it is a built-in property in general order-minimal structures [28]. It follows easily from the projection property that all (sub)level sets and (strict) epigraphs of a definable function is definable. For instance, the strict epigraph $\{(\mathbf{w}, t) \in \mathbb{R}^p \times \mathbb{R} : f(\mathbf{w}) < t\}$ is the projection of the set $\big[\{(\mathbf{w}, t, s) : f(\mathbf{w}) - t = s\}\big] \cap \big[(\mathbf{w}, t, s) : s < 0\big]$, which is definable since the function $h(\mathbf{w}, t) = f(\mathbf{w}) - t$ is definable hence having a definable graph.

To begin our proof, note first that if f is definable, the function $g(\mathbf{w}, \mu, \mathbf{z}) = f(\mathbf{z}) + \frac{1}{2\mu} \|\mathbf{z} - \mathbf{w}\|^2$ is definable in the product space $\mathbb{R}^p \times \mathbb{R}_{++} \times \mathbb{R}^p$, since the squared Euclidean norm is definable as well. (Recall that the squared Euclidean norm is semi-algebraic and we assume all semi-algebraic functions are definable.) Thus the strict epigraph of e^{μ}_f, $\{(\mathbf{w}, \mu, t) \in \mathbb{R}^p \times \mathbb{R}_{++} \times \mathbb{R} : \mathsf{e}^{\mu}_f(\mathbf{w}) < t\} = \{(\mathbf{w}, \mu, t) \in \mathbb{R}^p \times \mathbb{R}_{++} \times \mathbb{R} : \exists \mathbf{z}, g(\mathbf{w}, \mu, \mathbf{z}) < t\}$, as the projection of the strict epigraph of g, is definable. Similarly one can prove that the strict hypograph $\{(\mathbf{w}, t) \in \mathbb{R}^p \times \mathbb{R} : \mathsf{e}^{\mu}_f(\mathbf{w}) > t\}$ is definable too. So is thus the graph $\{(\mathbf{w}, \mu, t) \in \mathbb{R}^p \times \mathbb{R}_{++} \times \mathbb{R} : \mathsf{e}^{\mu}_f(\mathbf{w}) = t\}$ (since taking union or complement preserves definability). Hence $\mathsf{e}^{\mu}_f(\mathbf{w})$ is definable as a joint function of $(\mathbf{w}, \mu)$.

Conversely, let us assume e^{μ}_f is definable as a joint function of $(\mathbf{w}, \mu)$. We use the monotonic property of the envelope function, that is, $\mathsf{e}^{\mu}_f(\mathbf{w}) \uparrow f(\mathbf{w})$ for all $\mathbf{w}$ as $\mu \downarrow 0$. It follows that the strict hypograph of f, i.e. the set $\{(\mathbf{w}, t) : f(\mathbf{w}) > t\} = \{(\mathbf{w}, t) : \exists \mu > 0, \mathsf{e}^{\mu}_f(\mathbf{w}) > t\}$, is definable thanks to the projection property. Similarly the strict epigraph hence the graph of f is definable, namely f is definable.

Finally, thanks to our particular choice of the proximal average, we know from the previous result that $g(\mathbf{w}, \mu) = \sum_k \alpha_k \mathsf{e}^{\mu}_{f_k}(\mathbf{w})$ is definable as a joint function of $(\mathbf{w}, \mu)$. Using the previous result once again we know the proximal average $-\mathsf{e}^{\mu}_{(-g)}$ is definable as a joint function of $(\mathbf{w}, \mu)$. ■

G Proof of Proposition 6

Proof: It is clear that $\mathsf{e}^{\mu}_f \le f$ for all functions f. Thus we need only prove the first and the last inequalities.

The proof of the last inequality is the same as [14], which we reproduce for completeness. Observe that by the definition of the proximal average

$$\bar{f} - \mathsf{e}^{\mu}_{\mathsf{A}^{\mu}} = \sum_k \alpha_k (f_k - \mathsf{e}^{\mu}_{f_k}) \ge 0,$$

since $f \ge \mathsf{e}^{\mu}_f$ for any f. On the other hand

$$\begin{aligned}
\sup_{\mathbf{w}} \Big\{ f_k(\mathbf{w}) - \mathsf{e}^{\mu}_{f_k}(\mathbf{w}) \Big\} &= \sup_{\mathbf{w}} \Big\{ f_k(\mathbf{w}) - \Big[\inf_{\mathbf{z}} \tfrac{1}{2\mu} \|\mathbf{w} - \mathbf{z}\|^2 + f_k(\mathbf{z}) \Big] \Big\} \\
&= \sup_{\mathbf{w}, \mathbf{z}} \Big\{ f_k(\mathbf{w}) - f_k(\mathbf{z}) - \tfrac{1}{2\mu} \|\mathbf{w} - \mathbf{z}\|^2 \Big\} \\
&\le \sup_{\mathbf{w}, \mathbf{z}} \Big\{ M_k \|\mathbf{w} - \mathbf{z}\| - \tfrac{1}{2\mu} \|\mathbf{w} - \mathbf{z}\|^2 \Big\} \\
&\le \tfrac{\mu M_k^2}{2},
\end{aligned}$$

where the first inequality is due to the Lipschitz assumption on f_k. Therefore

$$\sup_{\mathbf{w}} \{ \bar{f}(\mathbf{w}) - \mathsf{e}^{\mu}_{\mathsf{A}^{\mu}}(\mathbf{w}) \} \le \sum_k \alpha_k \Big[\sup_{\mathbf{w}} f_k(\mathbf{w}) - \mathsf{e}^{\mu}_{f_k}(\mathbf{w}) \Big] \le \tfrac{\mu \overline{M^2}}{2}.$$

To prove $\mathsf{A}^\mu \leq \bar{f}$, we first use the concavity of the Moreau envelope map e^μ (cf. Proposition 1):

$$\mathsf{e}^\mu_{\bar{f}} = \mathsf{e}^\mu_{\left(\sum_k \alpha_k f_k\right)} \geq \sum_k \alpha_k \mathsf{e}^\mu_{f_k}.$$

Thanks to our choice of the proximal average, using the monotonicity of e^μ we have

$$\mathsf{A}^\mu = -\mathsf{e}^\mu_{\left(-\sum_k \alpha_k \mathsf{e}^\mu_{f_k}\right)} \leq -\mathsf{e}^\mu_{\left(-\mathsf{e}^\mu_{\bar{f}}\right)} \leq \bar{f},$$

where the last inequality is due to [26, Example 1.44]. ■

H Proof of Theorem 1

Proof: It follows from Proposition 4 that Algorithm 1 is the usual proximal gradient (a.k.a. forward-backward splitting) algorithm, applied to solve the approximate surrogate problem (9). Thanks to Proposition 5 and our assumption on the definability of ℓ and $\{f_k\}$, we know the objective in (9) is definable. All assumptions of [15, Proposition 3, p. 484] are met, hence follows our claim in Theorem 1. ■

For completeness, we briefly mention the main idea behind [15, Proposition 3, p. 484]. Basically, one first exploits the optimality condition of the proximal map to show that Algorithm 1 is making *sufficient* progress in each iteration. Thanks to a generalization of the celebrated Łojasiewicz gradient inequality [34], one then lower bounds the progress by the minimum norm of the subgradients. Together these allow one to show 0, asymptotically, is in the subdifferential, i.e., the algorithm converges to a critical point.

We emphasize that in general it is much harder to prove convergence in terms of iterates rather than the function values. This is not by chance. Indeed, Sard's celebrated theorem shows that there can only be few (Lebesgue null measure) possible function values at all critical points, while in contrast, the set of all critical points can be arbitrarily large. Think of the constant function: The function takes only a single value at all critical points, which consist of the whole space.

I Proof of Theorem 2

Proof: The proof is a slight generalization of that in [14] to the nonconvex setting.

If Algorithm 1 converges to an ϵ-local minimizer $\tilde{\mathbf{w}}$ of (9), then

$$\ell(\tilde{\mathbf{w}}) + \mathsf{A}^\mu(\tilde{\mathbf{w}}) \leq \ell(\mathbf{w}) + \mathsf{A}^\mu(\mathbf{w}) + \epsilon.$$

for all $\mathbf{w}$ in a neighborhood $N_{\tilde{\mathbf{w}}}$ of $\tilde{\mathbf{w}}$. Applying Proposition 6 we have

$$\begin{aligned}[\ell(\tilde{\mathbf{w}}) + \bar{f}(\tilde{\mathbf{w}})] - [\ell(\mathbf{w}) + \bar{f}(\mathbf{w})] &= [\ell(\tilde{\mathbf{w}}) + \mathsf{A}^\mu(\tilde{\mathbf{w}})] - [\ell(\mathbf{w}) + \mathsf{A}^\mu(\mathbf{w})] \\ &\quad + [\bar{f}(\tilde{\mathbf{w}}) - \mathsf{A}^\mu(\tilde{\mathbf{w}})] - [\bar{f}(\mathbf{w}) - \mathsf{A}^\mu(\mathbf{w})] \\ &\leq \epsilon + \epsilon + 0 = 2\epsilon.\end{aligned}$$

Therefore $\tilde{\mathbf{w}}$ is an (2ϵ)-local minimizer. The proof for $\tilde{\mathbf{w}}$ being a ϵ-global minimizer is similar. ■

J Derivation of the proximal map in Example 1

We derive the proximal map for the truncated graph-ℓ_1 norm. Note that the problem can be reduced to $\mathbb{R}^2$ by considering each edge separately.

J.1 Truncated ℓ_1

As a warm up, we first repeat the computation for the truncated ℓ_1 norm: $|w|_t = \min\{|w|, \tau\} := |w| \wedge \tau$. We remark that the explicit form of the proximal map (for $\mu = \tau$) has already appeared in [1, Fan's comment].

We use the following variational representation:

$$|w|_t = \min_{0 \le \eta \le 1} \eta |w| + (1-\eta)\tau. \tag{16}$$

Therefore the proximal map can be rewritten as

$$\min_z \tfrac{1}{2\mu}(w-z)^2 + |z|_t = \min_{0 \le \eta \le 1} \min_z \tfrac{1}{2\mu}(w-z)^2 + \eta|z| + (1-\eta)\tau. \tag{17}$$

For fixed η, clearly we have the soft-shrinkage operator

$$z = \mathrm{sign}(w) \cdot (|w| - \mu\eta)_+. \tag{18}$$

Plug in back to (17), we obtain

$$\min_{0 \le \eta \le 1} \tfrac{1}{2\mu} \Big[|w| \wedge (\mu\eta) \Big]^2 + \eta(|w| - \mu\eta)_+ + (1-\eta)\tau. \tag{19}$$

Once we find an optimal η, plugging it back to (18) immediately yields the proximal map. Clearly (19) is a piecewise quadratic function of η, thus we divide our discussion into several cases.

Case 1: $|w| \ge \mu$. In this case we obviously have $|w| \ge \mu\eta$ since $\eta \in [0,1]$. Therefore (19) simplifies to

$$\min_{0 \le \eta \le 1} -\tfrac{1}{2}\mu\eta^2 + \eta(|w| - \tau) + \tau.$$

Since $\mu \ge 0$, we are minimizing a *concave* quadratic, thus the minimizer must be attained at the extreme points 0 or 1. Comparing the resulting objective values gives us:

$$\eta = \begin{cases} 0, & \text{if } |w| > \frac{\mu}{2} + \tau \\ 1, & \text{if } |w| < \frac{\mu}{2} + \tau \\ \{0,1\}, & \text{if } |w| = \frac{\mu}{2} + \tau \end{cases}. \tag{20}$$

Case 2: $|w| \le \mu$. We need to further distinguish two subcases.

$$\min_{0 \le \eta \le |w|/\mu} -\tfrac{1}{2}\mu\eta^2 + \eta(|w| - \tau) + \tau \quad \text{v.s.} \quad \min_{|w|/\mu \le \eta \le 1} \tfrac{1}{2\mu} w^2 + (1-\eta)\tau$$

For the first subcase, again the minimizer is attained at one of the extreme points, namely, $\eta = 0$ with objective τ and $\eta = |w|/\mu$ with objective $\frac{w^2}{2\mu} - \frac{|w|\tau}{\mu} + \tau$. For the second subcase, clearly $\eta = 1$ with objective $\frac{1}{2\mu}w^2$. Note that

$$\tfrac{w^2}{2\mu} - \tfrac{|w|\tau}{\mu} + \tau \ge \tfrac{1}{2\mu} w^2$$

due to our assumption $|w| \le \mu$. Thus $\eta = 0$ only when $\tau \le \frac{1}{2\mu} w^2$, and $\eta = 1$ otherwise. To summarize,

$$\eta = \begin{cases} 0, & \text{if } |w| > \sqrt{2\mu\tau} \\ 1, & \text{if } |w| < \sqrt{2\mu\tau} \\ \{0,1\}, & \text{if } |w| = \sqrt{2\mu\tau} \end{cases}. \tag{21}$$

For a quick sanity check, consider when $\tau = 0$, in both cases we would have $\eta = 0$, resulting in $z = w$. Similarly when $\tau = \infty$, in both cases we would have $\eta = 1$, resulting in the soft-thresholding operator.

J.2 Truncated graph-ℓ_1

Next consider the slightly more complicated problem:

$$\min_{z_1, z_2} \tfrac{1}{2\mu}[(z_1 - w_1)^2 + (z_2 - w_2)^2] + |z_1 - z_2|_t \tag{22}$$

$$= \min_{0 \le \eta \le 1} \min_{z_1, z_2} \tfrac{1}{2\mu}[(z_1 - w_1)^2 + (z_2 - w_2)^2] + \eta|z_1 - z_2| + (1-\eta)\tau. \tag{23}$$

For any fixed η, we know from [14] that

$$z_1 = w_1 - \mathrm{sign}(w_1 - w_2)\left[(\mu\eta) \wedge \tfrac{|w_1 - w_2|}{2}\right] \tag{24}$$

$$z_2 = w_2 + \mathrm{sign}(w_1 - w_2)\left[(\mu\eta) \wedge \tfrac{|w_1 - w_2|}{2}\right]. \tag{25}$$

Therefore, we need only find an optimal η from

$$\min_{0 \le \eta \le 1} \tfrac{1}{\mu}\left[(\mu^2\eta^2) \wedge \tfrac{\delta^2}{4}\right] + \eta(\delta - 2\mu\eta)_+ + (1-\eta)\tau,$$

where for clarity we let $\delta := |w_1 - w_2|$. Like before, we will divide the analysis into several cases.

Case 1: $\delta \ge 2\mu$, implying that $\delta \ge 2\mu\eta$. Simplifying to obtain

$$\min_{0 \le \eta \le 1} -\mu\eta^2 + \eta(\delta - \tau) + \tau.$$

Comparing the objectives at the two extreme points yields

$$\eta = \begin{cases} 0, & \text{if } |w_1 - w_2| > \mu + \tau \\ 1, & \text{if } |w_1 - w_2| < \mu + \tau \\ \{0,1\}, & \text{if } |w_1 - w_2| = \mu + \tau \end{cases}. \tag{26}$$

Case 2: $\delta \le 2\mu$. Consider further the two subcases:

$$\min_{0 \le \eta \le \frac{\delta}{2\mu}} -\mu\eta^2 + \eta(\delta - \tau) + \tau \qquad \text{v.s.} \qquad \min_{\frac{\delta}{2\mu} \le \eta \le 1} \tfrac{\delta^2}{4\mu} + (1-\eta)\tau$$

For the first subcase, again the minimizer is attained at one of the extreme points, namely, $\eta = 0$ with objective τ and $\eta = \frac{\delta}{2\mu}$ with objective $\frac{\delta^2}{4\mu} - \frac{\delta\tau}{2\mu} + \tau$. For the second subcase, clearly $\eta = 1$ with objective $\frac{\delta^2}{4\mu}$. Note that

$$\tfrac{\delta^2}{4\mu} - \tfrac{\delta\tau}{2\mu} + \tau \ge \tfrac{\delta^2}{4\mu}$$

due to our assumption $\delta \le 2\mu$. Thus $\eta = 0$ only when $\tau \le \frac{\delta^2}{4\mu}$, and $\eta = 1$ otherwise. To summarize,

$$\eta = \begin{cases} 0, & \text{if } |w_1 - w_2| > 2\sqrt{\mu\tau} \\ 1, & \text{if } |w_1 - w_2| < 2\sqrt{\mu\tau} \\ \{0,1\}, & \text{if } |w_1 - w_2| = 2\sqrt{\mu\tau} \end{cases}. \tag{27}$$

Combining the two cases we have

$$\eta = \begin{cases} 0, & |w_1 - w_2| > 2\sqrt{\mu\tau} + \left((\sqrt{\tau} - \sqrt{\mu})_+\right)^2 \\ 1, & |w_1 - w_2| < 2\sqrt{\mu\tau} + \left((\sqrt{\tau} - \sqrt{\mu})_+\right)^2 \\ \{0,1\}, & |w_1 - w_2| = 2\sqrt{\mu\tau} + \left((\sqrt{\tau} - \sqrt{\mu})_+\right)^2 \end{cases}. \tag{28}$$

K Derivation of the proximal map in Example 2

We gives the details on how to compute the proximal map for the truncated hinge loss:

$$f(\mathbf{w}) = \min(\max(\rho - y\hat{y}, 0), \tau), \tag{29}$$

where $\hat{y} = \mathbf{w}^\top \mathbf{x}$. For simplicity we do not include the intercept. The margin parameter ρ is usually set to 1; we keep it variable here for flexibility.

K.1 The proximal map for the hinge loss

For completeness and ease of derivation, let us first compute the proximal map for the hinge loss (i.e., $\tau = \infty$). By definition,

$$\begin{aligned}
\min_{\mathbf{z}} \tfrac{1}{2\mu}\|\mathbf{w}-\mathbf{z}\|^2 + \max(\rho - y\mathbf{z}^\top\mathbf{x}, 0) &= \min_{\mathbf{z}} \max_{\alpha\in[0,1]} \tfrac{1}{2\mu}\|\mathbf{w}-\mathbf{z}\|^2 + \alpha(\rho - y\mathbf{z}^\top\mathbf{x}) \\
&= \max_{\alpha\in[0,1]} \min_{\mathbf{z}} \tfrac{1}{2\mu}\|\mathbf{w}-\mathbf{z}\|^2 + \alpha(\rho - y\mathbf{z}^\top\mathbf{x}) \\
(\mathbf{z} = \mathbf{w} + \mu\alpha y\mathbf{x}) \quad &= \max_{\alpha\in[0,1]} \alpha(\rho - y\mathbf{w}^\top\mathbf{x}) - \tfrac{\mu}{2} y\mathbf{x}^\top\mathbf{x}y\alpha^2
\end{aligned}$$

Without the constraint we should take $\alpha = \frac{\rho - y\mathbf{w}^\top\mathbf{x}}{\mu y\mathbf{x}^\top\mathbf{x}y}$. We have three cases, and we discuss them separately.

Case 1: $y\mathbf{w}^\top\mathbf{x} \geq \rho$. Then $\alpha = 0$, $\mathbf{z} = \mathbf{w}$, and the objective is 0.

Case 2: $y\mathbf{w}^\top\mathbf{x} + \mu y\mathbf{x}^\top\mathbf{x}y \leq \rho$. Then $\alpha = 1$, $\mathbf{z} = \mathbf{w} + \mu y\mathbf{x}$, and the objective is $\rho - y\mathbf{w}^\top\mathbf{x} - \frac{\mu}{2}y\mathbf{x}^\top\mathbf{x}y$.

Case 3: $\rho - \mu y\mathbf{x}^\top\mathbf{x}y \leq y\mathbf{w}^\top\mathbf{x} \leq \rho$. Then $\alpha = \frac{\rho - y\mathbf{w}^\top\mathbf{x}}{\mu y\mathbf{x}^\top\mathbf{x}y}$, $\mathbf{z} = \mathbf{w} + \frac{\rho - y\mathbf{w}^\top\mathbf{x}}{y\mathbf{x}^\top\mathbf{x}y}y\mathbf{x}$, and the objective is $\frac{1}{2\mu}\frac{(\rho - y\mathbf{w}^\top\mathbf{x})^2}{y\mathbf{x}^\top\mathbf{x}y}$.

Of course we can incorporate all three cases into a single formula: $\mathbf{z} = \mathbf{w} + \left[\frac{\rho - y\mathbf{w}^\top\mathbf{x}}{y\mathbf{x}^\top\mathbf{x}y}\right]_0^\mu \cdot y\mathbf{x}$, where $[\cdot]_0^\mu$ denotes the projection into the interval $[0, \mu]$.

K.2 The proximal map for the truncated hinge loss

Now we are ready for the truncated hinge loss:

$$\begin{aligned}
\min_{\mathbf{z}} \tfrac{1}{2\mu}\|\mathbf{w}-\mathbf{z}\|^2 + \min(\max(\rho - y\mathbf{z}^\top\mathbf{x}, 0), \tau) &= \min_{\eta\in[0,1]} \min_{\mathbf{z}} \tfrac{1}{2\mu}\|\mathbf{w}-\mathbf{z}\|^2 + \eta(\rho - y\mathbf{z}^\top\mathbf{x})_+ + (1-\eta)\tau \\
&= \min_{\eta\in[0,1]} (1-\eta)\tau + \eta\cdot\left[\min_{\mathbf{z}} \tfrac{1}{2\mu\eta}\|\mathbf{w}-\mathbf{z}\|^2 + (\rho - y\mathbf{z}^\top\mathbf{x})_+\right],
\end{aligned}$$

where the inner minimization is exactly what we have computed before (with the minor change $\mu \to \mu\eta$). Using our previous computation, we again have three cases.

Case 1: $y\mathbf{w}^\top\mathbf{x} \geq \rho$. Then $\eta = 1, \mathbf{z} = \mathbf{w}$ and the objective is 0.

Case 2: $y\mathbf{w}^\top\mathbf{x} + \mu\eta y\mathbf{x}^\top\mathbf{x}y \leq \rho$. Note that this gives us the additional constraint $\eta \leq \frac{\rho - y\mathbf{w}^\top\mathbf{x}}{\mu y\mathbf{x}^\top\mathbf{x}y}$. Let $a = \min\{1, \frac{\rho - y\mathbf{w}^\top\mathbf{x}}{\mu y\mathbf{x}^\top\mathbf{x}y}\}$. The objective now is simply

$$\min_{\eta\in[0,a]} (1-\eta)\tau + \eta(\rho - y\mathbf{w}^\top\mathbf{x} - \tfrac{\mu\eta}{2}y\mathbf{x}^\top\mathbf{x}y).$$

This is a *concave* quadratic function and we are *minimizing* it. Therefore the minimizer must be one of the extreme points, namely 0 or a. We simply compare their objectives and pick the smaller one. For concreteness, let us further divide into two subcases.

Case 2.1: $\frac{\rho - y\mathbf{w}^\top\mathbf{x}}{\mu y\mathbf{x}^\top\mathbf{x}y} \geq 1$. Then $a = 1$ and

$$\eta = \begin{cases} 0, & \text{if } \tau \leq \rho - y\mathbf{w}^\top\mathbf{x} - \frac{\mu}{2}y\mathbf{x}^\top\mathbf{x}y \\ 1, & \text{otherwise} \end{cases}. \tag{30}$$

And of course $\mathbf{z} = \mathbf{w} + \mu\eta y\mathbf{x}$ (recall that $\alpha = 1$ in this case).

Case 2.2: $\frac{\rho - y\mathbf{w}^\top\mathbf{x}}{\mu y\mathbf{x}^\top\mathbf{x}y} \leq 1$. Similarly we compare the objectives of the extreme points:

$$\eta = \begin{cases} 0, & \text{if } \tau \leq \mathit{blablabla} \\ \frac{\rho - y\mathbf{w}^\top\mathbf{x}}{\mu y\mathbf{x}^\top\mathbf{x}y}, & \text{otherwise} \end{cases}. \tag{31}$$

Note that there is no need to compute blablabla since it will be discarded anyway, as we will see. In the first case we have $\mathbf{z} = \mathbf{w}$ while in the second case $\mathbf{z} = \mathbf{w} + \frac{1 - y\mathbf{w}^\top\mathbf{x}}{y\mathbf{x}^\top\mathbf{x}y}y\mathbf{x}$.

Case 3: $\rho - \mu\eta y\mathbf{x}^\top\mathbf{x}y \le y\mathbf{w}^\top\mathbf{x} \le \rho$. This leads to the additional constraint $\eta \ge \frac{\rho - y\mathbf{w}^\top\mathbf{x}}{\mu y\mathbf{x}^\top\mathbf{x}y}$, therefore we must have $\frac{\rho - y\mathbf{w}^\top\mathbf{x}}{\mu y\mathbf{x}^\top\mathbf{x}y} \le 1$ for otherwise this case is vacuous. The objective is simply

$$\min_{\eta \in [\frac{\rho - y\mathbf{w}^\top\mathbf{x}}{\mu y\mathbf{x}^\top\mathbf{x}y}, 1]} (1-\eta)\tau + \frac{1}{2\mu}\frac{(\rho - y\mathbf{w}^\top\mathbf{x})^2}{y\mathbf{x}^\top\mathbf{x}y}. \tag{32}$$

Thus $\eta = 1$ and $\mathbf{z} = \mathbf{w} + \frac{\rho - y\mathbf{w}^\top\mathbf{x}}{y\mathbf{x}^\top\mathbf{x}y} y\mathbf{x}$, which happens to be the same as the second subcase in **case 2.2**. This implies that both have the objective shown in (32), which is decreasing in η. Since the second subcase in **case 2.2** has $\eta = \frac{\rho - y\mathbf{w}^\top\mathbf{x}}{\mu y\mathbf{x}^\top\mathbf{x}y}$ while the current case has $\eta = 1$, we will always pick the latter.

To be definite, let us summarize the above computations. The proximal map of the truncated hinge loss is given as follows:

$$\eta = \begin{cases} 1, & \text{if } \rho - y\mathbf{w}^\top\mathbf{x} \le 0 \\ 0, & \text{if } \frac{\rho - y\mathbf{w}^\top\mathbf{x}}{\mu y\mathbf{x}^\top\mathbf{x}y} \ge 1 \ \ \&\& \ \ \tau < \rho - y\mathbf{w}^\top\mathbf{x} - \frac{\mu}{2} y\mathbf{x}^\top\mathbf{x}y \\ 1, & \text{if } \frac{\rho - y\mathbf{w}^\top\mathbf{x}}{\mu y\mathbf{x}^\top\mathbf{x}y} \ge 1 \ \ \&\& \ \ \tau > \rho - y\mathbf{w}^\top\mathbf{x} - \frac{\mu}{2} y\mathbf{x}^\top\mathbf{x}y \\ \{0,1\}, & \text{if } \frac{\rho - y\mathbf{w}^\top\mathbf{x}}{\mu y\mathbf{x}^\top\mathbf{x}y} \ge 1 \ \ \&\& \ \ \tau = \rho - y\mathbf{w}^\top\mathbf{x} - \frac{\mu}{2} y\mathbf{x}^\top\mathbf{x}y \\ 0, & \text{if } 0 \le \frac{\rho - y\mathbf{w}^\top\mathbf{x}}{\mu y\mathbf{x}^\top\mathbf{x}y} \le 1 \ \ \&\& \ \ \tau < \frac{1}{2\mu}\frac{(\rho - y\mathbf{w}^\top\mathbf{x})^2}{y\mathbf{x}^\top\mathbf{x}y} \\ 1, & \text{if } 0 \le \frac{\rho - y\mathbf{w}^\top\mathbf{x}}{\mu y\mathbf{x}^\top\mathbf{x}y} \le 1 \ \ \&\& \ \ \tau > \frac{1}{2\mu}\frac{(\rho - y\mathbf{w}^\top\mathbf{x})^2}{y\mathbf{x}^\top\mathbf{x}y} \\ \{0,1\}, & \text{if } 0 \le \frac{\rho - y\mathbf{w}^\top\mathbf{x}}{\mu y\mathbf{x}^\top\mathbf{x}y} \le 1 \ \ \&\& \ \ \tau = \frac{1}{2\mu}\frac{(\rho - y\mathbf{w}^\top\mathbf{x})^2}{y\mathbf{x}^\top\mathbf{x}y} \end{cases}, \tag{33}$$

$$\mathbf{z} = \begin{cases} \mathbf{w} \\ \mathbf{w} \\ \mathbf{w} + \mu y\mathbf{x} \\ \{\mathbf{w}, \mathbf{w} + \mu y\mathbf{x}\} \\ \mathbf{w} \\ \mathbf{w} + \frac{\rho - y\mathbf{w}^\top\mathbf{x}}{y\mathbf{x}^\top\mathbf{x}y} y\mathbf{x} \\ \{\mathbf{w}, \mathbf{w} + \frac{\rho - y\mathbf{w}^\top\mathbf{x}}{y\mathbf{x}^\top\mathbf{x}y} y\mathbf{x}\} \end{cases}, \tag{34}$$

$$\mathsf{e}_f^\mu(\mathbf{w}) = \begin{cases} 0, & \text{if } \rho - y\mathbf{w}^\top\mathbf{x} \le 0 \\ \tau, & \text{if } \frac{\rho - y\mathbf{w}^\top\mathbf{x}}{\mu y\mathbf{x}^\top\mathbf{x}y} \ge 1 \ \ \&\& \ \ \tau \le \rho - y\mathbf{w}^\top\mathbf{x} - \frac{\mu}{2} y\mathbf{x}^\top\mathbf{x}y \\ \rho - y\mathbf{w}^\top\mathbf{x} - \frac{\mu}{2} y\mathbf{x}^\top\mathbf{x}y, & \text{if } \frac{\rho - y\mathbf{w}^\top\mathbf{x}}{\mu y\mathbf{x}^\top\mathbf{x}y} \ge 1 \ \ \&\& \ \ \tau \ge \rho - y\mathbf{w}^\top\mathbf{x} - \frac{\mu}{2} y\mathbf{x}^\top\mathbf{x}y \\ \tau, & \text{if } 0 \le \frac{\rho - y\mathbf{w}^\top\mathbf{x}}{\mu y\mathbf{x}^\top\mathbf{x}y} \le 1 \ \ \&\& \ \ \tau \le \frac{1}{2\mu}\frac{(\rho - y\mathbf{w}^\top\mathbf{x})^2}{y\mathbf{x}^\top\mathbf{x}y} \\ \frac{1}{2\mu}\frac{(\rho - y\mathbf{w}^\top\mathbf{x})^2}{y\mathbf{x}^\top\mathbf{x}y}, & \text{if } 0 \le \frac{\rho - y\mathbf{w}^\top\mathbf{x}}{\mu y\mathbf{x}^\top\mathbf{x}y} \le 1 \ \ \&\& \ \ \tau \ge \frac{1}{2\mu}\frac{(\rho - y\mathbf{w}^\top\mathbf{x})^2}{y\mathbf{x}^\top\mathbf{x}y} \end{cases}. \tag{35}$$

For a quick sanity check, let us see what happens when $\tau \to \infty$: only the 1st, 3rd, and 5th cases survive, which matches precisely what we had before for the hinge loss.

Additional References

[31] Jonathan M. Borwein and Xianfu Wang. "Distinct differentiable functions may share the same Clarke subdifferential at all points." *Proceedings of the American Mathematical Society*, vol. 125, no. 3 (1997), pp. 807–813 (cit. on p. 10).

[32] Robert R. Phelps. *Convex Functions, Monotone Operators and Differentiability*. 2nd. Springer, 1993 (cit. on pp. 13, 15).

[33] Anestis Antoniadis. "Wavelet methods in statistics: Some recent developments and their applications." *Statistical Surveys*, vol. 1 (2007), pp. 16–55 (cit. on p. 15).

[34] Jérôme Bolte, Aris Danilidis, Adrian Lewis, and Masahiro Shiota. "Clarke Subgradients of Stratifiable Functions." *SIAM Journal on Optimization*, vol. 18, no. 2 (2007), pp. 556–572 (cit. on p. 18).