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:

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

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:

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

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

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

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 $h = \frac{1}{2\mu}\left\|\cdot\right\|^2 - f$ be convex and finite valued. Then the function $j(\mathbf{w}) := h^*(\mu^{-1}\mathbf{w}) - \frac{1}{2\mu}\left\|\mathbf{w}\right\|^2$ 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 $0 \le \langle \mathbf{w}_n - \mathbf{z}_m, \mathbf{u}_n - \mathbf{v}_m \rangle \to \langle \mathbf{w} - \mathbf{z}, \mathbf{u} - \mathbf{v} \rangle$. ■

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 $\mathsf{h}_f^\mu := -\mathsf{e}_{(-\mathsf{e}_f^\mu)}^\mu$. 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 *$\mathsf{H}_f^\mu = \mathsf{P}_{\mathsf{h}_f^\mu}^\mu$.*

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 $\mathsf{P}_{\mathsf{h}_f^\mu}^\mu$ is maximal monotone. Since $\mathsf{e}_f^\mu = \mathsf{e}_{\mathsf{h}_f^\mu}^\mu$ [26, Example 1.44] and f agrees with h_f^μ on $\mathrm{Im}(\mathsf{P}_f^\mu)$, it easily follows that $\mathsf{P}_{\mathsf{h}_f^\mu}^\mu(\mathbf{w}) \supseteq \mathsf{P}_f^\mu(\mathbf{w})$ 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 $\mathsf{P}_{\mathsf{h}_f^\mu}^\mu$ 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

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

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

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

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

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

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

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

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

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

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

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

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

Therefore, we need only find an optimal η from

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

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:

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

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

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

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:

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

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

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

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).