Relaxed Clipping: A Global Training Method for Robust Regression and Classification (Supplementary Material)

Proof Details

For Proposition 1, recall the definitions from the main body of the paper (using the same equation numbers as in the main body)

$$\ell_c(y, \mathbf{x}^\top \boldsymbol{\theta}) = \min(1, \ell(y, \mathbf{x}^\top \boldsymbol{\theta})) \quad (2)$$

$$\ell_\rho(y, \mathbf{x}^\top \boldsymbol{\theta}) = \rho \ell(y, \mathbf{x}^\top \boldsymbol{\theta}) + 1 - \rho. \quad (4)$$

Proposition 1 *For any loss $\ell(y, \mathbf{x}^\top \boldsymbol{\theta})$, we have $\ell_c(y, \mathbf{x}^\top \boldsymbol{\theta}) = \min_{0 \le \rho \le 1} \ell_\rho(y, \mathbf{x}^\top \boldsymbol{\theta})$.*

Proof of Proposition 1

Notice that $\frac{d}{d\rho}\ell_\rho(y, \mathbf{x}^\top \boldsymbol{\theta}) = \ell(y, \mathbf{x}^\top \boldsymbol{\theta}) - 1$. Hence if $\ell(y, \mathbf{x}^\top \boldsymbol{\theta}) > 1$, then $\rho = 0$ will be the optimizer, obtaining a minimum objective value of 1. Otherwise if $\ell(y, \mathbf{x}^\top \boldsymbol{\theta}) < 1$, then $\rho = 1$ will be the optimizer, obtaining a minimum objective value of $\ell(y, \mathbf{x}^\top \boldsymbol{\theta})$. If $\ell(y, \mathbf{x}^\top \boldsymbol{\theta}) = 1$, then all feasible values of ρ yield the same objective value of 1. ∎

To prove Lemma 1, recall the definition from the main body of the paper

$$\ell^*(y, \alpha) = \sup_{\boldsymbol{\theta}} \alpha \mathbf{x}^\top \boldsymbol{\theta} - \ell(y, \mathbf{x}^\top \boldsymbol{\theta}). \quad (8)$$

Lemma 1 *For any convex differentiable loss function $\ell(y, \mathbf{x}^\top \boldsymbol{\theta})$ such that the level sets of $\ell_\alpha(\mathbf{v}) = \alpha \mathbf{x}^\top (\boldsymbol{\theta} - \mathbf{v}) + \ell(y, \mathbf{x}^\top \mathbf{v})$ are bounded*

$$\ell(y, \mathbf{x}^\top \boldsymbol{\theta}) = \sup_{\alpha} \alpha \mathbf{x}^\top \boldsymbol{\theta} - \ell^*(y, \alpha). \quad (9)$$

Proof of Lemma 1

First note that

$$\sup_\alpha \alpha \mathbf{x}^\top \boldsymbol{\theta} - \ell^*(y, \alpha) = \sup_\alpha \alpha \mathbf{x}^\top \boldsymbol{\theta} - \sup_{\mathbf{v}} \alpha \mathbf{x}^\top \mathbf{v} - \ell(y, \mathbf{x}^\top \mathbf{v}) \quad (18)$$

$$= \sup_\alpha \min_{\mathbf{v}} \alpha \mathbf{x}^\top (\boldsymbol{\theta} - \mathbf{v}) + \ell(y, \mathbf{x}^\top \mathbf{v}) \quad (19)$$

by the definition (8).[2] Since the inner objective has bounded level sets in $\mathbf{v}$ for all α by assumption, strong maximin duality holds in this case. [4, pp.281-2]. Hence

$$(19) = \min_{\mathbf{v}} \sup_\alpha \alpha \mathbf{x}^\top (\boldsymbol{\theta} - \mathbf{v}) + \ell(y, \mathbf{x}^\top \mathbf{v}) \quad (20)$$

Finally, notice that the inner supremum in (20) always achieves $+\infty$ unless $\mathbf{v} = \boldsymbol{\theta}$ (or $\boldsymbol{\theta}^\top \mathbf{x} = 0$), therefore the outer minimization in $\mathbf{v}$ must select $\mathbf{v} = \boldsymbol{\theta}$, yielding the final objective value $\ell(y, \mathbf{x}^\top \boldsymbol{\theta})$. ∎

For the proof of Theorem 1, first recall the main optimization objective and its equivalents that were formulated in the main body of the paper (using the same equation numbers as the main body)

$$\min_{\boldsymbol{\theta}} \frac{\gamma}{2}\|\boldsymbol{\theta}\|_2^2 + \sum_{i=1}^n \ell_c(y_i, X_{i:}\boldsymbol{\theta}) \quad (3)$$

$$= \min_{\boldsymbol{\theta}} \min_{\mathbf{0} \le \boldsymbol{\rho} \le \mathbf{1}} \frac{\gamma}{2}\|\boldsymbol{\theta}\|_2^2 + \sum_{i=1}^n \rho_i \ell(y_i, X_{i:}\boldsymbol{\theta}) + 1 - \rho_i \quad (5)$$

$$= \min_{0 \le \boldsymbol{\rho} \le 1} \min_{\boldsymbol{\theta}} \boldsymbol{\rho}^\top \boldsymbol{\ell}(\mathbf{y}, X\boldsymbol{\theta}) + \mathbf{1}^\top (\mathbf{1} - \boldsymbol{\rho}) + \frac{\gamma}{2}\|\boldsymbol{\theta}\|^2 \quad (6)$$

[2] We are using the fact that the inner objective is convex and has bounded level sets to note that the inner infimum can always be acheived, hence it can be re-expressed as a $\min$.

Theorem 1 *Let $K = XX^\top$ denote the kernel matrix over input data. Then*

$$(6) \quad = \quad \min_{-\frac{1}{\sqrt{n+1}}\mathbf{1} \leq \boldsymbol{\nu} \leq \frac{1}{\sqrt{n+1}}\mathbf{1}, \nu_1 = \frac{1}{\sqrt{n+1}}, \|\boldsymbol{\nu}\|=1} \quad \sup_{\boldsymbol{\alpha}} \quad -(n+1)\, \boldsymbol{\nu}^\top T(\boldsymbol{\alpha}) \boldsymbol{\nu} \tag{10}$$

where $\boldsymbol{\nu}$ is an $(n+1) \times 1$ vector, $\boldsymbol{\alpha}$ is an $n \times 1$ vector, and the matrix $T(\boldsymbol{\alpha})$ is given by

$$T(\boldsymbol{\alpha}) = \frac{1}{8\gamma} \begin{bmatrix} \mathbf{1}^\top \\ I \end{bmatrix} \Delta(\boldsymbol{\alpha}) K \Delta(\boldsymbol{\alpha}) \begin{bmatrix} \mathbf{1} & I \end{bmatrix} + \frac{1}{4} \begin{bmatrix} 2(\mathbf{1}^\top \boldsymbol{\ell}^*(\mathbf{y}, \boldsymbol{\alpha}) - n) & (\boldsymbol{\ell}^*(\mathbf{y}, \boldsymbol{\alpha}) + \mathbf{1})^\top \\ \boldsymbol{\ell}^*(\mathbf{y}, \boldsymbol{\alpha}) + \mathbf{1} & 0 \end{bmatrix}. \tag{11}$$

Proof of Theorem 1

First note that applying (9) from Lemma 1 to the main objective (6) yields

$$\begin{aligned} (6) \quad &= \quad \min_{0 \leq \boldsymbol{\rho} \leq 1} \min_{\boldsymbol{\theta}} \boldsymbol{\rho}^\top \boldsymbol{\ell}(\mathbf{y}, X\boldsymbol{\theta}) + \mathbf{1}^\top (\mathbf{1} - \boldsymbol{\rho}) + \frac{\gamma}{2} \|\boldsymbol{\theta}\|^2 \\ &= \quad \min_{0 \leq \boldsymbol{\rho} \leq 1} \min_{\boldsymbol{\theta}} \sup_{\boldsymbol{\alpha}} \; -\boldsymbol{\rho}^\top \boldsymbol{\ell}^*(\mathbf{y}, \boldsymbol{\alpha}) + \boldsymbol{\alpha}^\top \Delta(\boldsymbol{\rho}) X \boldsymbol{\theta} + \mathbf{1}^\top (\mathbf{1} - \boldsymbol{\rho}) + \frac{\gamma}{2} \boldsymbol{\theta}^\top \boldsymbol{\theta} \end{aligned} \tag{21}$$

where $\boldsymbol{\ell}^*(\mathbf{y}, \boldsymbol{\alpha})$ denotes the $n \times 1$ vector of dual values over training examples and $\Delta(\boldsymbol{\rho})$ denotes putting $\boldsymbol{\rho}$ on the main diagonal of a square matrix.

Next, observe that the inner minimization and maximization can be exchanged since the inner objective has bounded level sets in $\boldsymbol{\theta}$, meaning that strong minmax duality holds [4, pp.281-2]. Making the exchange then solving for the critical point in $\boldsymbol{\theta}$ yields

$$\boldsymbol{\theta} \quad = \quad -\frac{1}{\gamma} X^\top \Delta(\boldsymbol{\rho}) \boldsymbol{\alpha} \tag{12}$$

(the objective is convex in $\boldsymbol{\theta}$). Substituting this result back into (21) yields

$$(6) \quad = \quad \min_{0 \leq \boldsymbol{\rho} \leq 1} \sup_{\boldsymbol{\alpha}} \; -\boldsymbol{\rho}^\top \boldsymbol{\ell}^*(\mathbf{y}, \boldsymbol{\alpha}) + \mathbf{1}^\top (\mathbf{1} - \boldsymbol{\rho}) - \frac{1}{2\gamma} \boldsymbol{\rho}^\top \Delta(\boldsymbol{\alpha}) X X^\top \Delta(\boldsymbol{\alpha}) \boldsymbol{\rho}. \tag{22}$$

From here, deriving the reformulated objective merely involves algebraic manipulation. Consider

$$\begin{aligned} R \quad &= \quad -\boldsymbol{\rho}^\top \boldsymbol{\ell}^*(\mathbf{y}, \boldsymbol{\alpha}) - \mathbf{1}^\top (\boldsymbol{\rho} - \mathbf{1}) - \frac{1}{2\gamma} \boldsymbol{\rho}^\top \Delta(\boldsymbol{\alpha}) K \Delta(\boldsymbol{\alpha}) \boldsymbol{\rho} && (23) \\ &= \quad -\left(a + \boldsymbol{\rho}^\top \mathbf{b} + \boldsymbol{\rho}^\top C \boldsymbol{\rho}\right) && (24) \end{aligned}$$

where $a = -n$, $\mathbf{b} = \boldsymbol{\ell}^*(\mathbf{y}, \boldsymbol{\alpha}) + \mathbf{1}$, $C = \frac{1}{2\gamma} \Delta(\boldsymbol{\alpha}) K \Delta(\boldsymbol{\alpha})$, and $K = XX^\top$. Now let $\boldsymbol{\eta} = 2\boldsymbol{\rho} - \mathbf{1}$, which implies $\boldsymbol{\rho} = \frac{1}{2}(\boldsymbol{\eta} + \mathbf{1})$. Therefore, one can rewrite the objective as

$$\begin{aligned} R \quad &= \quad -\left(a + \tfrac{1}{2}\boldsymbol{\eta}^\top \mathbf{b} + \tfrac{1}{2}\mathbf{1}^\top \mathbf{b} + \tfrac{1}{4}\boldsymbol{\eta}^\top C \boldsymbol{\eta} + \tfrac{1}{2}\boldsymbol{\eta}^\top C \mathbf{1} + \tfrac{1}{4}\mathbf{1}^\top C \mathbf{1}\right) && (25) \\ &= \quad -\left(p + \boldsymbol{\eta}^\top \mathbf{s} + \boldsymbol{\eta}^\top S \boldsymbol{\eta}\right) && (26) \end{aligned}$$

where $p = a + \frac{1}{2}\mathbf{1}^\top \mathbf{b} + \frac{1}{4}\mathbf{1}^\top C \mathbf{1}$, $\mathbf{s} = \frac{1}{2}\mathbf{b} + \frac{1}{2} C \mathbf{1}$, and $S = \frac{1}{4} C$. Now note that if we let

$$\mathbf{v} \quad = \quad \begin{bmatrix} 1 \\ \boldsymbol{\eta} \end{bmatrix} \tag{27}$$

the objective can be rewritten as

$$R \quad = \quad -\mathbf{v}^\top T(\boldsymbol{\alpha}) \mathbf{v} \tag{28}$$

where

$$\begin{aligned} T(\boldsymbol{\alpha}) \quad &= \quad \begin{bmatrix} p & \frac{1}{2}\mathbf{s}^\top \\ \frac{1}{2}\mathbf{s} & S \end{bmatrix} && (29) \\ &= \quad \frac{1}{4} \begin{bmatrix} \mathbf{1}^\top \\ I \end{bmatrix} C \begin{bmatrix} \mathbf{1} & I \end{bmatrix} + \frac{1}{4} \begin{bmatrix} 4a + 2\mathbf{1}^\top \mathbf{b} & \mathbf{b}^\top \\ \mathbf{b} & 0 \end{bmatrix} && (30) \\ &= \quad (11). \end{aligned}$$

Finally, introduce the rescaling $\boldsymbol{\nu} = \frac{1}{\sqrt{n+1}}\mathbf{v}$, so that the objective becomes

$$R \quad = \quad -(n+1)\,\boldsymbol{\nu}^\top T(\boldsymbol{\alpha})\boldsymbol{\nu} \tag{31}$$

which gives the objective used in (10).

Next, to derive the transformed constraints, notice that $-\frac{1}{\sqrt{n+1}}\mathbf{1} \leq \boldsymbol{\nu} \leq \frac{1}{\sqrt{n+1}}\mathbf{1}$ and $\nu_1 = \frac{1}{\sqrt{n+1}}$ in (10) simply follow from the original constraints $0 \leq \boldsymbol{\rho} \leq 1$ in (6) plus the additional constraint $v_1 = 1$ introduced in (27), via the substitution chain $\boldsymbol{\rho} \mapsto \boldsymbol{\eta} \mapsto \mathbf{v} \mapsto \boldsymbol{\nu}$ given above.

The introduction of the final constraint $\|\boldsymbol{\nu}\| = 1$ requires an additional argument. Recall from the proof of Proposition 1 that the solution to (6) always has the property that $\boldsymbol{\rho} \in \{0,1\}^n$, hence adding this constraint to (6) does not change its optimal objective value. Given the above substitutions, imposing this discreteness constraint on $\boldsymbol{\rho}$ is equivalent to asserting $\nu_i \in \{\pm\frac{1}{\sqrt{n+1}}\}$ for all i. However, given the box constraints that have already been established on $\boldsymbol{\nu}$, this reduces to enforcing $\|\boldsymbol{\nu}\| = 1$. ∎

Before proving Theorem 2 we need to establish some useful relationships between the translated versions of the vectors and matrices introduced above.

Lemma 2 *Let M^* denote the matrix that participates in the optimal solution of (16). Consider the rescaling $M = (n+1)M^*$. If we define $\mathbf{m} = M_{:1}$ (i.e. the first column of M), and $\hat{\mathbf{m}} = \mathbf{m}_{2:n+1}$ (i.e. the subvector of $\mathbf{m}$ such that $\mathbf{m}^\top = [1\ \hat{\mathbf{m}}^\top]$), then from the feasibility constraints $M^* \succeq 0$ and $\boldsymbol{\delta}(M^*) = \frac{1}{n+1}\mathbf{1}$ it follows that $M \succeq 0$, $\delta(M) = \mathbf{1}$, $-1 \leq M_{ij} \leq 1$, $\mathbf{m}_1 = 1$, and $-1 \leq \mathbf{m}_i \leq 1$ for all i, j.*

From the definitions $M^* = \sum_{j=1}^k \sigma_j^* \boldsymbol{\nu}_j^* \boldsymbol{\nu}_j^{*\top}$ and $\bar{\boldsymbol{\nu}}^* = \sum_{j=1}^k \sigma_j^* \boldsymbol{\nu}_j^*$ given at the end of Section 4, and the fact that M^* satisfies the constraints in (16), if one defines $\mathbf{v}_j = \frac{1}{\nu_{1j}^*}\boldsymbol{\nu}_j^*$ and $\sigma_j = \sigma_j^*$, it follows that $\mathbf{v}_{1j} = 1$, $\sigma_j \geq 0$, $\sum_j \sigma_j = 1$, and therefore $M = \sum_j \sigma_j \mathbf{v}_j \mathbf{v}_j^\top$.

We will also use the definitions $\boldsymbol{\eta}_j = \mathbf{v}_{2:n+1,j}$ and $\bar{\boldsymbol{\eta}} = \sum_j \sigma_j \boldsymbol{\eta}_j$, so that $\mathbf{v}_j = [1\ \boldsymbol{\eta}_j^\top]^\top$, as in (27).

Finally, consider the definition of the rounded solution $\hat{\boldsymbol{\rho}} = \frac{1}{2}(\mathbf{1} + \bar{\boldsymbol{\nu}}_{2:n+1}^* \sqrt{n+1})$ given at the end of Section 4. Furthermore, define $\boldsymbol{\rho}_j = \frac{1}{2}(\mathbf{1} + \boldsymbol{\eta}_j)$ and $\bar{\boldsymbol{\rho}} = \sum_j \sigma_j \boldsymbol{\rho}_j$. From the collection of definitions and properties above, one can determine that $\hat{\boldsymbol{\rho}} = \frac{1}{2}(\mathbf{1} + \bar{\boldsymbol{\nu}}_{2:n+1}^* \sqrt{n+1}) = \frac{1}{2}(\mathbf{1} + \mathbf{m}_{2:n+1}) = \frac{1}{2}(\mathbf{1} + \hat{\mathbf{m}}) = \frac{1}{2}(\mathbf{1} + \bar{\boldsymbol{\eta}}) = \bar{\boldsymbol{\rho}}$. It also follows that $0 \leq \hat{\boldsymbol{\rho}}_i \leq 1$ for all i.

Theorem 2 (Part 1) *$\hat{R}(\hat{\boldsymbol{\rho}}, \boldsymbol{\alpha}^*) \leq 2R(\boldsymbol{\rho}^*, \boldsymbol{\theta}^*) \leq 2n$, where $\hat{R}(\hat{\boldsymbol{\rho}}, \boldsymbol{\alpha}^*)$ is the value of (10) at the rounded solution $\hat{\boldsymbol{\rho}} = \frac{1}{2}(\mathbf{1} + \bar{\boldsymbol{\nu}}_{2:n+1}^* \sqrt{n+1})$.*

Proof of Theorem 2 (Part 1)

Let $(\boldsymbol{\alpha}^*, M^*)$ denote the optimal solution to the relaxation (15) and let $(\boldsymbol{\rho}^*, \boldsymbol{\theta}^*)$ denote the optimal solution to the target problem (6). Consider the rescaling $M = (n+1)M^*$ discussed in Lemma 2 above, so that $M \succeq 0$ and $\boldsymbol{\delta}(M) = \mathbf{1}$. Define

$$\hat{R}(M, \boldsymbol{\alpha}^*) \quad = \quad -\text{tr}(MT(\boldsymbol{\alpha}^*)) \tag{32}$$

which gives the value of the relaxed objective achieved by $(M^*, \boldsymbol{\alpha}^*)$ in (15) where $M^* = M/(n+1)$. Recall that $R(\boldsymbol{\rho}^*, \boldsymbol{\theta}^*)$ denotes the optimum objective value obtained in (6). It then immediately follows that

$$\hat{R}(M, \boldsymbol{\alpha}^*) \quad \leq \quad R(\boldsymbol{\rho}^*, \boldsymbol{\theta}^*) \quad \leq \quad n. \tag{33}$$

since the optimal relaxed objective value is a lower bound by construction (13), and the second inequality follows trivially from the observation that $R(\mathbf{0}, \mathbf{0}) = n$. Our goal is to analyze the relaxed objective $\hat{R}(M, \boldsymbol{\alpha}^*)$ at the saddle point $(M, \boldsymbol{\alpha}^*)$ to bound the effects of rounding the solution to $\hat{\boldsymbol{\rho}}$.

We need some preliminary definitions to make things clearer. From the definition given in (11) rewrite $T(\boldsymbol{\alpha})$ as

$$T(\boldsymbol{\alpha}) \quad = \quad D(\boldsymbol{\alpha}) + G(\boldsymbol{\alpha}) - P \tag{34}$$

where

$$D(\boldsymbol{\alpha}) = \frac{1}{4}\begin{bmatrix} 2\mathbf{1}^\top \boldsymbol{\ell}^*(\mathbf{y},\boldsymbol{\alpha}) & \boldsymbol{\ell}^*(\mathbf{y},\boldsymbol{\alpha})^\top \\ \boldsymbol{\ell}^*(\mathbf{y},\boldsymbol{\alpha}) & 0 \end{bmatrix} \tag{35}$$

$$G(\boldsymbol{\alpha}) = \frac{1}{4}\begin{bmatrix} \mathbf{1}^\top \\ I \end{bmatrix} C(\boldsymbol{\alpha}) \begin{bmatrix} \mathbf{1} & I \end{bmatrix} \tag{36}$$

$$C(\boldsymbol{\alpha}) = \frac{1}{2\gamma}\Delta(\boldsymbol{\alpha})K\Delta(\boldsymbol{\alpha}) \tag{37}$$

$$P = \frac{1}{4}\begin{bmatrix} 2n & -\mathbf{1}^\top \\ -\mathbf{1} & 0 \end{bmatrix}. \tag{38}$$

One can then obtain a straightforward decompostion

$$\begin{aligned} \hat{R}(M,\boldsymbol{\alpha}^*) &= -\text{tr}(MT(\boldsymbol{\alpha}^*)) && (39)\\ &= \text{tr}\left(M(-D(\boldsymbol{\alpha}^*) - G(\boldsymbol{\alpha}^*) + P)\right) && (40)\\ &= \text{tr}\left(M(-D(\boldsymbol{\alpha}^*) - 2G(\boldsymbol{\alpha}^*))\right) + \text{tr}(MG(\boldsymbol{\alpha}^*)) + \text{tr}(MP). && (41) \end{aligned}$$

They key property of this decomposition is that each term must be nonnegative. That is, note that

$$\begin{aligned} \text{tr}(MP) &\geq 0 && (42)\\ \text{tr}(MG(\boldsymbol{\alpha}^*)) &\geq 0 && (43) \end{aligned}$$

since, in the first case, $M_{11} = 1$ and $-1 \leq M_{ij} \leq 1$, and in the second case, $G(\boldsymbol{\alpha}^*) \succeq 0$. Also, by Lemma 3 below we have

$$\text{tr}\left(M(-D(\boldsymbol{\alpha}^*) - 2G(\boldsymbol{\alpha}^*))\right) \geq 0. \tag{44}$$

Therefore, since each term in (41) is nonnegative and their sum is at most $R(\boldsymbol{\rho}^*, \boldsymbol{\theta}^*)$ (established in (33)), it must follow that each individual term in (41) is upper bounded by $R(\boldsymbol{\rho}^*, \boldsymbol{\theta}^*)$. In particular

$$\text{tr}(MG(\boldsymbol{\alpha}^*)) \leq R(\boldsymbol{\rho}^*, \boldsymbol{\theta}^*). \tag{45}$$

Finally, to bound the objective value obtained by $\hat{\boldsymbol{\rho}}$, we will work instead with its translation $\mathbf{m} = [1\ (2\hat{\boldsymbol{\rho}} - \mathbf{1})^\top] = M_{:1}$, defined in Lemma 2 above. In particular, using the definition (32) let

$$\hat{R}(\hat{\boldsymbol{\rho}}, \boldsymbol{\alpha}^*) = \hat{R}(\mathbf{m}\mathbf{m}^\top, \boldsymbol{\alpha}^*) = -\mathbf{m}^\top T(\boldsymbol{\alpha}^*)\mathbf{m} \tag{46}$$

where $\mathbf{m} = [1\ (2\hat{\boldsymbol{\rho}} - \mathbf{1})^\top] = M_{:1}$ (the latter equality follows from Lemma 2 above). Note that by the special structure of $D(\boldsymbol{\alpha}^*)$ and P we have

$$\begin{aligned} \text{tr}(MD(\boldsymbol{\alpha}^*)) &= \mathbf{m}^\top D(\boldsymbol{\alpha}^*)\mathbf{m} && (47)\\ \text{tr}(MP) &= \mathbf{m}^\top P\mathbf{m} && (48) \end{aligned}$$

Also since $\mathbf{m}^\top T(\boldsymbol{\alpha}^*)\mathbf{m} \leq \text{tr}(MT(\boldsymbol{\alpha}^*))$ by the optimality of $M^* = M/(n+1)$ (see (13)) we must also have

$$0 \leq \mathbf{m}^\top G(\boldsymbol{\alpha}^*)\mathbf{m} \leq \text{tr}(M^*G(\boldsymbol{\alpha}^*)) \leq R(\boldsymbol{\rho}^*, \boldsymbol{\theta}^*). \tag{49}$$

From (47)–(49) we conclude

$$\begin{aligned} \hat{R}(\hat{\boldsymbol{\rho}}, \boldsymbol{\alpha}^*) &= -\mathbf{m}^\top T(\boldsymbol{\alpha}^*)\mathbf{m} && (50)\\ &= -\mathbf{m}^\top \left(D(\boldsymbol{\alpha}^*) + G(\boldsymbol{\alpha}^*) - P\right)\mathbf{m} && (51)\\ &\leq -\mathbf{m}^\top \left(D(\boldsymbol{\alpha}^*) - P\right)\mathbf{m} && (52)\\ &= -\text{tr}\left(M(D(\boldsymbol{\alpha}^*) - P)\right) && (53)\\ &= -\text{tr}\left(M(T(\boldsymbol{\alpha}^*) - G(\boldsymbol{\alpha}^*))\right) && (54)\\ &= -\text{tr}(M(T(\boldsymbol{\alpha}^*)) + \text{tr}(MG(\boldsymbol{\alpha}^*)) && (55)\\ &\leq 2R(\boldsymbol{\rho}^*, \boldsymbol{\theta}^*). && (56) \end{aligned}$$

■

Lemma 3 *tr* $(M(-D(\boldsymbol{\alpha}^*) - 2G(\boldsymbol{\alpha}^*))) \geq 0$.

Proof of Lemma 3

The proof of this lemma uses the definitions and properties from Lemma 2 above. Note that by the special structure of $D(\boldsymbol{\alpha}^*)$ and $G(\boldsymbol{\alpha}^*)$ we have

$$\begin{aligned}
&\text{tr}(M(D(\boldsymbol{\alpha}^*) + 2G(\boldsymbol{\alpha}^*)) && (57)\\
&= \sum_j \sigma_j \mathbf{v}_j^\top (D(\boldsymbol{\alpha}^*) + 2G(\boldsymbol{\alpha}^*))\mathbf{v}_j && (58)\\
&= \tfrac{1}{2}\mathbf{1}^\top \boldsymbol{\ell}^*(\mathbf{y}, \boldsymbol{\alpha}^*) + \tfrac{1}{2}\bar{\boldsymbol{\eta}}^\top \boldsymbol{\ell}^*(\mathbf{y}, \boldsymbol{\alpha}^*) + \tfrac{1}{2}\mathbf{1}^\top C(\boldsymbol{\alpha}^*)\mathbf{1} + \bar{\boldsymbol{\eta}}^\top C(\boldsymbol{\alpha}^*)\mathbf{1} + \tfrac{1}{2}\sum_j \sigma_j \boldsymbol{\eta}_j^\top C(\boldsymbol{\alpha}^*)\boldsymbol{\eta}_j && (59)\\
&= \bar{\boldsymbol{\rho}}^\top \boldsymbol{\ell}^*(\mathbf{y}, \boldsymbol{\alpha}^*) + 2\sum_j \sigma_j \boldsymbol{\rho}_j^\top C(\boldsymbol{\alpha}^*)\boldsymbol{\rho}_j && (60)\\
&= \bar{\boldsymbol{\rho}}^\top \boldsymbol{\ell}^*(\mathbf{y}, \boldsymbol{\alpha}^*) + \frac{1}{\gamma}\sum_j \sigma_j \boldsymbol{\rho}_j^\top \Delta(\boldsymbol{\alpha}^*)K\Delta(\boldsymbol{\alpha}^*)\boldsymbol{\rho}_j && (61)\\
&= \bar{\boldsymbol{\rho}}^\top \boldsymbol{\ell}^*(\mathbf{y}, \boldsymbol{\alpha}^*) + \frac{1}{\gamma}\sum_j \sigma_j \boldsymbol{\alpha}^{*\top} \Delta(\boldsymbol{\rho}_j)K\Delta(\boldsymbol{\rho}_j)\boldsymbol{\alpha}^* && (62)\\
&= \bar{\boldsymbol{\rho}}^\top \boldsymbol{\ell}^*(\mathbf{y}, \boldsymbol{\alpha}^*) + \frac{1}{\gamma}\boldsymbol{\alpha}^{*\top} \bar{K}\boldsymbol{\alpha}^* && (63)
\end{aligned}$$

where

$$\bar{K} = \sum_j \sigma_j \Delta(\boldsymbol{\rho}_j)K\Delta(\boldsymbol{\rho}_j) \quad (64)$$

Note that $\bar{K} \succeq 0$ since $\sigma_j \geq 0$.

It remains to show (63) is nonpositive. Recall that $\boldsymbol{\alpha}^*$ was assumed to maximize (14), which means

$$\begin{aligned}
&-\bar{\boldsymbol{\rho}}^\top \boldsymbol{\ell}^*(\mathbf{y}, \boldsymbol{\alpha}^*) - \frac{1}{2\gamma}\boldsymbol{\alpha}^{*\top}\bar{K}\boldsymbol{\alpha}^* && (65)\\
&= \max_{\boldsymbol{\alpha}} -\bar{\boldsymbol{\rho}}^\top \boldsymbol{\ell}^*(\mathbf{y}, \boldsymbol{\alpha}) - \frac{1}{2\gamma}\boldsymbol{\alpha}^\top \bar{K}\boldsymbol{\alpha} && (66)\\
&= \max_{\boldsymbol{\alpha}} \min_{\boldsymbol{\theta}} \bar{\boldsymbol{\rho}}^\top \boldsymbol{\ell}(\mathbf{y}, X\boldsymbol{\theta}) - \boldsymbol{\alpha}^\top \Delta(\bar{\boldsymbol{\rho}})X\boldsymbol{\theta} - \frac{1}{2\gamma}\boldsymbol{\alpha}^\top \bar{K}\boldsymbol{\alpha} && (67)\\
&= \min_{\boldsymbol{\theta}} \max_{\boldsymbol{\alpha}} \bar{\boldsymbol{\rho}}^\top \boldsymbol{\ell}(\mathbf{y}, X\boldsymbol{\theta}) - \boldsymbol{\alpha}^\top \Delta(\bar{\boldsymbol{\rho}})X\boldsymbol{\theta} - \frac{1}{2\gamma}\boldsymbol{\alpha}^\top \bar{K}\boldsymbol{\alpha} && (68)\\
&= \min_{\boldsymbol{\theta}} \bar{\boldsymbol{\rho}}^\top \boldsymbol{\ell}(\mathbf{y}, X\boldsymbol{\theta}) + \frac{\gamma}{2}\boldsymbol{\theta}^\top X^\top \Delta(\bar{\boldsymbol{\rho}})\bar{K}^+\Delta(\bar{\boldsymbol{\rho}})X\boldsymbol{\theta} && (69)\\
&\geq 0
\end{aligned}$$

where we have used the fact that $\hat{\boldsymbol{\rho}} \geq 0$ and

$$\begin{aligned}
-\bar{\boldsymbol{\rho}}^\top \boldsymbol{\ell}^*(\mathbf{y}, \boldsymbol{\alpha}) &= \min_{\boldsymbol{\theta}} \bar{\boldsymbol{\rho}}^\top \boldsymbol{\ell}(\mathbf{y}, X\theta) - \boldsymbol{\alpha}^\top \Delta(\bar{\boldsymbol{\rho}})X\boldsymbol{\theta} && (70)\\
\bar{K}\boldsymbol{\alpha}^* &= -\gamma\Delta(\bar{\boldsymbol{\rho}})X\boldsymbol{\theta}^*. && (71)
\end{aligned}$$

Unfortunately, (65) is not quite the same as (63) (note the γ verus 2γ difference in the denominators). However, given the above results for (65) we can now return to (63) and observe

$$\begin{aligned}
&-\bar{\boldsymbol{\rho}}^\top \boldsymbol{\ell}^*(\mathbf{y}, \boldsymbol{\alpha}^*) - \frac{1}{\gamma}\boldsymbol{\alpha}^{*\top}\bar{K}\boldsymbol{\alpha}^* && (72)\\
&= -\bar{\boldsymbol{\rho}}^\top \boldsymbol{\ell}^*(\mathbf{y}, \boldsymbol{\alpha}^*) + \boldsymbol{\alpha}^{*\top}\Delta(\bar{\boldsymbol{\rho}})X\boldsymbol{\theta}^* \quad \text{by (71)} && (73)\\
&= \bar{\boldsymbol{\rho}}^\top \boldsymbol{\ell}(\mathbf{y}, X\boldsymbol{\theta}^*) - \boldsymbol{\alpha}^{*\top}\Delta(\bar{\boldsymbol{\rho}})X\boldsymbol{\theta}^* + \boldsymbol{\alpha}^{*\top}\Delta(\bar{\boldsymbol{\rho}})X\boldsymbol{\theta}^* \quad \text{by (70)} && (74)\\
&= \bar{\boldsymbol{\rho}}^\top \boldsymbol{\ell}(\mathbf{y}, X\boldsymbol{\theta}^*) && (75)\\
&\geq 0.
\end{aligned}$$

■

Theorem 2 (Part 2) *If the unclipped loss $\ell(y,\hat{y})$ is b-Lipschitz in $\hat{y}$ for $b < \infty$ and either $\mathbf{y}$ or K remains bounded, then there exists a $c < \infty$ such that $R(\hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\theta}}) \leq c$.*

Proof of Theorem 2 (Part 2)

For simplicity we will show the proof for the case when $\ell(y,\hat{y})$ is strictly convex in $\hat{y}$. Recall that $(M^*, \boldsymbol{\alpha}^*)$ solves the saddle point problem (14). Therefore, given M^*, the resulting objective must be at equilibrium in $\boldsymbol{\alpha}^*$, hence

$$\nabla_{\boldsymbol{\alpha}} \hat{R}(M^*, \boldsymbol{\alpha}^*) \quad = \quad -\Delta(\hat{\boldsymbol{\rho}}) \boldsymbol{\ell}^*(\mathbf{y}, \boldsymbol{\alpha}^*)' - \frac{1}{\gamma} \bar{K} \boldsymbol{\alpha}^* \quad = \quad 0 \tag{76}$$

where $\boldsymbol{\ell}^*(\mathbf{y}, \boldsymbol{\alpha}^*)'$ denotes the vector of derivatives of $\ell^*(y_i, \alpha_i^*)'$ with respect to α_i at α_i^*, and $\bar{K}$ is defined in (64). Therefore we have that $-\Delta(\hat{\boldsymbol{\rho}}) \boldsymbol{\ell}^*(\mathbf{y}, \boldsymbol{\alpha}^*)' = \bar{K}\boldsymbol{\alpha}^*/\gamma$. (The strict convexity of $\ell(y,\hat{y})$ in $\hat{y}$ ensures that $\ell^*(y,\alpha)$ is smooth in α, hence the derivative exists at α^* [22, Section 26].)

Now consider the definition

$$\mathbf{b}^* \quad = \quad -\Delta(\hat{\boldsymbol{\rho}}) \boldsymbol{\ell}^*(\mathbf{y}, \boldsymbol{\alpha}^*)' \quad = \quad \frac{1}{\gamma} \bar{K} \boldsymbol{\alpha}^*. \tag{77}$$

Note that under the assumptions of the theorem, there must exist a constant $c < \infty$ such that $\|\mathbf{b}^*\|_\infty \leq c$. We prove this fact in two cases. First, consider the case where $\mathbf{y}$ is bounded. Since $\ell(y,\hat{y})$ is b-Lipschitz in $\hat{y}$ by assumption, it follows that $\|\boldsymbol{\alpha}^*\|_\infty \leq b$ [22, Corollary 13.3.3]. Since both $\mathbf{y}$ and $\boldsymbol{\alpha}^*$ are bounded, $\boldsymbol{\ell}^*(\mathbf{y}, \boldsymbol{\alpha}^*)'$ must be bounded (since it was determined above that $\ell^*(y,\alpha)$ is smooth in α). Finally, $0 \leq \hat{\boldsymbol{\rho}} \leq 1$ was established in Lemma 2 above; therefore $\mathbf{b}^*$ is bounded. Second, consider the case where K is bounded. Then $\bar{K}$ must be bounded, since both σ_j and $\boldsymbol{\rho}_j$ are bounded. Finally, since $\boldsymbol{\alpha}^*$ is bounded (as established above) it follows that $\mathbf{b}^* = \bar{K}\boldsymbol{\alpha}^*/\gamma$ is bounded.

Finally, consider the objective value obtained by fixing $\hat{\boldsymbol{\rho}}$ and re-optimizing $\boldsymbol{\theta}$

$$\begin{aligned}
R(\hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\theta}}) &= \min_{\boldsymbol{\theta}} R(\hat{\boldsymbol{\rho}}, \boldsymbol{\theta}) && (78)\\
&= \min_{\boldsymbol{\theta}} \hat{\boldsymbol{\rho}}^\top \boldsymbol{\ell}(\mathbf{y}, X\boldsymbol{\theta}) + \frac{\gamma}{2}\|\boldsymbol{\theta}\|^2 + \mathbf{1}^\top(\mathbf{1} - \hat{\boldsymbol{\rho}}) && (79)\\
&= \max_{\|\boldsymbol{\alpha}\|_\infty \leq b} -\hat{\boldsymbol{\rho}}^\top \boldsymbol{\ell}^*(\mathbf{y}, \boldsymbol{\alpha}) - \frac{1}{2\gamma}\boldsymbol{\alpha}^\top \Delta(\hat{\boldsymbol{\rho}}) K \Delta(\hat{\boldsymbol{\rho}}) \boldsymbol{\alpha} + \mathbf{1}^\top(\mathbf{1} - \hat{\boldsymbol{\rho}}) && (80)\\
&\leq \max_{\|\boldsymbol{\alpha}\|_\infty \leq b} -\hat{\boldsymbol{\rho}}^\top \boldsymbol{\ell}^*(\mathbf{y}, \boldsymbol{\alpha}) + \mathbf{1}^\top(\mathbf{1} - \hat{\boldsymbol{\rho}}) && (81)\\
&\leq \max_{\|\boldsymbol{\alpha}\|_\infty \leq b} -\hat{\boldsymbol{\rho}}^\top \boldsymbol{\ell}^*(\mathbf{y}, \boldsymbol{\alpha}^*) + (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^\top \mathbf{b}^* + \mathbf{1}^\top(\mathbf{1} - \hat{\boldsymbol{\rho}}) && (82)\\
&= \max_{\|\boldsymbol{\alpha}\|_\infty \leq b} -\hat{\boldsymbol{\rho}}^\top \boldsymbol{\ell}^*(\mathbf{y}, \boldsymbol{\alpha}^*) - \frac{1}{\gamma}\boldsymbol{\alpha}^{*\top} \bar{K} \boldsymbol{\alpha}^* + \boldsymbol{\alpha}^\top \mathbf{b}^* + \mathbf{1}^\top(\mathbf{1} - \hat{\boldsymbol{\rho}}) && (83)
\end{aligned}$$

Note that by Lemma 3 and the observation after (44) we have that the two first terms in (83) are bounded by $0 \leq -\hat{\boldsymbol{\rho}}^\top \boldsymbol{\ell}^*(\mathbf{y}, \boldsymbol{\alpha}^*) - \frac{1}{\gamma}\boldsymbol{\alpha}^{*\top} \bar{K} \boldsymbol{\alpha}^* \leq R(\boldsymbol{\rho}^*, \boldsymbol{\theta}^*)$. Therefore

$$\begin{aligned}
(83) \quad &\leq \max_{\|\boldsymbol{\alpha}\|_\infty \leq b} R(\boldsymbol{\rho}^*, \boldsymbol{\theta}^*) + \boldsymbol{\alpha}^\top \mathbf{b}^* + \mathbf{1}^\top(\mathbf{1} - \hat{\boldsymbol{\rho}}) && (84)\\
&\leq \max_{\|\boldsymbol{\alpha}\|_\infty \leq b} 2n + \boldsymbol{\alpha}^\top \mathbf{b}^* && (85)
\end{aligned}$$

Since it was established above that $\mathbf{b}^*$ is bounded, it follows that $R(\hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\theta}})$ is bounded. ∎