Loss Max-Pooling for Semantic Image Segmentation
Supplementary Material

Samuel Rota Bulò[⋆,†] Gerhard Neuhold[†] Peter Kontschieder[†]

[⋆]FBK - Trento, Italy - rotabulo@fbk.eu

[†]Mapillary - Graz, Austria - {samuel,gerhard,pkontschieder}@mapillary.com

Abstract

This document provides the following, additional contributions to our CVPR 2017 submission:

- *in Section A we provide the proofs of Theorem 1 and other results referenced from the main paper;*
- *in Section B we provide the derivation of the gradient.*

A. Proofs and Auxiliary Results

Proof of Thm. 1. Let $\alpha^\star = \max\{\alpha \in \mathbb{R} : \eta(\alpha) = 0\}$ and let $\mathcal{J}^\star = \{u \in \mathcal{I} : \eta(\ell_{\hat{y}y}(u)) > \alpha^\star\}$. We start proving that $\mathcal{J}^\star = \mathcal{J}^*$. If $u \in \mathcal{J}^*$, then by definition of $\mathcal{J}^*$ we have $\eta(\ell_{\hat{y}y}(u)) > 0$, which implies $\ell_{\hat{y}y}(u) > \alpha^\star$ by Proposition 8 (take $\alpha_1 = \ell_{\hat{y}y}(u)$ and $\alpha_2 = \alpha^\star$). Consequently, $u \in \mathcal{J}^\star$ and, therefore, $\mathcal{J}^* \subseteq \mathcal{J}^\star$. If $u \in \mathcal{J}^\star$, then by definition of $\mathcal{J}^\star$ we have $\eta(\ell_{\hat{y}y}(u)) > \alpha^\star$, which implies $\eta(\ell_{\hat{y}y}(u)) > 0$ since $\alpha^\star \geq 0$ hold by Proposition 5. Therefore, $u \in \mathcal{J}^*$. So, $\mathcal{J}^\star \subseteq \mathcal{J}^*$. We have thus proved that $\mathcal{J}^\star = \mathcal{J}^*$.

We obtain α^*, as given in the theorem, by solving equation $\eta(\alpha^\star) = 0$ for variable $\alpha^\star$, after having replaced $\mathcal{J}^\star$ with the equivalent $\mathcal{J}^*$. The equation admits a unique solution, because $|\mathcal{J}^*| = |\mathcal{J}^\star| < m$ by Proposition 5. Accordingly, $\alpha^* = \alpha^\star$. Then, by Proposition 1 we have that $\lambda^* = |\ell_{\hat{y}y} - \alpha^*|_+$ is a solution to (6), from which we derive

$$\begin{aligned} L_{\mathcal{W}}(\hat{y}, y) = g(\lambda^*) &= \tau\langle\lambda^*\rangle + \gamma\|\ell_{\hat{y}y} - \lambda^*\|_q = \tau\langle\lambda^*\rangle + \gamma(\langle\ell^q_{\hat{y}y}\rangle_{\overline{\mathcal{J}}^*} + \alpha^{*q}|\mathcal{J}^*|)^{1/q} \\ &= \tau\langle\lambda^*\rangle + \gamma(m\alpha^{*q} - \underbrace{\eta(\alpha^*)}_{=0})^{1/q} = \tau(\langle\ell_{\hat{y}y}\rangle_{\mathcal{J}^*} - |\mathcal{J}^*|\alpha^*) + \underbrace{\gamma m^{1/q}}_{=\tau m}\alpha^* = \tau\left[\langle\ell_{\hat{y}y}\rangle_{\mathcal{J}^*} + (m - |\mathcal{J}^*|)\alpha^*\right] . \end{aligned}$$

As for w^*, we have that $L_{\mathcal{W}}(\hat{y}, y) \geq w^* \cdot \ell_{\hat{y},y}$ holds in general. Now, if $\alpha^* = 0$, then $L_{\mathcal{W}}(\hat{y}, y) \geq w^* \cdot \ell_{\hat{y}y} = \tau\langle\ell_{\hat{y}y}\rangle_{\mathcal{J}^*} = L_{\mathcal{W}}(\hat{y}, y)$. If $\alpha^* > 0$, then

$$L_{\mathcal{W}}(\hat{y}, y) \geq w^* \cdot \ell_{\hat{y}y} = \tau\langle\ell_{\hat{y}y}\rangle_{\mathcal{J}^*} + \frac{\tau}{(\alpha^*)^{q-1}}\left\langle\ell^q_{\hat{y}y}\right\rangle_{\overline{\mathcal{J}}^*} = \tau\left[\langle\ell_{\hat{y}y}\rangle_{\mathcal{J}^*} + (m - |\mathcal{J}^*|)\alpha^*\right] = L_{\mathcal{W}}(\hat{y}, y) ,$$

where the last equality follows from the observation that $(m - |\mathcal{J}^*|)\alpha^{*q} = \langle\ell^q_{\hat{y}y}\rangle_{\overline{\mathcal{J}}^*}$, by definition of α^*. Hence, w^* is an optimal solution to the maximization in (4). □

Proposition 1. *Let $1 \leq q < \infty$ and $1 \leq m \leq n$. If λ^* is a solution to (6), then $\alpha^* = m^{-1/q}\|\ell_{\hat{y}y} - \lambda^*\|_q$ is a root of η. If $\alpha^* = \max\{\alpha \in \mathbb{R} : \eta(\alpha) = 0\}$, then $\lambda^* = |\ell_{\hat{y}y} - \alpha^*|_+$ is a solution to (6).*

Proof. Let λ^* be a solution to (6). It follows from Propositions 3 and 2 that $\alpha^* = m^{-1/q}\|\ell_{\hat{y}y} - \lambda^*\|_q$ is a root of η.

Let $\alpha^* = \max\{\alpha \in \mathbb{R} : \eta(\alpha) = 0\}$ and $\lambda^* = |\ell_{\hat{y}y} - \alpha^*|_+$. Then

$$m^{-1/q}\|\ell_{\hat{y}y} - \lambda^*\|_q = m^{-1/q}(\langle\ell^q_{\hat{y}y}\rangle_{\overline{\mathcal{J}}^*} + \alpha^{*q}|\mathcal{J}^*|)^{1/q} = m^{-1/q}(m\alpha^{*q} - \underbrace{\eta(\alpha^*)}_{=0})^{1/q} = \alpha^* .$$

Now, let $\lambda^\star \in \arg\max\{\|\ell_{\hat{y}y} - \lambda\|_q \,:\, \lambda \text{ satisfies (8)}\}$ and let $\alpha^\star = m^{-1/q}\|\ell_{\hat{y}y} - \lambda^\star\|_q$. By Proposition 2 we have that λ^* satisfies (8), hence

$$\alpha^* = m^{-1/q}\|\ell_{\hat{y}y} - \lambda^*\|_q \leq m^{-1/q}\|\ell_{\hat{y}y} - \lambda^\star\|_q = \alpha^\star\,.$$

By Proposition 2, $\alpha^\star$ is a root of η. However, this implies $\alpha^* = \alpha^\star$ by definition of α^* and, therefore, $\lambda^* = \lambda^\star$. Finally, by Proposition 3 we conclude that λ^* is a solution to (6). □

Theorem 2. *Let $p = 1$ and $1 \leq m \leq n$. Then*

$$L_{\mathcal{W}}(\hat{y}, y) = \tau[\langle \ell_{\hat{y}y}\rangle_{\mathcal{J}^*} + (m - |\mathcal{J}^*|)\alpha^*]\,, \tag{11}$$

where $\mathcal{J}^* \in \arg\max\{\langle \ell_{\hat{y}y}\rangle_{\mathcal{J}} \,:\, \mathcal{J} \subseteq \mathcal{I}, |\mathcal{J}| = \lfloor m \rfloor\}$, i.e. *$\mathcal{J}^*$ contains the pixels with the $\lfloor m \rfloor$ highest losses, and $\alpha^* = \|\ell_{\hat{y}y}\|_{\infty,\overline{\mathcal{J}}^*}$,* i.e. *it corresponds to the highest loss in $\overline{\mathcal{J}}^*$ or zero if $\overline{\mathcal{J}}^*$ is empty. Moreover,*

$$w^*(u) = \begin{cases} \tau & \text{if } u \in \mathcal{J}^* \\ \tau(m - \lfloor m \rfloor)\mu(u) & \text{if } u \in \mathcal{J}^+ \\ 0 & \text{otherwise} \end{cases}$$

is an optimal solution for the maximization in (4), for any probability distribution μ defined over $\mathcal{J}^+ = \{u \in \mathcal{I} : \ell_{\hat{y}y}(u) = \alpha^*\} \setminus \mathcal{J}^*$.

Proof. Assume $w^\star$ to be the maximizer in (4), *i.e.* $L_{\mathcal{W}}(\hat{y}, y) = w^\star \cdot \ell_{\hat{y}y}$. Then it has to be nonnegative and should satisfy $\|w^\star\|_1 = \gamma$. Otherwise, we could construct $w^\dagger = \gamma|w^\star|/\|w^\star\|_1$, which would satisfy $w^\dagger \cdot \ell_{\hat{y}y} > w^\star \cdot \ell_{\hat{y}y}$ contradicting $L_{\mathcal{W}}(\hat{y}, y) = w^\star \cdot \ell_{\hat{y}y}$.

Now, from (4) and the definition of w^* we can derive

$$L_{\mathcal{W}}(\hat{y}, y) \geq w^* \cdot \ell_{\hat{y}y} = \tau[\langle \ell_{\hat{y}y}\rangle_{\mathcal{J}^*} + (m - \underbrace{\lfloor m \rfloor}_{=|\mathcal{J}^*|})\underbrace{\langle \mu \ell_{\hat{y}y}\rangle_{\mathcal{J}^+}}_{=\alpha^*}].$$

Assume by contradiction that strict inequality holds, or in other terms that $w^\star \cdot \ell_{\hat{y}y} > w^* \cdot \ell_{\hat{y}y}$. Let $\mathcal{A}^+ = \{u \in \mathcal{I} : w^\star(u) > w^*(u)\}$ and $\mathcal{A}^- = \{u \in \mathcal{I} : w^\star(u) < w^*(u)\}$. Note that $\mathcal{A}^+ \neq \emptyset$ and $w^\star \neq w^*$, otherwise $w^\star \cdot \ell_{\hat{y}y} \leq w^* \cdot \ell_{\hat{y}y}$ holds yielding a contradiction, and $\mathcal{A}^+ \subseteq \overline{\mathcal{J}}^*$, because $w^\star$ is upper bounded by τ and $w^*(u) = \tau$ for $u \in \mathcal{J}^*$. It follows by definition of α^* that $\ell_{\hat{y}y}(u) \leq \alpha^*$ for any $u \in \mathcal{A}^+$. Additionally, $\mathcal{A}^- \neq \emptyset$, otherwise $\|w^\star\|_1 > \|w^*\|_1 = \gamma$ holds contradicting $w^\star \in \mathcal{W}$, and necessarily $\mathcal{A}^- \subseteq \mathcal{J}^* \cup \mathcal{J}^+$ because $w^\star$ is lower bounded by 0 and $w^*(u) = 0$ for $u \notin \mathcal{A}^-$. It follows by definition of $\mathcal{J}^*$ and $\mathcal{J}^+$ that $\ell_{\hat{y}y}(u) \geq \alpha^*$ for any $v \in \mathcal{A}^-$. But then

$$(w^\star - w^*) \cdot \ell_{\hat{y}y} = \underbrace{\langle (w^\star - w^*)\ell_{\hat{y}y}\rangle_{\mathcal{A}^+}}_{\leq \alpha^* \langle w^\star - w^*\rangle_{\mathcal{A}^+}} + \underbrace{\langle (w^\star - w^*)\ell_{\hat{y}y}\rangle_{\mathcal{A}^-}}_{\leq \alpha^* \langle w^\star - w^*\rangle_{\mathcal{A}^-}} \leq \alpha^* \langle w^\star - w^*\rangle = \alpha^*(\underbrace{\langle w^\star\rangle}_{=\gamma} - \underbrace{\langle w^*\rangle}_{=\gamma}) = 0\,,$$

yielding a contradiction. Hence, $L_{\mathcal{W}}(\hat{y}, y) = w^* \cdot \ell_{\hat{y}y}$, *i.e.* w^* is an optimal solution for the maximization in (4), and (11) holds. □

Proposition 2. *Let $1 \leq q < \infty$ and $1 \leq m \leq n$. If λ satisfies (8), then $\alpha = m^{-1/q}\|\ell_{\hat{y}y} - \lambda\|_q$ is a root of η. If α is a root of η, then $\lambda = |\ell_{\hat{y}y} - \alpha|_+$ satisfies (8).*

Proof. Let $\alpha = m^{-1/q}\|\ell_{\hat{y}y} - \lambda\|_q$. If λ satisfies (8), then $\lambda = |\ell_{\hat{y}y} - \alpha|_+$. By substituting it back into α we obtain:

$$\begin{aligned} \alpha^q &= m^{-1}\|\ell_{\hat{y}y} - \lambda\|_q^q \\ m\alpha^q &= \left[|\mathcal{J}_\alpha|\alpha^q + \langle \ell_{\hat{y}y}^q\rangle_{\overline{\mathcal{J}}_\alpha}\right] \\ 0 &= (m - |\mathcal{J}_\alpha|)\alpha^q - \langle \ell_{\hat{y}y}^q\rangle_{\overline{\mathcal{J}}_\alpha} = \eta(\alpha)\,. \end{aligned}$$

Hence, α is a root of η.

Let α be a root of η and let $\lambda = |\ell_{\hat{y}y} - \alpha|_+$. By following the previous relation bottom-up, we obtain $\alpha = m^{-1/q}\|\ell_{\hat{y}y} - \lambda\|_q$, and by substituting it back into λ we obtain (8). □

Proposition 3. *Let $1 \leq q < \infty$ and $1 \leq m \leq n$. If λ^* is a solution to (6), then it satisfies (8). If $\lambda^* \in \arg\max\{\|\ell_{\hat{y}y} - \lambda\|_q : \lambda$ satisfies (8)$\}$, then it is a solution to (6).*

Proof. Let λ^* be a solution to (6). If $\lambda^* = \ell_{\hat{y}y}$, then (8) is trivially satisfied. Otherwise, it is satisfied by Proposition 4.

Let $\lambda^* \in \arg\max\{\|\ell_{\hat{y}y} - \lambda\|_q : \lambda$ satisfies (8)$\}$. If $\lambda^* \neq \ell_{\hat{y}y}$, then λ^* is a solution to (6) by Proposition 4. If $\lambda^* = \ell_{\hat{y}y}$, then it is the only point satisfying (8). It follows from Proposition 4 that no solution to (6) exists where g is differentiable. However, at least one solution has to exist because the minimization problem in (6) admits a finite solution. So, it has to be a point where g is non-differentiable, but the only one is $\lambda^* = \ell_{\hat{y}y}$. Therefore, λ^* is a solution to (6). □

Proposition 4. *Let $1 \leq q < \infty$, $1 \leq m \leq n$ and $\lambda^* \neq \ell_{\hat{y}y}$. Then λ^* is a solution to (6) if and only if it satisfies (8).*

Proof. ($\Rightarrow$) If λ^* is a solution to (6) and $\lambda^* \neq \ell_{\hat{y}y}$, then λ^* is a point where g is differentiable and the Karush-Kuhn-Tucker (KKT) necessary conditions [1] for optimality are satisfied. Specifically, there exists $\nu \succeq 0$ satisfying

$$\tau - \gamma \left(\frac{\ell_{\hat{y}y} - \lambda^*}{\|\ell_{\hat{y}y} - \lambda^*\|_q} \right)^{q-1} - \nu = 0\,, \qquad \nu \cdot \lambda^* = 0\,.$$

The complementarity constraint and the nonnegativity of ν imply that $\nu(u) = 0$ if $\lambda^*(u) > 0$. By using this fact, we can derive after simple algebraic manipulations the following equivalent relation, which holds for all $u \in \mathcal{I}$:

$$\lambda^*(u) = \begin{cases} \ell_{\hat{y}y}(u) - m^{-1/q}\|\ell_{\hat{y}y} - \lambda^*\|_q & \text{if } \lambda^*(u) > 0 \\ 0 & \text{otherwise,} \end{cases}$$

and this corresponds to

$$\lambda^* = |\ell_{\hat{y}y} - m^{-1/q}\|\ell_{\hat{y}y} - \lambda^*\|_q|_+\,.$$

($\Leftarrow$) By following the derivation above in reversed order, we have that if $\lambda^* \neq \ell_{\hat{y}y}$ satisfies (8) then the KKT conditions for optimality are satisfied. Since g is convex, those conditions are also sufficient [1] and, therefore, λ^* is a solution to (6). □

Proposition 5. *Let $1 \leq q < \infty$ and $1 \leq m \leq n$. If $\alpha^\star = \max\{\alpha \in \mathbb{R} : \eta(\alpha) = 0\}$, then $|\mathcal{J}^\star| < m$ and $\alpha^\star \geq 0$, where $\mathcal{J}^\star = \{u \in \mathcal{I} : \eta(\ell_{\hat{y}y}(u)) > \alpha^\star\}$.*

Proof. Let λ^* be a solution to (6). By Proposition 1 we have that $\alpha^* = m^{-1/q}\|\ell_{\hat{y}y} - \lambda^*\|_q$ is a root of η. If $\lambda^* = \ell_{\hat{y}y}$ we have $\alpha^* = 0$ and by Proposition 6 we have that $\ell_{\hat{y}y}$ has at most $\lfloor m \rfloor$ positive elements. It follows that $|\mathcal{J}^*| \leq m$. If $\lambda^* \neq \ell_{\hat{y}y}$, we have $\alpha^* > 0$. This and $\eta(\alpha^*) = 0$ imply $|\mathcal{J}^*| \leq m$. Accordingly, $|\mathcal{J}^*| \leq m$ always holds and since $\alpha^\star \geq \alpha^* \geq 0$, we have $|\mathcal{J}^\star| \leq |\mathcal{J}^*| \leq m$ and $\alpha^\star \geq 0$.

Finally, assume by contradiction that $|\mathcal{J}^\star| = m$ holds. Then $\eta(\alpha^\star) = 0$ implies $\langle \ell^q_{\hat{y}y} \rangle_{\overline{\mathcal{J}}^\star} = 0$. Take $\underline{\alpha} = \min\{\ell_{\hat{y}y}(u) : u \in \mathcal{J}^\star\}$, then $\mathcal{J}_{\underline{\alpha}} \subset \mathcal{J}^\star$, where $\mathcal{J}_{\underline{\alpha}} = \{u \in \mathcal{I} : \eta(\ell_{\hat{y}y}(u)) > \underline{\alpha}\}$. Let $\Delta = \mathcal{J}^\star \setminus \mathcal{J}_{\underline{\alpha}}$, which is necessarily non-empty and contains only pixels with loss $\underline{\alpha}$. Then

$$\eta(\underline{\alpha}) = (m - |\mathcal{J}_{\underline{\alpha}}|)\underline{\alpha}^q - \langle \ell^q_{\hat{y}y} \rangle_{\overline{\mathcal{J}}_{\underline{\alpha}}} = (\underbrace{m - |\mathcal{J}^\star|}_{=0} + |\Delta|)\underline{\alpha}^q - \underbrace{\langle \ell^q_{\hat{y}y} \rangle_{\overline{\mathcal{J}}^\star}}_{=0} - \langle \ell^q_{\hat{y}y} \rangle_{\Delta} = |\Delta|\underline{\alpha}^q - \langle \ell^q_{\hat{y}y} \rangle_{\Delta} = 0\,,$$

where the last equality follows from the fact that $\ell_{\hat{y}y}(u) = \underline{\alpha}$ for all $u \in \Delta$. However, $\eta(\underline{\alpha}) = 0$ implies $\alpha^\star \geq \underline{\alpha}$ by definition of $\alpha^\star$, which yields a contradiction because $\underline{\alpha} > \alpha^\star$ follows from the definition of $\underline{\alpha}$. Hence, $|\mathcal{J}^\star| < m$. □

Proposition 6. *Let $p > 1$ and $1 \leq m \leq n$. If $\lambda^* = \ell_{\hat{y}y}$ is a minimizer of (6), then $\ell_{\hat{y}y}$ has at most $\lfloor m \rfloor$ positive elements.*

Proof. If $p = \infty$ then $m = n$ and the result is trivially true. Otherwise ($1 < p < \infty$), assume by contradiction that there exist at least $\lfloor m \rfloor + 1$ positive elements in $\ell_{\hat{y}y}$, say with indices in $\mathcal{J}$. Then, the dual objective yields $g(\lambda^*) = \tau\langle \ell_{\hat{y}y} \rangle_{\mathcal{J}}$. However, this value cannot be attained by the primal formulation because there exist at most $\lfloor m \rfloor$ elements in $\mathcal{J}$ with weight τ, while the remaining element would have a weight not exceeding $\tau(m - \lfloor m \rfloor)^{1/p} < \tau$ (due to the constraint $\|w\|_p \leq \gamma$). This implies $L_{\mathcal{W}}(\hat{y}, y) < g(\lambda^*)$, which contradicts strong duality, implied by the satisfaction of the Slater's condition [1]. □

Proposition 7. *Let $1 \leq q < \infty$ and $1 \leq m \leq n$. If $0 \leq \alpha_1 < \alpha_2$, $|\mathcal{J}_1| < m$ and $\eta(\alpha_1) \geq 0$, then $\eta(\alpha_2) > 0$, where $J_1 = \{u \in \mathcal{I} : \ell_{\hat{y}y}(u) > \alpha_1\}$.*

Proof. Let $J_2 = \{u \in \mathcal{I} : \ell_{\hat{y}y}(u) > \alpha_2\}$. The hypothesis $\alpha_1 < \alpha_2$ implies that $\mathcal{J}_2 \subseteq \mathcal{J}_1$. Hence, we can write

$$\eta(\alpha_2) = (m - |\mathcal{J}_1|)\alpha_2^q + |\Delta|\alpha_2^q - \langle \ell_{\hat{y}y}^q \rangle_{\overline{\mathcal{J}}_1} - \langle \ell_{\hat{y}y}^q \rangle_{\Delta} = \underbrace{\eta(\alpha_1)}_{\geq 0} + \underbrace{(m - |\mathcal{J}_1|)\delta}_{a} + \underbrace{|\Delta|\alpha_2^q - \langle \ell_{\hat{y}y}^q \rangle_{\Delta}}_{b},$$

where $\delta = \alpha_2^q - \alpha_1^q$ and $\Delta = \mathcal{J}_1 \setminus \mathcal{J}_2$. Expression a is positive, because $\delta > 0$ and $|\mathcal{J}_1| < m$. Moreover, since Δ has no element in $\mathcal{J}_2$, we have by definition of $\mathcal{J}_2$ that $\ell_{\hat{y}y}(u) \leq \alpha_2$ for all $u \in \Delta$ and, therefore, b is nonnegative. It follows that $\eta(\alpha_2) > 0$. □

Proposition 8. *Let $1 \leq q < \infty$ and $1 \leq m \leq n$. If $\alpha_1 \geq 0$, $\eta(\alpha_1) > 0$ and $\eta(\alpha_2) \leq 0$, then $\alpha_2 < \alpha_1$.*

Proof. By Proposition 9 we have $|\mathcal{J}_1| < m$, where $J_1 = \{u \in \mathcal{I} : \ell_{\hat{y}y}(u) > \alpha_1\}$. The result then follows from the contrapositive of Proposition 7. □

Proposition 9. *Let $1 \leq q < \infty$ and $1 \leq m \leq n$. If $\alpha \geq 0$ and $\eta(\alpha) > 0$, then $|\mathcal{J}_\alpha| < m$, where $J_\alpha = \{u \in \mathcal{I} : \ell_{\hat{y}y}(u) > \alpha\}$.*

Proof. If $\eta(\alpha) > 0$, then $(m - |\mathcal{J}_\alpha|)\alpha^q$ must be positive. But this is the case only if $|\mathcal{J}_\alpha| < m$, since α is nonnegative. □

Proposition 10. *Let $1 \leq q < \infty$, let π be a bijective function $\pi \in \mathcal{I}^{\{1,\ldots,n\}}$ satisfying $\ell_{\hat{y}y}(\pi_i) \leq \ell_{\hat{y}y}(\pi_j)$ if $i < j$, and let*

$$\tau = \arg\min \quad \{i \in \{1, \ldots, n\} : \eta_i > 0\} \cup \{n + 1\},$$

where $\eta_i = (m - n + i)\ell_{\hat{y}y}^q(\pi_i) - \sum_{j=1}^{i} \ell_{\hat{y}y}^q(\pi_j)$. Then

$$\{\pi_j \,:\, \tau \leq j \leq n\} = \{u \in \mathcal{I} : \eta(\ell_{\hat{y}y}(u)) > 0\}.$$

Proof. Take $i \in \{1, \ldots, n\}$, let $\mathcal{J}_i = \{u \in \mathcal{I} : \eta(\ell_{\hat{y}y}(u)) > \ell_{\hat{y}y}(\pi_i)\}$ and let $u = \max\{j \in \{1, \ldots, n\} : \ell_{\hat{y}y}(\pi_j) = \ell_{\hat{y}y}(\pi_i)\}$. Then

$$\begin{aligned} \eta(\ell_{\hat{y}y}(\pi_i)) = (m - |\mathcal{J}_i|)\ell_{\hat{y}y}^q(\pi_i) - \langle \ell_{\hat{y}y}^q \rangle_{\overline{\mathcal{J}}_i} &= (m - n + u)\ell_{\hat{y}y}^q(\pi_i) - \sum_{j=1}^{u} \ell_{\hat{y}y}^q(\pi_j) \\ &\overset{(*)}{=} (m - n + i)\ell_{\hat{y}y}^q(\pi_i) - \sum_{j=1}^{i} \ell_{\hat{y}y}^q(\pi_j) = \eta_i, \end{aligned} \tag{12}$$

holds, where in $(*)$ we used the fact that $\ell_{\hat{y}y}^q(\pi_j)$ is constant for $i \leq j \leq u$. It follows that $\eta(\ell_{\hat{y}y}(\pi_i)) \leq 0$ for all $1 \leq i < \tau$. Now, if $\tau = n + 1$ then the theorem trivially holds, for sets $\{\pi_j \,:\, \tau \leq j \leq n\}$ and $\{u \in \mathcal{I} : \eta(\ell_{\hat{y}y}(u)) > 0\}$ are empty. If $\tau \leq n$ then $\eta_\tau > 0$ by definition of τ and, hence, $\eta(\ell_{\hat{y}y}(\pi_\tau)) > 0$ by (12). Consequently, by Proposition 9 and Proposition 7 we have that $\eta(\ell_{\hat{y}y}(\pi_j)) > 0$ for all $\tau \leq j \leq n$. □

B. Derivation of Gradient

As discussed in the main paper, the gradient $\frac{\partial L_{\mathcal{W}}}{\partial \hat{y}}(\hat{y}, y)$ exists almost everywhere. For the points where it exists, the gradient takes the form:

$$\frac{\partial L_{\mathcal{W}}}{\partial \hat{y}}(\hat{y}, y) = w^* \cdot \frac{\partial \ell_{\hat{y}y}}{\partial \hat{y}} + \frac{\partial w^*}{\partial \hat{y}} \cdot \ell_{\hat{y}y}.$$

In general we consider gradients in directions that leave $\mathcal{J}^*$ and $\mathcal{J}^+$ unchanged. Under this assumption we have that $\frac{\partial w^*}{\partial \hat{y}} \cdot \ell_{\hat{y}y} = 0$ so that $\frac{\partial L_{\mathcal{W}}}{\partial \hat{y}}(\hat{y}, y) = w^* \cdot \frac{\partial \ell_{\hat{y}y}}{\partial \hat{y}}$.

Indeed, $\frac{\partial w^*}{\partial \hat{y}} = 0$ holds for the case $p = 1$. For the case $p > 1$, note that the optimal w^* always satisfies $\|w^*\|_p = \gamma$, which implies that $\frac{\partial w^*}{\partial \hat{y}} \cdot w^{*(p-1)} = 0$ has to be satisfied, indeed

$$\begin{aligned} \frac{\partial}{\partial \hat{y}} \|w^*\|_p &= \frac{\partial}{\partial \hat{y}} \gamma \\ \frac{\partial w^*}{\partial \hat{y}} \cdot \left(\frac{w^*}{\|w^*\|_p} \right)^{p-1} &= 0 \\ \frac{\partial w^*}{\partial \hat{y}} \cdot \left(\frac{w^*}{\gamma} \right)^{p-1} &= 0 \\ \frac{\partial w^*}{\partial \hat{y}} \cdot w^{*(p-1)} &= 0. \end{aligned}$$

Now, if we consider w^* as per Theorem 1 in the main paper, we have that $\frac{\partial w^*}{\partial \hat{y}}(u) = 0$ for $u \in \mathcal{J}^*$. Hence $\frac{\partial w^*}{\partial \hat{y}} \cdot w^{*(p-1)} = 0$ implies

$$\frac{\partial w^*}{\partial \hat{y}} \cdot w^{*(p-1)} = \sum_{u \in \mathcal{J}^*} \underbrace{\frac{\partial w^*}{\partial \hat{y}}(u)}_{=0} w^{*(p-1)}(u) + \sum_{u \in \overline{\mathcal{J}}^*} \frac{\partial w^*}{\partial \hat{y}}(u) \tau^{p-1} \frac{\ell_{\hat{y}y}(u)}{\alpha^*} = 0$$

$$\sum_{u \in \overline{\mathcal{J}}^*} \frac{\partial w^*}{\partial \hat{y}}(u) \ell_{\hat{y}y}(u) = 0$$

$$\sum_{u \in \mathcal{J}^*} \underbrace{\frac{\partial w^*}{\partial \hat{y}}(u)}_{=0} \ell_{\hat{y}y}(u) + \sum_{u \in \overline{\mathcal{J}}^*} \frac{\partial w^*}{\partial \hat{y}}(u) \ell_{\hat{y}y}(u) = 0$$

$$\frac{\partial w^*}{\partial \hat{y}} \cdot \ell_{\hat{y}y} = 0\,.$$

References

[1] S. P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. 3