Supplementary Material

A PROXIMITY OPERATORS AND MOREAU PROJECTIONS

Throughout, we let $\varphi : \mathbb{R}^p \to \bar{\mathbb{R}}$ (where $\bar{\mathbb{R}} \triangleq \mathbb{R} \cup \{+\infty\}$) be a convex, lower semicontinuous (lsc) (the epigraph epi$\varphi \triangleq \{(x,t) \in \mathbb{R}^p \times \mathbb{R} \,|\, \varphi(x) \le t\}$ is closed in $\mathbb{R}^p \times \mathbb{R}$), and proper ($\exists \mathbf{x} : \varphi(\mathbf{x}) \neq +\infty$) function. The *Fenchel conjugate* of φ is $\varphi^\star : \mathbb{R}^p \to \bar{\mathbb{R}}$, $\varphi^\star(\mathbf{y}) \triangleq \sup_{\mathbf{x}} \mathbf{y}^\top \mathbf{x} - \varphi(\mathbf{x})$. Let:

$$M_\varphi(\mathbf{y}) \triangleq \inf_{\mathbf{x}} \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|^2 + \varphi(\mathbf{x}), \quad \text{and} \quad \mathrm{prox}_\varphi(\mathbf{y}) = \arg\inf_{\mathbf{x}} \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|^2 + \varphi(\mathbf{x});$$

the function $M_\varphi : \mathbb{R}^p \to \bar{\mathbb{R}}$ is called the *Moreau envelope* of φ, and the map $\mathrm{prox}_\varphi : \mathbb{R}^p \to \mathbb{R}^p$ is the *proximity operator* of φ (Combettes and Wajs, 2006; Moreau, 1962). Proximity operators generalize Euclidean projectors: consider the case $\varphi = \iota_{\mathcal{C}}$, where $\mathcal{C} \subseteq \mathbb{R}^p$ is a convex set and $\iota_{\mathcal{C}}$ denotes its indicator (*i.e.*, $\varphi(\mathbf{x}) = 0$ if $\mathbf{x} \in \mathcal{C}$ and $+\infty$ otherwise). Then, prox_φ is the Euclidean projector onto $\mathcal{C}$ and M_φ is the residual. Two other important examples of proximity operators follow:

- if $\varphi(\mathbf{x}) = (\lambda/2)\|\mathbf{x}\|^2$, then $\mathrm{prox}_\varphi(\mathbf{y}) = \mathbf{y}/(1+\lambda)$;
- if $\varphi(\mathbf{x}) = \tau\|\mathbf{x}\|_1$, then $\mathrm{prox}_\varphi(\mathbf{y}) = \mathrm{soft}(\mathbf{y}, \tau)$ is the *soft-threshold* function (Wright et al., 2009), defined as $[\mathrm{soft}(\mathbf{y},\tau)]_k = \mathrm{sgn}(y_k) \cdot \max\{0, |y_k| - \tau\}$.

If $\varphi : \mathbb{R}^{d_1} \times \ldots \times \mathbb{R}^{d_p} \to \bar{\mathbb{R}}$ is (group-)separable, *i.e.*, $\varphi(\mathbf{x}) = \sum_{k=1}^p \varphi_k(\mathbf{x}_k)$, where $\mathbf{x}_k \in \mathbb{R}^{d_k}$, then its proximity operator inherits the same (group-)separability: $[\mathrm{prox}_\varphi(\mathbf{x})]_k = \mathrm{prox}_{\varphi_k}(\mathbf{x}_k)$ (Wright et al., 2009). For example, the proximity operator of the mixed $\ell_{2,1}$-norm, which is group-separable, has this form. The following proposition extends this result by showing how to compute proximity operators of functions (maybe not separable) that only depend on the ℓ_2-norms of groups of components; *e.g.*, the proximity operator of the squared $\ell_{2,1}$-norm reduces to that of squared ℓ_1.

Proposition 5 *Let $\varphi : \mathbb{R}^{d_1} \times \ldots \times \mathbb{R}^{d_p} \to \bar{\mathbb{R}}$ be of the form $\varphi(\mathbf{x}_1, \ldots, \mathbf{x}_p) = \psi(\|\mathbf{x}_1\|, \ldots, \|\mathbf{x}_p\|)$ for some $\psi : \mathbb{R}^p \to \bar{\mathbb{R}}$. Then, $M_\varphi(\mathbf{x}_1, \ldots, \mathbf{x}_p) = M_\psi(\|\mathbf{x}_1\|, \ldots, \|\mathbf{x}_p\|)$ and $[\mathrm{prox}_\varphi(\mathbf{x}_1, \ldots, \mathbf{x}_p)]_k = [\mathrm{prox}_\psi(\|\mathbf{x}_1\|, \ldots, \|\mathbf{x}_p\|)]_k(\mathbf{x}_k/\|\mathbf{x}_k\|)$.*

Proof: We have respectively:

$$\begin{aligned}
M_\varphi(\mathbf{x}_1, \ldots, \mathbf{x}_p) &= \min_{\mathbf{y}} \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|^2 + \varphi(\mathbf{y}) \\
&= \min_{\mathbf{y}_1, \ldots, \mathbf{y}_p} \frac{1}{2}\sum_{k=1}^p \|\mathbf{y}_k - \mathbf{x}_k\|^2 + \psi(\|\mathbf{y}_1\|, \ldots, \|\mathbf{y}_p\|) \\
&= \min_{\mathbf{u} \in \mathbb{R}_+^p} \psi(u_1, \ldots, u_p) + \min_{\mathbf{y} : \|\mathbf{y}_k\| = u_k, \forall k} \frac{1}{2}\sum_{k=1}^p \|\mathbf{y}_k - \mathbf{x}_k\|^2 \\
&= \min_{\mathbf{u} \in \mathbb{R}_+^p} \psi(u_1, \ldots, u_p) + \frac{1}{2}\sum_{k=1}^p \min_{\mathbf{y}_k : \|\mathbf{y}_k\| = u_k} \|\mathbf{y}_k - \mathbf{x}_k\|^2 \quad (^*) \\
&= \min_{\mathbf{u} \in \mathbb{R}_+^p} \psi(u_1, \ldots, u_p) + \frac{1}{2}\sum_{k=1}^p \left\| \frac{u_k}{\|\mathbf{x}_k\|}\mathbf{x}_k - \mathbf{x}_k \right\|^2 \\
&= \min_{\mathbf{u} \in \mathbb{R}_+^p} \psi(u_1, \ldots, u_p) + \frac{1}{2}\sum_{k=1}^p (u_k - \|\mathbf{x}_k\|)^2 \\
&= M_\psi(\|\mathbf{x}_1\|, \ldots, \|\mathbf{x}_p\|),
\end{aligned} \tag{18}$$

where the solution of the innermost minimization problem in $(^*)$ is $\mathbf{y}_k = \frac{u_k}{\|\mathbf{x}_k\|}\mathbf{x}_k$, and therefore $[\mathrm{prox}_\varphi(\mathbf{x}_1, \ldots, \mathbf{x}_p)]_k = [\mathrm{prox}_\psi(\|\mathbf{x}_1\|, \ldots, \|\mathbf{x}_p\|)]_k \frac{\mathbf{x}_k}{\|\mathbf{x}_k\|}$. ■

Finally, we recall the *Moreau decomposition*, relating the proximity operators of Fenchel conjugate functions (Combettes and Wajs, 2006) and present a corollary that is the key to our regret bound in §3.3.

Proposition 6 (Moreau (1962)) *For any convex, lsc, proper function $\varphi : \mathbb{R}^p \to \bar{\mathbb{R}}$,*

$$\mathbf{x} = \mathrm{prox}_\varphi(\mathbf{x}) + \mathrm{prox}_{\varphi^\star}(\mathbf{x}) \quad \textit{and} \quad \|\mathbf{x}\|^2/2 = M_\varphi(\mathbf{x}) + M_{\varphi^\star}(\mathbf{x}). \tag{19}$$

Corollary 7 *Let $\varphi : \mathbb{R}^p \to \bar{\mathbb{R}}$ be as in Prop. 6, and $\bar{\mathbf{x}} \triangleq \mathrm{prox}_\varphi(\mathbf{x})$. Then, any $\mathbf{y} \in \mathbb{R}^p$ satisfies*

$$\|\mathbf{y} - \bar{\mathbf{x}}\|^2 - \|\mathbf{y} - \mathbf{x}\|^2 \leq 2(\varphi(\mathbf{y}) - \varphi(\bar{\mathbf{x}})). \tag{20}$$

Proof: We start by stating and proving the following lemma:

Lemma 8 *Let $\varphi : \mathbb{R}^p \to \bar{\mathbb{R}}$ be as in Prop. 6, and let $\bar{\mathbf{x}} \triangleq \mathrm{prox}_\varphi(\mathbf{x})$. Then, any $\mathbf{y} \in \mathbb{R}^p$ satisfies*

$$(\bar{\mathbf{x}} - \mathbf{y})^\top (\bar{\mathbf{x}} - \mathbf{x}) \leq \varphi(\mathbf{y}) - \varphi(\bar{\mathbf{x}}) \tag{21}$$

Proof (of the Lemma): From (19), we have that

$$\begin{aligned}
\frac{1}{2}\|\mathbf{x}\|^2 &= \frac{1}{2}\|\bar{\mathbf{x}} - \mathbf{x}\|^2 + \varphi(\bar{\mathbf{x}}) + \frac{1}{2}\|\bar{\mathbf{x}}\|^2 + \varphi^*(\mathbf{x} - \bar{\mathbf{x}}) \\
&= \frac{1}{2}\|\bar{\mathbf{x}} - \mathbf{x}\|^2 + \varphi(\bar{\mathbf{x}}) + \frac{1}{2}\|\bar{\mathbf{x}}\|^2 + \sup_{\mathbf{u} \in \mathbb{R}^p} \left(\mathbf{u}^\top(\mathbf{x} - \bar{\mathbf{x}}) - \varphi(\mathbf{u})\right) \\
&\geq \frac{1}{2}\|\bar{\mathbf{x}} - \mathbf{x}\|^2 + \varphi(\bar{\mathbf{x}}) + \frac{1}{2}\|\bar{\mathbf{x}}\|^2 + \mathbf{y}^\top(\mathbf{x} - \bar{\mathbf{x}}) - \varphi(\mathbf{y}) \\
&= \frac{1}{2}\|\mathbf{x}\|^2 + \bar{\mathbf{x}}^\top(\bar{\mathbf{x}} - \mathbf{x}) + \mathbf{y}^\top(\mathbf{x} - \bar{\mathbf{x}}) - \varphi(\mathbf{y}) + \varphi(\bar{\mathbf{x}}) \\
&= \frac{1}{2}\|\mathbf{x}\|^2 + (\bar{\mathbf{x}} - \mathbf{y})^\top(\bar{\mathbf{x}} - \mathbf{x}) - \varphi(\mathbf{y}) + \varphi(\bar{\mathbf{x}}),
\end{aligned}$$

from which (21) follows. ■

Now, take Lemma 8 and bound the left hand side as:

$$\begin{aligned}
(\bar{\mathbf{x}} - \mathbf{y})^\top(\bar{\mathbf{x}} - \mathbf{x}) &\geq (\bar{\mathbf{x}} - \mathbf{y})^\top(\bar{\mathbf{x}} - \mathbf{x}) - \frac{1}{2}\|\bar{\mathbf{x}} - \mathbf{x}\|^2 \\
&= (\bar{\mathbf{x}} - \mathbf{y})^\top(\bar{\mathbf{x}} - \mathbf{x}) - \frac{1}{2}\|\bar{\mathbf{x}}\|^2 - \frac{1}{2}\|\mathbf{x}\|^2 + \bar{\mathbf{x}}^\top\mathbf{x} \\
&= \frac{1}{2}\|\bar{\mathbf{x}}\|^2 - \mathbf{y}^\top(\bar{\mathbf{x}} - \mathbf{x}) - \frac{1}{2}\|\mathbf{x}\|^2 \\
&= \frac{1}{2}\|\mathbf{y} - \bar{\mathbf{x}}\|^2 - \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|^2.
\end{aligned}$$

This concludes the proof of Corollary 7. ■

Note that although the Fenchel dual $\varphi^\star$ does not show up in (20), it has a crucial role in this proof.

B PROOF OF LEMMA 2

Let $u(\bar{\boldsymbol{\theta}}, \boldsymbol{\theta}) \triangleq \lambda\Omega(\bar{\boldsymbol{\theta}}) - \lambda\Omega(\boldsymbol{\theta})$. We have successively:

$$\begin{aligned}
\|\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_{t+1}\|^2 &\leq^{\text{(i)}} \|\bar{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}_{t+1}\|^2 \\
&\leq^{\text{(ii)}} \|\bar{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}_t\|^2 + 2\eta_t\lambda \textstyle\sum_{j=1}^J (\Omega_j(\bar{\boldsymbol{\theta}}) - \Omega_j(\tilde{\boldsymbol{\theta}}_{t+j/J})) \\
&\leq^{\text{(iii)}} \|\bar{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}_t\|^2 + 2\eta_t u(\bar{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}_{t+1}) \\
&\leq^{\text{(iv)}} \|\bar{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}_t\|^2 + 2\eta_t u(\bar{\boldsymbol{\theta}}, \boldsymbol{\theta}_{t+1}) \\
&= \|\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_t\|^2 + \|\boldsymbol{\theta}_t - \tilde{\boldsymbol{\theta}}_t\|^2 + 2(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_t)^\top(\boldsymbol{\theta}_t - \tilde{\boldsymbol{\theta}}_t) + 2\eta_t u(\bar{\boldsymbol{\theta}}, \boldsymbol{\theta}_{t+1}) \\
&= \|\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_t\|^2 + \eta_t^2\|\boldsymbol{g}\|^2 + 2\eta_t(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_t)^\top\boldsymbol{g} + 2\eta_t u(\bar{\boldsymbol{\theta}}, \boldsymbol{\theta}_{t+1}) \\
&\leq^{\text{(v)}} \|\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_t\|^2 + \eta_t^2\|\boldsymbol{g}\|^2 + 2\eta_t(L(\bar{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}_t)) + 2\eta_t u(\bar{\boldsymbol{\theta}}, \boldsymbol{\theta}_{t+1}) \\
&\leq \|\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_t\|^2 + \eta_t^2 G^2 + 2\eta_t(L(\bar{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}_t)) + 2\eta_t u(\bar{\boldsymbol{\theta}}, \boldsymbol{\theta}_{t+1}),
\end{aligned} \tag{22}$$

where the inequality (i) is due to the nonexpansiveness of the projection operator, (ii) follows from applying Corollary 7 J times, (iii) follows from applying the inequality $\Omega_j(\tilde{\boldsymbol{\theta}}_{t+l/J}) \geq \Omega_j(\tilde{\boldsymbol{\theta}}_{t+(l+1)/J})$ for $l = j, \ldots, J-1$, (iv) results from the fact that $\Omega(\tilde{\boldsymbol{\theta}}_{t+1}) \geq \Omega(\Pi_\Theta(\tilde{\boldsymbol{\theta}}_{t+1}))$, and (v) results from the subgradient inequality of convex functions, which has an extra term $\frac{\sigma}{2}\|\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_t\|^2$ if L is σ-strongly convex.

C PROOF OF PROPOSITION 3

Invoke Lemma 2 and sum for $t = 1, \ldots, T$, which gives

$$\begin{aligned}
\sum_{t=1}^{T}(L(\boldsymbol{\theta}_t; x_t, y_t) + \lambda\Omega(\boldsymbol{\theta}_t)) &= \sum_{t=1}^{T}(L(\boldsymbol{\theta}_t; x_t, y_t) + \lambda\Omega(\boldsymbol{\theta}_{t+1})) - \lambda(\Omega(\boldsymbol{\theta}_{T+1}) - \Omega(\boldsymbol{\theta}_1)) \\
&\leq^{\text{(i)}} \sum_{t=1}^{T}(L(\boldsymbol{\theta}_t; x_t, y_t) + \lambda\Omega(\boldsymbol{\theta}_{t+1})) \\
&\leq \sum_{t=1}^{T}(L(\boldsymbol{\theta}^*; x_t, y_t) + \lambda\Omega(\boldsymbol{\theta}^*)) + \frac{G^2}{2}\sum_{t=1}^{T}\eta_t + \sum_{t=1}^{T}\frac{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_t\|^2 - \|\boldsymbol{\theta}^* - \boldsymbol{\theta}_{t+1}\|^2}{2\eta_t} \\
&= \sum_{t=1}^{T}(L(\boldsymbol{\theta}^*; x_t, y_t) + \lambda\Omega(\boldsymbol{\theta}^*)) + \frac{G^2}{2}\sum_{t=1}^{T}\eta_t + \frac{1}{2}\sum_{t=2}^{T}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\cdot\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_t\|^2 \\
&\quad + \frac{1}{2\eta_1}\cdot\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_1\|^2 - \frac{1}{2\eta_T}\cdot\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_{T+1}\|^2
\end{aligned} \tag{23}$$

where the inequality (i) is due to the fact that $\boldsymbol{\theta}_1 = \mathbf{0}$. Noting that the third term vanishes for a constant learning rate and that the last term is non-positive suffices to prove the first part. For the second part, we continue as:

$$\begin{aligned}
\sum_{t=1}^{T}(L(\boldsymbol{\theta}_t; x_t, y_t) + \lambda\Omega(\boldsymbol{\theta}_t)) &\leq \sum_{t=1}^{T}(L(\boldsymbol{\theta}^*; x_t, y_t) + \lambda\Omega(\boldsymbol{\theta}^*)) + \frac{G^2}{2}\sum_{t=1}^{T}\eta_t + \frac{F^2}{2}\sum_{t=2}^{T}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + \frac{F^2}{2\eta_1} \\
&= \sum_{t=1}^{T}(L(\boldsymbol{\theta}^*; x_t, y_t) + \lambda\Omega(\boldsymbol{\theta}^*)) + \frac{G^2}{2}\sum_{t=1}^{T}\eta_t + \frac{F^2}{2\eta_T} \\
&\leq^{\text{(ii)}} \sum_{t=1}^{T}(L(\boldsymbol{\theta}^*; x_t, y_t) + \lambda\Omega(\boldsymbol{\theta}^*)) + G^2\eta_0(\sqrt{T} - 1/2) + \frac{F^2\sqrt{T}}{2\eta_0} \\
&\leq \sum_{t=1}^{T}(L(\boldsymbol{\theta}^*; x_t, y_t) + \lambda\Omega(\boldsymbol{\theta}^*)) + \left(G^2\eta_0 + \frac{F^2}{2\eta_0}\right)\sqrt{T},
\end{aligned} \tag{24}$$

where equality (ii) is due to the fact that $\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \leq 2\sqrt{T} - 1$. For the third part, continue after inequality (i) as:

$$\begin{aligned}
\sum_{t=1}^{T}(L(\boldsymbol{\theta}_t; x_t, y_t) + \lambda\Omega(\boldsymbol{\theta}_t)) &\leq \sum_{t=1}^{T}(L(\boldsymbol{\theta}^*; x_t, y_t) + \lambda\Omega(\boldsymbol{\theta}^*)) + \frac{G^2}{2}\sum_{t=1}^{T}\eta_t + \frac{1}{2}\sum_{t=2}^{T}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \sigma\right)\cdot\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_t\|^2 \\
&\quad + \frac{1}{2}\left(\frac{1}{\eta_1} - \sigma\right)\cdot\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_1\|^2 - \frac{1}{2\eta_T}\cdot\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_{T+1}\|^2 \\
&= \sum_{t=1}^{T}(L(\boldsymbol{\theta}^*; x_t, y_t) + \lambda\Omega(\boldsymbol{\theta}^*)) + \frac{G^2}{2\sigma}\sum_{t=1}^{T}\frac{1}{t} - \frac{\sigma T}{2}\cdot\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_{T+1}\|^2 \\
&\leq \sum_{t=1}^{T}(L(\boldsymbol{\theta}^*; x_t, y_t) + \lambda\Omega(\boldsymbol{\theta}^*)) + \frac{G^2}{2\sigma}\sum_{t=1}^{T}\frac{1}{t} \\
&\leq^{\text{(iii)}} \sum_{t=1}^{T}(L(\boldsymbol{\theta}^*; x_t, y_t) + \lambda\Omega(\boldsymbol{\theta}^*)) + \frac{G^2}{2\sigma}(1 + \log T),
\end{aligned} \tag{25}$$

where the equality (iii) is due to the fact that $\sum_{t=1}^{T}\frac{1}{t} \leq 1 + \log T$.

D LIPSCHITZ CONSTANTS FOR SOME LOSS FUNCTIONS

Let $\boldsymbol{\theta}^*$ be a solution of the problem (9) with $\Theta = \mathcal{H}$. For certain loss functions, we may obtain bounds of the form $\|\boldsymbol{\theta}^*\| \leq \gamma$ for some $\gamma > 0$, as the next proposition illustrates. Therefore, we may redefine $\Theta = \{\boldsymbol{\theta} \in \mathcal{H} \mid \|\boldsymbol{\theta}\| \leq \gamma\}$ (a vacuous constraint) without affecting the solution of (9).

Proposition 9 *Let $\Omega(\boldsymbol{\theta}) = \frac{1}{2}(\sum_{m=1}^M \|\boldsymbol{\theta}_m\|)^2$. Let L_{SVM} and L_{CRF} be the structured hinge and logistic losses (4). Assume that the average cost function (in the SVM case) or the average entropy (in the CRF case) are bounded by some $\Lambda \geq 0$, i.e.,*[13]

$$\frac{1}{N}\sum_{i=1}^N \max_{y_i' \in \in \mathcal{Y}(x_t)} c(y_i'; y_i) \leq \Lambda \quad \text{or} \quad \frac{1}{N}\sum_{i=1}^N H(Y_i) \leq \Lambda. \tag{26}$$

Then:

1. *The solution of (9) with $\Theta = \mathcal{H}$ satisfies $\|\boldsymbol{\theta}^*\| \leq \sqrt{2\Lambda/\lambda}$.*
2. *L is G-Lipschitz on $\mathcal{H}$, with $G = 2\max_{u \in \mathcal{U}} \|\boldsymbol{\phi}(u)\|$.*
3. *Consider the following problem obtained from (9) by adding a quadratic term:*

$$\min_{\boldsymbol{\theta}} \frac{\sigma}{2}\|\boldsymbol{\theta}\|^2 + \lambda\Omega(\boldsymbol{\theta}) + \frac{1}{N}\sum_{i=1}^N L(\boldsymbol{\theta}; x_i, y_i). \tag{27}$$

   The solution of this problem satisfies $\|\boldsymbol{\theta}^*\| \leq \sqrt{2\Lambda/(\lambda+\sigma)}$.

4. *The modified loss $\tilde{L} = L + \frac{\sigma}{2}\|.\|^2$ is $\tilde{G}$-Lipschitz on $\left\{\boldsymbol{\theta} \mid \|\boldsymbol{\theta}\| \leq \sqrt{2\Lambda/(\lambda+\sigma)}\right\}$, where $\tilde{G} = G + \sqrt{2\sigma^2\Lambda/(\lambda+\sigma)}$.*

Proof: Let $F_{\mathrm{SVM}}(\boldsymbol{\theta})$ and $F_{\mathrm{CRF}}(\boldsymbol{\theta})$ be the objectives of (9) for the SVM and CRF cases. We have

$$F_{\mathrm{SVM}}(\mathbf{0}) = \lambda\Omega(\mathbf{0}) + \frac{1}{N}\sum_{i=1}^N L_{\mathrm{SVM}}(\mathbf{0}; x_i, y_i) = \frac{1}{N}\sum_{i=1}^N \max_{y_i' \in \mathcal{Y}(x_i)} c(y_i'; y_i) \leq \Lambda_{\mathrm{SVM}} \tag{28}$$

$$F_{\mathrm{CRF}}(\mathbf{0}) = \lambda\Omega(\mathbf{0}) + \frac{1}{N}\sum_{i=1}^N L_{\mathrm{CRF}}(\mathbf{0}; x_i, y_i) = \frac{1}{N}\sum_{i=1}^N \log|\mathcal{Y}(x_i)| \leq \Lambda_{\mathrm{CRF}} \tag{29}$$

Using the facts that $F(\boldsymbol{\theta}^*) \leq F(\mathbf{0})$, that the losses are non-negative, and that $(\sum_i |x_i|)^2 \geq \sum_i x_i^2$, we obtain $\frac{\lambda}{2}\|\boldsymbol{\theta}^*\|^2 \leq \lambda\Omega(\boldsymbol{\theta}^*) \leq F(\boldsymbol{\theta}^*) \leq F(\mathbf{0})$, which proves the first statement.

To prove the second statement for the SVM case, note that a subgradient of L_{SVM} at $\boldsymbol{\theta}$ is $\boldsymbol{g}_{\mathrm{SVM}} = \boldsymbol{\phi}(x, \hat{y}) - \boldsymbol{\phi}(x, y)$, where $\hat{y} = \arg\max_{y' \in \mathcal{Y}(x)} \boldsymbol{\theta}^\top(\boldsymbol{\phi}(x, y') - \boldsymbol{\phi}(x, y)) + c(y'; y)$; and that the gradient of L_{CRF} at $\boldsymbol{\theta}$ is $\boldsymbol{g}_{\mathrm{CRF}} = \mathbb{E}_{\boldsymbol{\theta}}\boldsymbol{\phi}(x, Y) - \boldsymbol{\phi}(x, y)$. Applying Jensen's inequality, we have that $\|\boldsymbol{g}_{\mathrm{CRF}}\| \leq \mathbb{E}_{\boldsymbol{\theta}}\|\boldsymbol{\phi}(x, Y) - \boldsymbol{\phi}(x, y)\|$. Therefore, both $\|\boldsymbol{g}_{\mathrm{SVM}}\|$ and $\|\boldsymbol{g}_{\mathrm{CRF}}\|$ are upper bounded by $\max_{x \in \mathcal{X}, y, y' \in \mathcal{Y}(x)} \|\boldsymbol{\phi}(x, y') - \boldsymbol{\phi}(x, y)\| \leq 2\max_{u \in \mathcal{U}} \|\boldsymbol{\phi}(u)\|$.

The same rationale can be used to prove the third and fourth statements. ■

E COMPUTING THE PROXIMITY OPERATOR OF THE (NON-SEPARABLE) SQUARED ℓ_1

We present an algorithm (Alg. 4) that computes the Moreau projection of the *squared*, *weighted* ℓ_1-norm. Denote by $\odot$ the Hadamard product, $[\mathbf{a} \odot \mathbf{b}]_k = a_k b_k$. Letting $\lambda, \mathbf{d} \geq 0$, and $\phi_{\mathbf{d}}(\mathbf{x}) \triangleq \frac{1}{2}\|\mathbf{d} \odot \mathbf{x}\|_1^2$, the underlying optimization problem is:

$$M_{\lambda\phi_{\mathbf{d}}}(\mathbf{x}_0) \triangleq \min_{\mathbf{x} \in \mathbb{R}^M} \frac{1}{2}\|\mathbf{x} - \mathbf{x}_0\|^2 + \frac{\lambda}{2}\left(\sum_{m=1}^M d_m|x_m|\right)^2. \tag{30}$$

[13] In sequence binary labeling, we have $\Lambda = \bar{P}$ for the CRF case and for the SVM case with a Hamming cost function, where $\bar{P}$ is the average sequence length. Observe that the entropy of a distribution over labelings of a sequence of length P is upper bounded by $\log 2^P = P$.

Algorithm 4 Moreau projection for the squared weighted ℓ_1-norm

Input: A vector $\mathbf{x}_0 \in \mathbb{R}^M$, a weight vector $\mathbf{d} \geq 0$, and a parameter $\lambda > 0$
Set $u_{0m} = |x_{0m}|/d_m$ and $a_m = d_m^2$ for each $m = 1, \ldots, M$
Sort $\mathbf{u}_0$: $u_{0(1)} \geq \ldots \geq u_{0(M)}$
Find $\rho = \max \left\{ j \in \{1, \ldots, M\} \mid u_{0(j)} - \frac{\lambda}{1+\lambda \sum_{r=1}^{j} a_{(r)}} \sum_{r=1}^{j} a_{(r)} u_{0(r)} > 0 \right\}$
Compute $\mathbf{u} = \mathrm{soft}(\mathbf{u}_0, \tau)$, where $\tau = \frac{\lambda}{1+\lambda \sum_{r=1}^{\rho} a_{(r)}} \sum_{r=1}^{\rho} a_{(r)} u_{0(r)}$
Output: $\mathbf{x}$ s.t. $x_r = \mathrm{sign}(x_{0r}) d_r u_r$.

This includes the squared ℓ_1-norm as a particular case, when $\mathbf{d} = \mathbf{1}$ (the case addressed in Alg. 2). The proof is somewhat technical and follows the same procedure employed by Duchi et al. (2008) to derive an algorithm for projecting onto the ℓ_1-ball. The runtime is $O(M \log M)$ (the amount of time that is necessary to sort the vector), but a similar trick as the one described by (Duchi et al., 2008) can be employed to yield $O(M)$ runtime.

Lemma 10 *Let $\mathbf{x}^* = \mathrm{prox}_{\lambda\phi_{\mathbf{d}}}(\mathbf{x}_0)$ be the solution of (30). Then:*

1. *$\mathbf{x}^*$ agrees in sign with $\mathbf{x}_0$,* i.e.*, each component satisfies $x_{0i} \cdot x_i^* \geq 0$.*
2. *Let $\boldsymbol{\sigma} \in \{-1, 1\}^M$. Then $\mathrm{prox}_{\lambda\phi_{\mathbf{d}}}(\boldsymbol{\sigma} \odot \mathbf{x}_0) = \boldsymbol{\sigma} \odot \mathrm{prox}_{\lambda\phi_{\mathbf{d}}}(\mathbf{x}_0)$,* i.e.*, flipping a sign in $\mathbf{x}_0$ produces a $\mathbf{x}^*$ with the same sign flipped.*

Proof: Suppose that $x_{0i} \cdot x_i^* < 0$ for some i. Then, $\mathbf{x}$ defined by $x_j = x_j^*$ for $j \neq i$ and $x_i = -x_i^*$ achieves a lower objective value than $\mathbf{x}^*$, since $\phi_{\mathbf{d}}(\mathbf{x}) = \phi_{\mathbf{d}}(\mathbf{x}^*)$ and $(x_i - x_{0i})^2 < (x_i^* - x_{0i})^2$; this contradicts the optimality of $\mathbf{x}^*$. The second statement is a simple consequence of the first one and that $\phi_{\mathbf{d},\lambda}(\boldsymbol{\sigma} \odot \mathbf{x}) = \phi_{\mathbf{d},\lambda}(\boldsymbol{\sigma} \odot \mathbf{x}^*)$. ∎

Lemma 10 enables reducing the problem to the non-negative orthant, by writing $\mathbf{x}_0 = \boldsymbol{\sigma} \cdot \tilde{\mathbf{x}}_0$, with $\tilde{\mathbf{x}}_0 \geq \mathbf{0}$, obtaining a solution $\tilde{\mathbf{x}}^*$ and then recovering the true solution as $\mathbf{x}^* = \boldsymbol{\sigma} \cdot \tilde{\mathbf{x}}^*$. It therefore suffices to solve (30) with the constraint $\mathbf{x} \geq \mathbf{0}$, which in turn can be transformed into:

$$\min_{\mathbf{u} \geq \mathbf{0}} F(\mathbf{u}) \triangleq \frac{1}{2} \sum_{m=1}^{M} a_m (u_m - u_{0m})^2 + \frac{\lambda}{2} \left(\sum_{m=1}^{M} a_m u_m \right)^2, \tag{31}$$

where we made the change of variables $a_m \triangleq d_m^2$, $u_{0m} \triangleq x_{0m}/d_m$ and $u_m \triangleq x_m/d_m$.

The Lagrangian of (31) is $\mathcal{L}(\mathbf{u}, \boldsymbol{\xi}) = \frac{1}{2} \sum_{m=1}^{M} a_m (u_m - u_{0m})^2 + \frac{\lambda}{2} \left(\sum_{m=1}^{M} a_m u_m \right)^2 - \boldsymbol{\xi}^\top \mathbf{u}$, where $\boldsymbol{\xi} \geq \mathbf{0}$ are Lagrange multipliers. Equating the gradient (w.r.t. $\mathbf{u}$) to zero gives

$$\mathbf{a} \odot (\mathbf{u} - \mathbf{u}_0) + \lambda \sum_{m=1}^{M} a_m u_m \mathbf{a} - \boldsymbol{\xi} = \mathbf{0}. \tag{32}$$

From the complementary slackness condition, $u_j > 0$ implies $\xi_j = 0$, which in turn implies

$$a_j (u_j - u_{0j}) + \lambda a_j \sum_{m=1}^{M} a_m u_m = 0. \tag{33}$$

Thus, if $u_j > 0$, the solution is of the form $u_j = u_{0j} - \tau$, with $\tau = \lambda \sum_{m=1}^{M} a_m u_m$. The next lemma shows the existence of a split point below which some coordinates vanish.

Lemma 11 *Let $\mathbf{u}^*$ be the solution of (31). If $u_k^* = 0$ and $u_{0j} < u_{0k}$, then we must have $u_j^* = 0$.*

Proof: Suppose that $u_j^* = \epsilon > 0$. We will construct a $\tilde{\mathbf{u}}$ whose objective value is lower than $F(\mathbf{u}^*)$, which contradicts the optimality of $\mathbf{u}^*$: set $\tilde{u}_l = u_l^*$ for $l \notin \{j, k\}$, $\tilde{u}_k = \epsilon c$, and $\tilde{u}_j = \epsilon (1 - c a_k / a_j)$, where $c = \min\{a_j / a_k, 1\}$. We have $\sum_{m=1}^{M} a_m u_m^* = \sum_{m=1}^{M} a_m \tilde{u}_m$, and therefore

$$\begin{aligned} 2(F(\tilde{\mathbf{u}}) - F(\mathbf{u}^*)) &= \sum_{m=1}^{M} a_m (\tilde{u}_m - u_{0m})^2 - \sum_{m=1}^{M} a_m (u_m^* - u_{0m})^2 \\ &= a_j (\tilde{u}_j - u_{0j})^2 - a_j (u_j^* - u_{0j})^2 + a_k (\tilde{u}_k - u_{0k})^2 - a_k (u_k^* - u_{0k})^2. \end{aligned} \tag{34}$$

Consider the following two cases: (i) if $a_j \leq a_k$, then $\tilde{u}_k = \epsilon a_j/a_k$ and $\tilde{u}_j = 0$. Substituting in (34), we obtain $2(F(\tilde{\mathbf{u}}) - F(\mathbf{u}^*)) = \epsilon^2 \left(a_j^2/a_k - a_j\right) \leq 0$, which leads to the contradiction $F(\tilde{\mathbf{u}}) \leq F(\mathbf{u}^*)$. If (ii) $a_j > a_k$, then $\tilde{u}_k = \epsilon$ and $\tilde{u}_j = \epsilon\left(1 - a_k/a_j\right)$. Substituting in (34), we obtain $2(F(\tilde{\mathbf{u}}) - F(\mathbf{u}^*)) = a_j\epsilon^2\left(1 - a_k/a_j\right)^2 + 2a_k\epsilon u_{0j} - 2a_k\epsilon u_{0k} + a_k\epsilon^2 - a_j\epsilon^2 < a_k^2/a_j\epsilon^2 - 2a_k\epsilon^2 + a_k\epsilon^2 = \epsilon^2\left(a_k^2/a_j - a_k\right) < 0$, which also leads to a contradiction. ■

Let $u_{0(1)} \geq \ldots \geq u_{0(M)}$ be the entries of $\mathbf{u}_0$ sorted in decreasing order, and let $u^*_{(1)}, \ldots, u^*_{(M)}$ be the entries of $\mathbf{u}^*$ under the same permutation. Let ρ be the number of nonzero entries in $\mathbf{u}^*$, i.e., $u^*_{(\rho)} > 0$, and, if $\rho < M$, $u^*_{(\rho+1)} = 0$. Summing (33) for $(j) = 1, \ldots, \rho$, we get

$$\sum_{r=1}^{\rho} a_{(r)}u^*_{(r)} - \sum_{r=1}^{\rho} a_{(r)}u_{0(r)} + \left(\sum_{r=1}^{\rho} a_{(r)}\right)\lambda\sum_{r=1}^{\rho} a_{(r)}u^*_{(r)} = 0, \tag{35}$$

which implies

$$\sum_{m=1}^{M} u^*_m = \sum_{r=1}^{\rho} u^*_{(r)} = \frac{1}{1+\lambda\sum_{r=1}^{\rho} a_{(r)}}\sum_{r=1}^{\rho} a_{(r)}u_{0(r)}, \tag{36}$$

and therefore $\tau = \frac{\lambda}{1+\lambda\sum_{r=1}^{\rho} a_{(r)}}\sum_{r=1}^{\rho} a_{(r)}u_{0(r)}$. The complementary slackness conditions for $r = \rho$ and $r = \rho+1$ imply

$$u^*_{(\rho)} - u_{0(\rho)} + \lambda\sum_{r=1}^{\rho} a_{(r)}u^*_{(r)} = 0 \quad \text{and} \quad -u^*_{0(\rho+1)} + \lambda\sum_{r=1}^{\rho} a_{(r)}u^*_{(r)} = \xi_{(\rho+1)} \geq 0; \tag{37}$$

therefore $u_{0(\rho)} > u_{0(\rho)} - u^*_{(\rho)} = \tau \geq u_{0(\rho+1)}$. This implies that ρ is such that

$$u_{0(\rho)} > \frac{\lambda}{1+\lambda\sum_{r=1}^{\rho} a_{(r)}}\sum_{r=1}^{\rho} a_{(r)}u_{0(r)} \geq u_{0(\rho+1)}. \tag{38}$$

The next proposition goes farther by exactly determining ρ.

Proposition 12 *The quantity ρ can be determined via:*

$$\rho = \max\left\{ j \in \{1, \ldots, M\} \;\middle|\; u_{0(j)} - \frac{\lambda}{1+\lambda\sum_{r=1}^{j} a_{(r)}}\sum_{r=1}^{j} a_{(r)}u_{0(r)} > 0 \right\}. \tag{39}$$

Proof: Let $\rho^* = \max\{j | u^*_{(j)} > 0\}$. We have that $u^*_{(r)} = u_{0(r)} - \tau^*$ for $r \leq \rho^*$, where $\tau^* = \frac{\lambda}{1+\lambda\sum_{r=1}^{\rho^*} a_{(r)}}\sum_{r=1}^{\rho^*} a_{(r)}u_{0(r)}$, and therefore $\rho \geq \rho^*$. We need to prove that $\rho \leq \rho^*$, which we will do by contradiction. Assume that $\rho > \rho^*$. Let $\mathbf{u}$ be the vector induced by the choice of ρ, i.e., $u_{(r)} = 0$ for $r > \rho$ and $u_{(r)} = u_{0(r)} - \tau$ for $r \leq \rho$, where $\tau = \frac{\lambda}{1+\lambda\sum_{r=1}^{\rho} a_{(r)}}\sum_{r=1}^{\rho} a_{(r)}u_{0(r)}$. From the definition of ρ, we have $u_{(\rho)} = u_{0(\rho)} - \tau > 0$, which implies $u_{(r)} = u_{0(r)} - \tau > 0$ for each $r \leq \rho$. In addition,

$$\begin{aligned}
\sum_{r=1}^{M} a_r u_r &= \sum_{r=1}^{\rho} a_{(r)}u_{0(r)} - \sum_{r=1}^{\rho} a_{(r)}\tau = \left(1 - \frac{\lambda\sum_{r=1}^{\rho} a_{(r)}}{1+\lambda\sum_{r=1}^{\rho} a_{(r)}}\right)\sum_{r=1}^{\rho} a_{(r)}u_{0(r)} \\
&= \frac{1}{1+\lambda\sum_{r=1}^{\rho} a_{(r)}}\sum_{r=1}^{\rho} a_{(r)}u_{0(r)} = \frac{\tau}{\lambda}, \qquad (40)\\
\sum_{r=1}^{M} a_r(u_r - u_{0r})^2 &= \sum_{r=1}^{\rho^*} a_{(r)}\tau^2 + \sum_{r=\rho^*+1}^{\rho} a_{(r)}\tau^2 + \sum_{r=\rho+1}^{M} a_{(r)}u^2_{0(r)} \\
&< \sum_{r=1}^{\rho^*} a_{(r)}\tau^2 + \sum_{r=\rho^*+1}^{M} a_{(r)}u^2_{0(r)}. \qquad (41)
\end{aligned}$$

We next consider two cases:

1. $\boxed{\tau^* \geq \tau.}$ From (41), we have that $\sum_{r=1}^M a_r(u_r - u_{0r})^2 < \sum_{r=1}^{\rho^*} a_{(r)}\tau^2 + \sum_{r=\rho^*+1}^M a_{(r)}u_{0(r)}^2 \leq \sum_{r=1}^{\rho^*} a_{(r)}(\tau^*)^2 + \sum_{r=\rho^*+1}^M a_{(r)}u_{0(r)}^2 = \sum_{r=1}^M a_r(u_r^* - u_{0r})^2$. From (40), we have that $\left(\sum_{r=1}^M a_r u_r\right)^2 = \tau^2/\lambda^2 \leq (\tau^*)^2/\lambda^2$. Summing the two inequalities, we get $F(\mathbf{u}) < F(\mathbf{u}^*)$, which leads to a contradiction.

2. $\boxed{\tau^* < \tau.}$ We will construct a vector $\tilde{\mathbf{u}}$ from $\mathbf{u}^*$ and show that $F(\tilde{\mathbf{u}}) < F(\mathbf{u}^*)$. Define

$$\tilde{u}_{(r)} = \begin{cases} u^*_{(\rho^*)} - \frac{2a_{(\rho^*+1)}}{a_{(\rho^*)}+a_{(\rho^*+1)}}\epsilon, & \text{if } r = \rho^* \\ \frac{2a_{(\rho^*)}}{a_{(\rho^*)}+a_{(\rho^*+1)}}\epsilon, & \text{if } r = \rho^*+1 \\ u^*_{(r)} & \text{otherwise,} \end{cases} \tag{42}$$

where $\epsilon = (u_{0(\rho^*+1)} - \tau^*)/2$. Note that $\sum_{r=1}^M a_r\tilde{u}_r = \sum_{r=1}^M a_r u_r^*$. From the assumptions that $\tau^* < \tau$ and $\rho^* < \rho$, we have that $u^*_{(\rho^*+1)} = u_{0(\rho^*+1)} - \tau > 0$, which implies that $\tilde{u}_{(\rho^*+1)} = \frac{a_{(\rho^*)}(u_{0(\rho^*+1)} - \tau^*)}{a_{(\rho^*)}+a_{(\rho^*+1)}} > \frac{a_{(\rho^*)}(u_{0(\rho^*+1)} - \tau)}{a_{(\rho^*)}+a_{(\rho^*+1)}} = \frac{a_{(\rho^*)}u^*_{(\rho^*+1)}}{a_{(\rho^*)}+a_{(\rho^*+1)}} > 0$, and that $u^*_{(\rho^*)} = u_{0(\rho^*)} - \tau^* - \frac{a_{(\rho^*+1)}(u_{0(\rho^*+1)} - \tau^*)}{a_{(\rho^*)}+a_{(\rho^*+1)}} = u_{0(\rho^*)} - \frac{a_{(\rho^*+1)}u_{0(\rho^*+1)}}{a_{(\rho^*)}+a_{(\rho^*+1)}} - \left(1 - \frac{a_{(\rho^*+1)}}{a_{(\rho^*)}+a_{(\rho^*+1)}}\right)\tau^* >^{(\mathrm{i})} \left(1 - \frac{a_{(\rho^*+1)}}{a_{(\rho^*)}+a_{(\rho^*+1)}}\right)(u_{0(\rho^*+1)} - \tau) = \left(1 - \frac{a_{(\rho^*+1)}}{a_{(\rho^*)}+a_{(\rho^*+1)}}\right)(u^*_{(\rho^*+1)}) > 0$, where inequality (i) is justified by the facts that $u_{0(\rho^*)} \geq u_{0(\rho^*+1)}$ and $\tau > \tau^*$. This ensures that $\tilde{\mathbf{u}}$ is well defined. We have:

$$\begin{aligned} 2(F(\mathbf{u}^*) - F(\tilde{\mathbf{u}})) &= \sum_{r=1}^M a_r(u_r^* - u_{0r})^2 - \sum_{r=1}^M a_r(\tilde{u}_r - u_{0r})^2 \\ &= a_{(\rho^*)}(\tau^*)^2 + a_{(\rho^*+1)}u_{0(\rho^*+1)}^2 - a_{(\rho^*)}\left(\tau^* + \frac{2a_{(\rho^*+1)}\epsilon}{a_{(\rho^*)}+a_{(\rho^*+1)}}\right)^2 \\ &\quad -a_{(\rho^*+1)}\left(u_{0(\rho^*+1)} - \frac{2a_{(\rho^*)}\epsilon}{a_{(\rho^*)}+a_{(\rho^*+1)}}\right)^2 \\ &= -\frac{4a_{(\rho^*)}a_{(\rho^*+1)}\epsilon}{a_{(\rho^*)}+a_{(\rho^*+1)}}\underbrace{(\tau^* - u_{0(\rho^*+1)})}_{-2\epsilon} - \frac{4a_{(\rho^*)}a_{(\rho^*+1)}^2\epsilon^2}{\left(a_{(\rho^*)}+a_{(\rho^*+1)}\right)^2} - \frac{4a_{(\rho^*)}^2a_{(\rho^*+1)}\epsilon^2}{\left(a_{(\rho^*)}+a_{(\rho^*+1)}\right)^2} \\ &= \frac{4a_{(\rho^*)}a_{(\rho^*+1)}\epsilon^2}{a_{(\rho^*)}+a_{(\rho^*+1)}} \geq 0, \end{aligned} \tag{43}$$

which leads to a contradiction and completes the proof.

■