A. Proofs

A.1. Proof of Theorem 1

If the composite loss $\widetilde{\ell}(z)$ is convex, it is linear.

Proof: The composite loss is an odd function:

$$\widetilde{\ell}(-z) = \ell(-z) - \ell(z) = -\widetilde{\ell}(z),$$

Therefore, $\frac{\mathrm{d}^2}{\mathrm{d}z^2}\widetilde{\ell}(z) = -\frac{\mathrm{d}^2}{\mathrm{d}z^2}\widetilde{\ell}(-z)$. If the composite loss $\widetilde{\ell}(z)$ is convex, $\frac{\mathrm{d}^2}{\mathrm{d}z^2}\widetilde{\ell}(z) \geq 0$ holds for all z. Since the convexity of $\widetilde{\ell}(z)$ implies the convexity of $\widetilde{\ell}(-z)$, $\frac{\mathrm{d}^2}{\mathrm{d}z^2}\widetilde{\ell}(-z) \geq 0$ should also hold for all z. However, if $\frac{\mathrm{d}^2}{\mathrm{d}z^2}\widetilde{\ell}(z) > 0$, then $\frac{\mathrm{d}^2}{\mathrm{d}z^2}\widetilde{\ell}(-z) < 0$ holds, which is contradictory to the convexity of $\widetilde{\ell}(-z)$. Therefore, $\frac{\mathrm{d}^2}{\mathrm{d}z^2}\widetilde{\ell}(z) = 0$ should hold, which is satisfied only when $\widetilde{\ell}(z)$ is linear. □

A.2. Proof of Lemma 2

$J_{\mathrm{S}}(\boldsymbol{\alpha})$ is strongly convex in $\boldsymbol{\alpha}$ with parameter at least λ, and thus

$$\begin{aligned} J_{\mathrm{S}}(\boldsymbol{\alpha}) &\geq J_{\mathrm{S}}(\boldsymbol{\alpha}^*_{\mathrm{S}}) + \nabla J_{\mathrm{S}}(\boldsymbol{\alpha}^*_{\mathrm{S}})^\top(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*_{\mathrm{S}}) + \lambda\|\boldsymbol{\alpha} - \boldsymbol{\alpha}^*_{\mathrm{S}}\|_2^2 \\ &\geq J_{\mathrm{S}}(\boldsymbol{\alpha}^*_{\mathrm{S}}) + \lambda\|\boldsymbol{\alpha} - \boldsymbol{\alpha}^*_{\mathrm{S}}\|_2^2, \end{aligned}$$

where we use the optimality condition $\nabla J_{\mathrm{S}}(\boldsymbol{\alpha}^*_{\mathrm{S}}) = \mathbf{0}$. Similarly, we can prove the other two inequalities. □

A.3. Proof of Lemma 3

The difference function can be written as

$$J_{\mathrm{S}}(\boldsymbol{\alpha}, \boldsymbol{u}) - J_{\mathrm{S}}(\boldsymbol{\alpha}) = \frac{1}{4}\boldsymbol{\alpha}^\top \boldsymbol{u}_1 \boldsymbol{\alpha} + \frac{1}{2}\boldsymbol{u}_2^\top \boldsymbol{\alpha} - \pi \boldsymbol{u}_3^\top \boldsymbol{\alpha},$$

with a partial gradient

$$\frac{\partial}{\partial \boldsymbol{\alpha}}(J_{\mathrm{S}}(\boldsymbol{\alpha}, \boldsymbol{u}) - J_{\mathrm{S}}(\boldsymbol{\alpha})) = \frac{1}{2}\boldsymbol{u}_1\boldsymbol{\alpha} + \frac{1}{2}\boldsymbol{u}_2 - \pi\boldsymbol{u}_3.$$

Given the δ-ball of $\boldsymbol{\alpha}^*_{\mathrm{S}}$, i.e., $B_\delta(\boldsymbol{\alpha}^*_{\mathrm{S}}) = \{\boldsymbol{\alpha} \mid \|\boldsymbol{\alpha} - \boldsymbol{\alpha}^*_{\mathrm{S}}\|_2 \leq \delta\}$, it is easy to see that for any $\boldsymbol{\alpha} \in B_\delta(\boldsymbol{\alpha}^*_{\mathrm{S}})$,

$$\|\boldsymbol{\alpha}\|_2 \leq \|\boldsymbol{\alpha} - \boldsymbol{\alpha}^*_{\mathrm{S}}\|_2 + \|\boldsymbol{\alpha}^*_{\mathrm{S}}\|_2 \leq 1 + M_\alpha,$$

and then

$$\left\|\frac{\partial}{\partial \boldsymbol{\alpha}}(J_{\mathrm{S}}(\boldsymbol{\alpha}, \boldsymbol{u}) - J_{\mathrm{S}}(\boldsymbol{\alpha}))\right\|_2 \leq \frac{1}{2}(1 + M_\alpha)\|\boldsymbol{u}_1\|_{\mathrm{Fro}} + \frac{1}{2}\|\boldsymbol{u}_2\|_2 + \pi\|\boldsymbol{u}_3\|_2.$$

This means that $J_{\mathrm{S}}(\cdot, \boldsymbol{u}) - J_{\mathrm{S}}(\cdot)$ is Lipschitz continuous on $B_\delta(\boldsymbol{\alpha}^*_{\mathrm{S}})$ with a Lipschitz constant of order $\mathcal{O}(\|\boldsymbol{u}_1\|_{\mathrm{Fro}} + \|\boldsymbol{u}_2\|_2 + \|\boldsymbol{u}_3\|_2)$. □

A.4. Proof of Lemma 5

The difference function can be written as

$$J_{\mathrm{LL}}(\boldsymbol{\alpha}, \boldsymbol{u}) - J_{\mathrm{LL}}(\boldsymbol{\alpha}) = -\pi\boldsymbol{u}_3^\top\boldsymbol{\alpha} + u_4(\boldsymbol{\alpha}).$$

Given $\boldsymbol{\alpha} \in B_\delta(\boldsymbol{\alpha}^*_{\mathrm{LL}})$, we have known that $-\pi\boldsymbol{u}_3^\top\boldsymbol{\alpha}$ is Lipschitz continuous with a Lipschitz constant of order $\mathcal{O}(\|\boldsymbol{u}_3\|_2)$ in the proof of Lemma 3. Consequently, $J_{\mathrm{LL}}(\cdot, \boldsymbol{u}) - J_{\mathrm{LL}}(\cdot)$ is Lipschitz continuous on $B_\delta(\boldsymbol{\alpha}^*_{\mathrm{LL}})$ with a Lipschitz constant of order $\mathcal{O}(\|\boldsymbol{u}_3\|_2 + \mathrm{Lip}(u_4))$. □

A.5. Proof of Lemma 7

Same as the proof of Lemma 5. □

A.6. Proof of Theorem 4

Let $\boldsymbol{u}_1$, $\boldsymbol{u}_2$ and $\boldsymbol{u}_3$ be defined as in Eq. (13). According to the *central limit theorem*,

$$\|\boldsymbol{u}_1\|_{\mathrm{Fro}} = \mathcal{O}_p(n'^{-1/2}), \quad \|\boldsymbol{u}_2\|_2 = \mathcal{O}_p(n'^{-1/2}), \quad \|\boldsymbol{u}_3\|_2 = \mathcal{O}_p(n^{-1/2}),$$

as $n, n' \to \infty$. Thus, we have

$$\begin{aligned} \|\widehat{\boldsymbol{\alpha}}_{\mathrm{S}} - \boldsymbol{\alpha}^*_{\mathrm{S}}\|_2 &\leq \lambda^{-1}\omega(\boldsymbol{u}) \\ &= \mathcal{O}(\|\boldsymbol{u}_1\|_{\mathrm{Fro}} + \|\boldsymbol{u}_2\|_2 + \|\boldsymbol{u}_3\|_2) \\ &= \mathcal{O}_p(n^{-1/2} + n'^{-1/2}) \end{aligned}$$

by Lemma 2, Lemma 3, and Proposition 6.1 in Bonnans & Shapiro (1998, p. 19).

On the other hand,

$$|\widehat{J}_{\mathrm{S}}(\widehat{\boldsymbol{\alpha}}_{\mathrm{S}}) - J_{\mathrm{S}}(\boldsymbol{\alpha}^*_{\mathrm{S}})| \leq |\widehat{J}_{\mathrm{S}}(\widehat{\boldsymbol{\alpha}}_{\mathrm{S}}) - \widehat{J}_{\mathrm{S}}(\boldsymbol{\alpha}^*_{\mathrm{S}})| + |\widehat{J}_{\mathrm{S}}(\boldsymbol{\alpha}^*_{\mathrm{S}}) - J_{\mathrm{S}}(\boldsymbol{\alpha}^*_{\mathrm{S}})|,$$

in which

$$\begin{aligned} \widehat{J}_{\mathrm{S}}(\widehat{\boldsymbol{\alpha}}_{\mathrm{S}}) - \widehat{J}_{\mathrm{S}}(\boldsymbol{\alpha}^*_{\mathrm{S}}) &= (\widehat{\boldsymbol{\alpha}}_{\mathrm{S}} + \boldsymbol{\alpha}^*_{\mathrm{S}})^\top \left(\frac{1}{4n'} \sum_{i=1}^{n'} \boldsymbol{\varphi}(\boldsymbol{x}'_i)\boldsymbol{\varphi}(\boldsymbol{x}'_i)^\top + \frac{\lambda}{2} I_m \right) (\widehat{\boldsymbol{\alpha}}_{\mathrm{S}} - \boldsymbol{\alpha}^*_{\mathrm{S}}) \\ &\quad + \left(\frac{1}{2n'} \sum_{i=1}^{n'} \boldsymbol{\varphi}(\boldsymbol{x}'_i) \right)^\top (\widehat{\boldsymbol{\alpha}}_{\mathrm{S}} - \boldsymbol{\alpha}^*_{\mathrm{S}}) - \pi \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\varphi}(\boldsymbol{x}_i) \right)^\top (\widehat{\boldsymbol{\alpha}}_{\mathrm{S}} - \boldsymbol{\alpha}^*_{\mathrm{S}}), \\ \widehat{J}_{\mathrm{S}}(\boldsymbol{\alpha}^*_{\mathrm{S}}) - J_{\mathrm{S}}(\boldsymbol{\alpha}^*_{\mathrm{S}}) &= \frac{1}{4} \boldsymbol{\alpha}^{*\top}_{\mathrm{S}} \boldsymbol{u}_1 \boldsymbol{\alpha}^*_{\mathrm{S}} + \frac{1}{2} \boldsymbol{u}_2 \boldsymbol{\alpha}^*_{\mathrm{S}} - \pi \boldsymbol{u}_3 \boldsymbol{\alpha}^*_{\mathrm{S}}. \end{aligned}$$

Since $0 \leq \varphi_j(\boldsymbol{x}) \leq 1$, $\|\boldsymbol{\alpha}^*_{\mathrm{S}}\|_2 \leq M_\alpha$ and $\|\widehat{\boldsymbol{\alpha}}_{\mathrm{S}}\|_2 \leq M_\alpha$,

$$\begin{aligned} |\widehat{J}_{\mathrm{S}}(\widehat{\boldsymbol{\alpha}}_{\mathrm{S}}) - J_{\mathrm{S}}(\boldsymbol{\alpha}^*_{\mathrm{S}})| &\leq |\widehat{J}_{\mathrm{S}}(\widehat{\boldsymbol{\alpha}}_{\mathrm{S}}) - \widehat{J}_{\mathrm{S}}(\boldsymbol{\alpha}^*_{\mathrm{S}})| + |\widehat{J}_{\mathrm{S}}(\boldsymbol{\alpha}^*_{\mathrm{S}}) - J_{\mathrm{S}}(\boldsymbol{\alpha}^*_{\mathrm{S}})| \\ &\leq \mathcal{O}_p(\|\widehat{\boldsymbol{\alpha}}_{\mathrm{S}} - \boldsymbol{\alpha}^*_{\mathrm{S}}\|_2) + \mathcal{O}_p(\|\boldsymbol{u}_1\|_{\mathrm{Fro}} + \|\boldsymbol{u}_2\|_2 + \|\boldsymbol{u}_3\|_2) \\ &= \mathcal{O}_p(n^{-1/2} + n'^{-1/2}), \end{aligned}$$

which completes the proof. □

A.7. Proof of Theorem 6

Let $\boldsymbol{u}_3$ and $u_4(\boldsymbol{\alpha})$ be defined as in Eq. (14). The gradient of u_4 is given by

$$\nabla u_4(\boldsymbol{\alpha}) = \frac{1}{n'} \sum_{i=1}^{n'} \frac{\boldsymbol{\varphi}(\boldsymbol{x}'_i)}{1 + \exp(-\boldsymbol{\varphi}(\boldsymbol{x}'_i)^\top \boldsymbol{\alpha})} - \int \frac{\boldsymbol{\varphi}(\boldsymbol{x})}{1 + \exp(-\boldsymbol{\varphi}(\boldsymbol{x})^\top \boldsymbol{\alpha})} p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

According to the central limit theorem,

$$\|\boldsymbol{u}_3\|_2 = \mathcal{O}_p(n^{-1/2}), \quad \mathrm{Lip}(u_4) = \mathcal{O}_p(n'^{-1/2}),$$

as $n, n' \to \infty$, since $\mathrm{Lip}(u_4) = \sup_{\boldsymbol{\alpha}} \|\nabla u_4(\boldsymbol{\alpha})\|_2$ and

$$\sup_{\boldsymbol{\alpha} \in \mathbb{R}^m, \boldsymbol{x} \in \mathbb{R}^d} \left\| \frac{\boldsymbol{\varphi}(\boldsymbol{x})}{1 + \exp(-\boldsymbol{\varphi}(\boldsymbol{x})^\top \boldsymbol{\alpha})} \right\|_2 \leq m^{1/2} < \infty.$$

Thus, we have

$$\begin{aligned} \|\widehat{\boldsymbol{\alpha}}_{\mathrm{LL}} - \boldsymbol{\alpha}^*_{\mathrm{LL}}\|_2 &\leq \lambda^{-1}\omega(\boldsymbol{u}) \\ &= \mathcal{O}(\|\boldsymbol{u}_3\|_2 + \mathrm{Lip}(u_4)) \end{aligned}$$

$$= \mathcal{O}_p(n^{-1/2} + n'^{-1/2})$$

by Lemma 2, Lemma 5, and Proposition 6.1 in Bonnans & Shapiro (1998, p. 19).

On the other hand,

$$|\widehat{J}_{\mathrm{LL}}(\widehat{\boldsymbol{\alpha}}_{\mathrm{LL}}) - J_{\mathrm{LL}}(\boldsymbol{\alpha}^*_{\mathrm{LL}})| \le |\widehat{J}_{\mathrm{LL}}(\widehat{\boldsymbol{\alpha}}_{\mathrm{LL}}) - \widehat{J}_{\mathrm{LL}}(\boldsymbol{\alpha}^*_{\mathrm{LL}})| + |\widehat{J}_{\mathrm{LL}}(\boldsymbol{\alpha}^*_{\mathrm{LL}}) - J_{\mathrm{LL}}(\boldsymbol{\alpha}^*_{\mathrm{LL}})|.$$

For the second term,

$$\begin{aligned} |\widehat{J}_{\mathrm{LL}}(\boldsymbol{\alpha}^*_{\mathrm{LL}}) - J_{\mathrm{LL}}(\boldsymbol{\alpha}^*_{\mathrm{LL}})| &= |-\pi \boldsymbol{u}_3^\top \boldsymbol{\alpha}^*_{\mathrm{LL}} + u_4(\boldsymbol{\alpha}^*_{\mathrm{LL}})| \\ &\le \pi M_\alpha \|\boldsymbol{u}_3\|_2 + |u_4(\boldsymbol{\alpha}^*_{\mathrm{LL}})| \\ &= \mathcal{O}_p(n^{-1/2} + n'^{-1/2}) \end{aligned}$$

according to the central limit theorem. For the first term, it is a bit more complex:

$$\begin{aligned} |\widehat{J}_{\mathrm{LL}}(\widehat{\boldsymbol{\alpha}}_{\mathrm{LL}}) - \widehat{J}_{\mathrm{LL}}(\boldsymbol{\alpha}^*_{\mathrm{LL}})| \le& \left|\frac{\lambda}{2}(\widehat{\boldsymbol{\alpha}}_{\mathrm{LL}} + \boldsymbol{\alpha}^*_{\mathrm{LL}})^\top(\widehat{\boldsymbol{\alpha}}_{\mathrm{LL}} - \boldsymbol{\alpha}^*_{\mathrm{LL}})\right| + \left|\pi\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\varphi}(\boldsymbol{x}_i)\right)^\top(\widehat{\boldsymbol{\alpha}}_{\mathrm{LL}} - \boldsymbol{\alpha}^*_{\mathrm{LL}})\right| \\ &+ \frac{1}{n'}\sum_{i=1}^{n'} |\ln(1+\exp(\boldsymbol{\varphi}(\boldsymbol{x}'_i)^\top \widehat{\boldsymbol{\alpha}}_{\mathrm{LL}})) - \ln(1+\exp(\boldsymbol{\varphi}(\boldsymbol{x}'_i)^\top \boldsymbol{\alpha}^*_{\mathrm{LL}}))|. \end{aligned}$$

Let $f(z,t) = \ln(1+\exp(z+t))$, then $\lim_{t\to 0} f(z,t) = f(z,0)$ and

$$\lim_{t\to 0}\frac{f(z,t)-f(z,0)}{t} = \lim_{t\to 0}\frac{\partial}{\partial t}f(z,t) = \frac{1}{1+\exp(-z-t)} < \infty,$$

where we use *L'Hôpital's rule*. In other words, $f(z,t)$ approaches $f(z,0)$ in $\mathcal{O}(t)$ as $t \to 0$. Subsequently, for any $\boldsymbol{x} \in \mathbb{R}^d$, by $z = \boldsymbol{\varphi}(\boldsymbol{x})^\top \boldsymbol{\alpha}^*_{\mathrm{LL}}$ and $t = \boldsymbol{\varphi}(\boldsymbol{x})^\top \widehat{\boldsymbol{\alpha}}_{\mathrm{LL}} - \boldsymbol{\varphi}(\boldsymbol{x})^\top \boldsymbol{\alpha}^*_{\mathrm{LL}}$ we can obtain

$$\begin{aligned} |\ln(1+\exp(\boldsymbol{\varphi}(\boldsymbol{x})^\top \widehat{\boldsymbol{\alpha}}_{\mathrm{LL}})) - \ln(1+\exp(\boldsymbol{\varphi}(\boldsymbol{x})^\top \boldsymbol{\alpha}^*_{\mathrm{LL}}))| &= \mathcal{O}(|\boldsymbol{\varphi}(\boldsymbol{x})^\top \widehat{\boldsymbol{\alpha}}_{\mathrm{LL}} - \boldsymbol{\varphi}(\boldsymbol{x})^\top \boldsymbol{\alpha}^*_{\mathrm{LL}}|) \\ &= \mathcal{O}(m^{1/2}\|\widehat{\boldsymbol{\alpha}}_{\mathrm{LL}} - \boldsymbol{\alpha}^*_{\mathrm{LL}}\|_2), \end{aligned}$$

which results in $|\widehat{J}_{\mathrm{LL}}(\widehat{\boldsymbol{\alpha}}_{\mathrm{LL}}) - \widehat{J}_{\mathrm{LL}}(\boldsymbol{\alpha}^*_{\mathrm{LL}})| = \mathcal{O}_p(n^{-1/2} + n'^{-1/2})$. □

A.8. Proof of Theorem 8

The proof goes along the same line as that of Theorem 6. Let $\boldsymbol{u}_3$ and $u_5(\boldsymbol{\alpha})$ be defined as in Eq. (15). Note that the function $\max\{0,(1+z)/2,z\}$ is piecewise linear in z, differentiable almost everywhere, and $0 \le (\mathrm{d}/\mathrm{d}z)\max\{0,(1+z)/2,z\} \le 1$. As a result,

$$\|\boldsymbol{u}_3\|_2 = \mathcal{O}_p(n^{-1/2}), \quad \mathrm{Lip}(u_5) = \mathcal{O}_p(n'^{-1/2}),$$

as $n, n' \to \infty$, and

$$\begin{aligned} \|\widehat{\boldsymbol{\alpha}}_{\mathrm{DH}} - \boldsymbol{\alpha}^*_{\mathrm{DH}}\|_2 &\le \lambda^{-1}\omega(\boldsymbol{u}) \\ &= \mathcal{O}(\|\boldsymbol{u}_3\|_2 + \mathrm{Lip}(u_5)) \\ &= \mathcal{O}_p(n^{-1/2} + n'^{-1/2}) \end{aligned}$$

by Lemma 2, Lemma 7, and Proposition 6.1 in Bonnans & Shapiro (1998, p. 19).

On the other hand,

$$\begin{aligned} |\widehat{J}_{\mathrm{DH}}(\widehat{\boldsymbol{\alpha}}_{\mathrm{DH}}) - J_{\mathrm{DH}}(\boldsymbol{\alpha}^*_{\mathrm{DH}})| &\le |\widehat{J}_{\mathrm{DH}}(\widehat{\boldsymbol{\alpha}}_{\mathrm{DH}}) - \widehat{J}_{\mathrm{DH}}(\boldsymbol{\alpha}^*_{\mathrm{DH}})| + |\widehat{J}_{\mathrm{DH}}(\boldsymbol{\alpha}^*_{\mathrm{DH}}) - J_{\mathrm{DH}}(\boldsymbol{\alpha}^*_{\mathrm{DH}})| \\ &\le \frac{1}{n'}\sum_{i=1}^{n'} |\max\{0,(1+\boldsymbol{\varphi}(\boldsymbol{x}'_i)^\top \widehat{\boldsymbol{\alpha}}_{\mathrm{LL}})/2, \boldsymbol{\varphi}(\boldsymbol{x}'_i)^\top \widehat{\boldsymbol{\alpha}}_{\mathrm{LL}}\} \end{aligned}$$

$$- \max\{0, (1 + \boldsymbol{\varphi}(\boldsymbol{x}'_i)^\top \boldsymbol{\alpha}^*_{\mathrm{LL}})/2, \boldsymbol{\varphi}(\boldsymbol{x}'_i)^\top \boldsymbol{\alpha}^*_{\mathrm{LL}}\}| + \mathcal{O}_p(n^{-1/2} + n'^{-1/2}).$$

Let $f(z,t) = \max\{0, (1+z+t)/2, z+t\}$, then $\lim_{t\to 0} f(z,t) = f(z,0)$ and for $z \in \mathbb{R} \setminus \{0,1\}$,

$$\lim_{t\to 0} \frac{f(z,t) - f(z,0)}{t} = \lim_{t\to 0} \frac{\partial}{\partial t} f(z,t) \in \left\{0, \frac{1}{2}, 1\right\}.$$

In other words, $f(z,t)$ approaches $f(z,0)$ in $\mathcal{O}(t)$ as $t \to 0$ almost surely. Subsequently, for any $\boldsymbol{x} \in \mathbb{R}^d$, by $z = \boldsymbol{\varphi}(\boldsymbol{x})^\top \boldsymbol{\alpha}^*_{\mathrm{DH}}$ and $t = \boldsymbol{\varphi}(\boldsymbol{x})^\top \widehat{\boldsymbol{\alpha}}_{\mathrm{DH}} - \boldsymbol{\varphi}(\boldsymbol{x})^\top \boldsymbol{\alpha}^*_{\mathrm{DH}}$ we can obtain

$$\begin{aligned}
|\max\{0, (1 + \boldsymbol{\varphi}(\boldsymbol{x})^\top \widehat{\boldsymbol{\alpha}}_{\mathrm{LL}})/2, \boldsymbol{\varphi}(\boldsymbol{x})^\top \widehat{\boldsymbol{\alpha}}_{\mathrm{LL}}\} - \max\{0, (1 + \boldsymbol{\varphi}(\boldsymbol{x})^\top \boldsymbol{\alpha}^*_{\mathrm{LL}})/2, \boldsymbol{\varphi}(\boldsymbol{x})^\top \boldsymbol{\alpha}^*_{\mathrm{LL}}\}| &= \mathcal{O}(|\boldsymbol{\varphi}(\boldsymbol{x})^\top \widehat{\boldsymbol{\alpha}}_{\mathrm{LL}} - \boldsymbol{\varphi}(\boldsymbol{x})^\top \boldsymbol{\alpha}^*_{\mathrm{LL}}|) \\
&= \mathcal{O}(m^{1/2} \|\widehat{\boldsymbol{\alpha}}_{\mathrm{LL}} - \boldsymbol{\alpha}^*_{\mathrm{LL}}\|_2) \\
&= \mathcal{O}_p(n^{-1/2} + n'^{-1/2}),
\end{aligned}$$

which completes the proof. □

B. Optimization problems

In this section, we give exact optimization problems for the optimization methods presented in the paper. The logistic regression and logistic loss method is solved with a quasi-Newton method, and therefore we provide the derivatives in Sec. B.1.

The Hinge loss and Double Hinge loss result in quadratic problems. The ramp-loss is solved via a sequence of quadratic problems. All quadratic problems are expressed in the form

$$\begin{aligned}
\min_{\boldsymbol{\alpha}} \quad & \tfrac{1}{2} \boldsymbol{\alpha}^\top H \boldsymbol{\alpha} + \boldsymbol{f}^\top \boldsymbol{\alpha} \\
\text{s.t.} \quad & L\boldsymbol{\alpha} \preceq \boldsymbol{k} \\
& \boldsymbol{l} \preceq \boldsymbol{\alpha}
\end{aligned}$$

This standard form can then just be plugged into an off-the-shelf optimization package such as Gurobi, IBM CPLEX or MATLAB's internal 'quadprog' function.

B.1. Logistic loss

The gradient for the objective function in Eq. (8) is

$$\begin{aligned}
\frac{\partial \widehat{J}_{\mathrm{LL}}(\boldsymbol{\alpha}, b)}{\partial \boldsymbol{\alpha}} &= -\frac{\pi}{n} \Phi_{\mathrm{P}}^\top \mathbf{1} + \lambda \boldsymbol{\alpha} \\
&\quad - \frac{1}{n'} \sum_{j=1}^{n'} \ell'_{\mathrm{LL}} \left(-\boldsymbol{\alpha}^\top \boldsymbol{\varphi}(\boldsymbol{x}'_j) - b\right) \boldsymbol{\varphi}(\boldsymbol{x}'_j),
\end{aligned}$$

where $\ell'_{\mathrm{LL}}(z)$ is the derivative of $\ell_{\mathrm{LL}}(z)$:

$$\ell'_{\mathrm{LL}}(z) = -\frac{\exp(-z)}{1 + \exp(-z)}.$$

The derivative with respect to the unregularized constant b is

$$\frac{\partial \widehat{J}_{\mathrm{LL}}(\boldsymbol{\alpha}, b)}{\partial b} = -\pi - \frac{1}{n'} \sum_{j=1}^{n'} \ell'_{\mathrm{LL}} \left(-\boldsymbol{\alpha}^\top \boldsymbol{\varphi}(\boldsymbol{x}'_j) - b\right).$$

B.2. Double Hinge Loss - PU Learning

The objective function can be expressed as

$$-\frac{\pi}{n} \sum_{i=1}^{n} g(\boldsymbol{x}_i) + \frac{1}{n'} \sum_{j=1}^{n'} \max\left(0, \max\left(g(\boldsymbol{x}'_j), \frac{1}{2} + \frac{1}{2} g(\boldsymbol{x}'_j)\right)\right) + \frac{\lambda}{2} \|g\|_2^2$$

$$= -\frac{\pi}{n}\sum_{i=1}^{n}\left(\sum_{\ell=1}^{m}\alpha_\ell\varphi_\ell(\boldsymbol{x}_i)+b\right)+\frac{1}{n'}\sum_{j=1}^{n'}\max\left(0,\max\left(\sum_{\ell=1}^{m}\alpha_\ell\varphi_\ell(\boldsymbol{x}'_j)+b,\frac{1}{2}+\frac{1}{2}\left(\sum_{\ell=1}^{m}\alpha_\ell\varphi_\ell(\boldsymbol{x}'_j)+b\right)\right)\right)+\frac{\lambda}{2}\sum_{\ell=1}^{m}\alpha_\ell^2$$

The objective function can then be expressed as

$$\begin{aligned}\min_{\boldsymbol{\alpha},b,\boldsymbol{\xi}} \quad & -\tfrac{\pi}{n}\mathbf{1}^\top\Phi_{\mathrm{P}}\boldsymbol{\alpha}-\pi b+\tfrac{1}{n'}\mathbf{1}^\top\boldsymbol{\xi}+\tfrac{\lambda}{2}\boldsymbol{\alpha}^\top\boldsymbol{\alpha}\\ \text{s.t.} \quad & \boldsymbol{\xi}\succeq\mathbf{0},\\ & \boldsymbol{\xi}\succeq\tfrac{1}{2}\mathbf{1}+\tfrac{1}{2}\Phi_{\mathrm{U}}\boldsymbol{\alpha}+\tfrac{1}{2}b\mathbf{1},\\ & \boldsymbol{\xi}\succeq\Phi_{\mathrm{U}}\boldsymbol{\alpha}+b\mathbf{1},\end{aligned}$$

Let

$$\boldsymbol{\gamma}=\begin{bmatrix}\boldsymbol{\alpha}_{b\times 1}\\ b\\ \boldsymbol{\xi}_{n'\times 1}\end{bmatrix}.$$

Then H is defined as

$$H=\begin{bmatrix}\lambda I_{m\times m} & O_{m\times 1} & O_{m'\times n'}\\ O_{1\times m} & 0 & O_{1\times n'}\\ O_{n'\times m} & O_{n'\times 1} & O_{n'\times n'}\end{bmatrix},$$

where $O_{n\times m}$ is a zero matrix of n rows and m columns. The linear part of the objective is

$$\boldsymbol{f}=\begin{bmatrix}-\frac{\pi}{n}\Phi_{\mathrm{P}}^\top\mathbf{1}\\ -\pi\\ \frac{1}{n'}\mathbf{1}_{n'\times 1}\end{bmatrix}$$

The lower-bound is

$$\boldsymbol{l}=\begin{bmatrix}-\infty_{m\times 1}\\ -\infty\\ \mathbf{0}_{n'\times 1}\end{bmatrix}.$$

The first linear constraint is

$$\begin{aligned}\boldsymbol{\xi}&\succeq\frac{1}{2}\mathbf{1}+\frac{1}{2}\Phi_{\mathrm{U}}\boldsymbol{\alpha}+\frac{1}{2}b\mathbf{1}\\ \frac{1}{2}\Phi_{\mathrm{U}}\boldsymbol{\alpha}+\frac{1}{2}b\mathbf{1}-\boldsymbol{\xi}&\preceq-\frac{1}{2}\mathbf{1}\\ \begin{bmatrix}\frac{1}{2}\Phi_{\mathrm{U}} & \frac{1}{2}\mathbf{1}_{n'\times 1} & -I_{n'\times n'}\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha}\\ u\\ \boldsymbol{\xi}\end{bmatrix}&\preceq-\frac{1}{2}\mathbf{1}_{n'\times 1}.\end{aligned}$$

The second linear constraint is

$$\begin{aligned}\boldsymbol{\xi}&\succeq\Phi_{\mathrm{U}}\boldsymbol{\alpha}+b\mathbf{1}\\ \Phi_{\mathrm{U}}\boldsymbol{\alpha}+b\mathbf{1}-\boldsymbol{\xi}&\preceq\mathbf{0}_{n'\times 1}\\ \begin{bmatrix}\Phi_{\mathrm{U}} & \mathbf{1}_{n'\times 1} & -I_{n'\times n'}\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha}\\ b\\ \boldsymbol{\xi}\end{bmatrix}&\preceq\mathbf{0}_{n'\times 1}.\end{aligned}$$

Combining the two sets of inequalities, we get

$$L=\begin{bmatrix}\frac{1}{2}\Phi_{\mathrm{U}} & \frac{1}{2}\mathbf{1}_{n'\times 1} & -I_{n'\times n'}\\ \Phi_{\mathrm{U}} & \mathbf{1}_{n'\times 1} & -I_{n'\times n'}\end{bmatrix},$$

and

$$\boldsymbol{k}=\begin{bmatrix}-\frac{1}{2}\mathbf{1}_{n'\times 1}\\ \mathbf{0}_{n'\times 1}\end{bmatrix}.$$

B.3. Weighted hinge loss classifier

We want a cost-sensitive classifier with a per-sample weighting. Using the model

$$g(\boldsymbol{x}) = \sum_{\ell=1}^{m} \alpha_\ell \varphi_\ell(\boldsymbol{x}_)+b,$$

where

$$\{\boldsymbol{c}_1,\dots,\boldsymbol{c}_m\} := \{\boldsymbol{x}_1,\dots,\boldsymbol{x}_n\},$$

we wish to minimize

$$\begin{aligned} J(g) &= \frac{1}{n}\sum_{i=1}^{b} w_i \ell_{\mathrm{H}}\left(y_i \sum_{\ell=1}^{m}\alpha_\ell\varphi_\ell(\boldsymbol{x}_i)+b\right)+\frac{\lambda}{2}\boldsymbol{\alpha}^\top\boldsymbol{\alpha}, \\ &= \frac{1}{2n}\sum_{i=1}^{n} w_i \max\left(0, 1-y_i\sum_{\ell=1}^{m}\alpha_\ell\varphi_\ell(\boldsymbol{x}_i)+b\right)+\frac{\lambda}{2}\boldsymbol{\alpha}^\top\boldsymbol{\alpha}. \end{aligned}$$

This gives a QP of

$$\begin{array}{ll} \min\limits_{\boldsymbol{\alpha},b,\boldsymbol{\xi}} & \frac{1}{2n}\boldsymbol{w}^\top\boldsymbol{\xi}+\frac{\lambda}{2}\boldsymbol{\alpha}^\top R\boldsymbol{\alpha} \\ \text{s.t.} & \xi_i \geq 0, \qquad \forall i=1,\dots,n \\ & \xi_i \geq 1-y_i\sum_{\ell=1}^{b}\alpha_\ell k(\boldsymbol{x}_i,\boldsymbol{c}_\ell)+u \qquad \forall i=1,\dots,n. \end{array}$$

We then set

$$\boldsymbol{\gamma} = \begin{bmatrix} \boldsymbol{\alpha} \\ b \\ \boldsymbol{\xi} \end{bmatrix}.$$

H is then

$$H = \begin{bmatrix} \lambda I & O_{m\times 1} & O_{m\times n} \\ O_{1\times n} & 0 & O_{1\times n} \\ O_{n\times n} & O_{n\times 1} & O_{n\times n} \end{bmatrix}.$$

The linear term is

$$\boldsymbol{f} = \begin{bmatrix} \boldsymbol{0}_{m\times 1} \\ 0 \\ \frac{1}{2n}\boldsymbol{w} \end{bmatrix}$$

The lower bound is

$$\boldsymbol{l} = \begin{bmatrix} -\infty_{m\times 1} \\ -\infty \\ \boldsymbol{0}_{n\times 1} \end{bmatrix}$$

Define $\bar{\Phi}$ as

$$\bar{\Phi}_{i\ell} = y_i\varphi_\ell(\boldsymbol{x}_i).$$

The constraint can be written in matrix form as

$$\begin{gathered} \boldsymbol{\xi} \succeq \boldsymbol{1}_{n\times 1} - \left(\bar{\Phi}\boldsymbol{\alpha}+b\boldsymbol{y}\right) \\ -\bar{\Phi}\boldsymbol{\alpha}-b\boldsymbol{y}-\boldsymbol{\xi} \preceq -\boldsymbol{1}_{n\times 1} \end{gathered}$$

The matrix is then

$$L = \begin{bmatrix} -\Phi & -\boldsymbol{y} & -I_{n\times n} \end{bmatrix},$$

and $\boldsymbol{k}$ is

$$\boldsymbol{k} = \left[-\boldsymbol{1}_{n\times 1}\right].$$

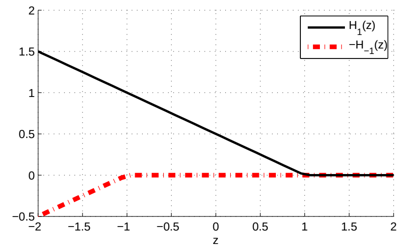

Figure 6. Decomposition of the ramp-loss into convex and concave parts.

B.4. Weighted ramp-loss classifier (CCCP)

Classification with the ramp-loss is difficult, due to the the non-convexity of the loss function. One popular method to perform optimization is to split the non-convex function into a convex and concave part. The concave part is then upper-bounded by a linear function, and optimization is iteratively performed: minimization of the upper-bound, and tightening of the upper-bound around the new minima. We minimize the ramp-loss problem here using this approach. This is a straightforward application of the convex-concave procedure (CCCP) in Yuille & Rangarajan (2002) and is essentially the same as Collobert et al. (2006).

We wish to minimize the following non-convex objective function:

$$J(\boldsymbol{\alpha}, b) = \frac{1}{n}\sum_{i=1}^{n} w_i \ell_{\mathrm{R}}\left(y_i \sum_{\ell=1}^{m} \alpha_\ell \varphi_\ell(\boldsymbol{x}_i) + b\right) + \frac{\lambda}{2}\boldsymbol{\alpha}^\top\boldsymbol{\alpha}, \tag{16}$$

where the ramp loss $\ell_{\mathrm{R}}(z)$ is defined as

$$\ell_{\mathrm{R}}(z) = \max\left(0, \min\left(1, \frac{1}{2} - \frac{1}{2}z\right)\right) = \frac{1}{2}\max\left(0, \min(2, 1-z)\right).$$

By defining the following (slightly more general) hinge loss

$$H_\epsilon(z) = \frac{1}{2}\max(0, \epsilon - z),$$

the ramp loss $\ell_{\mathrm{R}}(z)$ can be decomposed as:

$$\ell_{\mathrm{R}}(z) = H_1(z) - H_{-1}(z).$$

This is illustrated in Fig. 6. The objective Eq. (16) can therefore be decomposed as

$$\begin{aligned}
J(\boldsymbol{\alpha}, b) &= J_{\mathrm{vex}}(\boldsymbol{\alpha}, b) + J_{\mathrm{cave}}(\boldsymbol{\alpha}, b),\\
J_{\mathrm{vex}}(\boldsymbol{\alpha}, b) &= \frac{1}{n}\sum_{i=1}^{n} w_i H_1\left(\sum_{\ell=1}^{m} \alpha_\ell \varphi_\ell(\boldsymbol{x}_i) + b\right) + \frac{\lambda}{2}\boldsymbol{\alpha}^\top\boldsymbol{\alpha},\\
J_{\mathrm{cave}}(\boldsymbol{\alpha}, b) &= -\frac{1}{n}\sum_{i=1}^{n} w_i H_{-1}\left(\sum_{\ell=1}^{m} \alpha_\ell \varphi_\ell(\boldsymbol{x}_i) + b\right)
\end{aligned}$$

The following self-evident relation can be used to upper-bound the concave part

$$\begin{aligned}
tz - f(z) &\leq \sup_{y\in\mathbb{R}} yt - f(y)\\
\Rightarrow f(z) &\geq tz - f^*(t),
\end{aligned} \tag{17}$$

where

$$f^*(t) = \sup_{y\in\mathbb{R}} yt - f(y).$$

The inequality in Eq.(17) is known as the *Fenchel inequality* and the function $f^*(z)$ is known as the *Fenchel dual* or *convex conjugate*. Applying the above inequality to $H_\epsilon(z)$, we can obtain a bound as

$$\begin{aligned} H_\epsilon(z) &\geq zt - H^*_\epsilon(t), \\ -H_\epsilon(z) &\leq H^*_\epsilon(t) - zt, \end{aligned}$$

where $H^*_\epsilon(t)$ is the Fenchel dual of $H_\epsilon(z)$. The Fenchel dual of $H_{-1}(t)$ is (the full calculation is given in Appendix B.4.3)

$$H^*_{-1}(t) = \begin{cases} -t & -\frac{1}{2} \leq t \leq 0, \\ \infty & \text{otherwise.} \end{cases}$$

We can minimize the upper-bound as

$$\arg\min_t H^*_{-1}(t) - tz = \begin{cases} t = 0 & z > -1. \\ t = -\frac{1}{2} & z \leq -1. \end{cases}$$

The concave part is then bounded, with the parameter $\boldsymbol{a}$ as

$$\bar{J}_{\text{cave}}(\boldsymbol{\alpha}, b, \boldsymbol{a}) = \frac{1}{n}\sum_{i=1}^{n} w_i \left(H^*_1(a_i) - a_i y_i \left(\sum_{\ell=1}^{m} \alpha_\ell \varphi_\ell(\boldsymbol{x}_i) + b \right) \right),$$

where $J_{\text{cave}}(\boldsymbol{\alpha}, u) \leq \bar{J}_{\text{cave}}(\boldsymbol{\alpha}, b, \boldsymbol{a})$, for any $\boldsymbol{a}$.

B.4.1. Tightening of the upper-bound

The upperbound is minimized (tightened) when

$$a_i = \begin{cases} -\frac{1}{2} & y_i \left(\sum_{\ell=1}^{m} \alpha_\ell \varphi_\ell(\boldsymbol{x}_i) + b \right) \leq -1, \\ 0 & \text{otherwise.} \end{cases}$$

B.4.2. Minimizing the objective

We wish to minimize the convex part and the upper bound $\bar{J}(\boldsymbol{\alpha}, u, \boldsymbol{a}) = J_{\text{vex}}(\boldsymbol{\alpha}, u) + \bar{J}_{\text{cave}}(\boldsymbol{\alpha}, u, \boldsymbol{a})$ with respect to $\boldsymbol{a}$. This gives an objective of

$$\bar{J}(\boldsymbol{\alpha}, b, \boldsymbol{a}) = \frac{1}{n}\sum_{i=1}^{n} w_i H_1 \left(y_i \left(\sum_{\ell=1}^{m} \alpha_\ell \varphi_\ell(\boldsymbol{x}_i) + b \right) \right) + \frac{\lambda}{2}\boldsymbol{\alpha}^\top \boldsymbol{\alpha} - \frac{1}{n}\sum_{i=1}^{n} w_i a_i y_i \left(\sum_{\ell=1}^{m} \alpha_\ell \varphi_\ell(\boldsymbol{x}_i) + b \right).$$

We define the following matrices:

$$\begin{aligned} \Phi_{i,\ell} &= y_i k(\boldsymbol{x}_i, \boldsymbol{c}_\ell), \\ \bar{\Phi}_{i,\ell} &= w_i a_i y_i k(\boldsymbol{x}_i, \boldsymbol{c}_\ell), \end{aligned}$$

The QP for this is then

$$\begin{aligned} \min_{\boldsymbol{\alpha}, b, \boldsymbol{\xi}} \quad & \tfrac{1}{2n}\boldsymbol{w}^\top \boldsymbol{\xi} + \tfrac{\lambda}{2}\boldsymbol{\alpha}^\top \boldsymbol{\alpha} - \tfrac{1}{n}\boldsymbol{1}^\top \bar{\Phi}\boldsymbol{\alpha} - b\tfrac{1}{n}\textstyle\sum_{i=1}^{n} w_i a_i y_i. \\ s.t. \quad & \xi_i \geq 0 \qquad \forall i = 1, \ldots, n \\ & \xi_i \geq 1 - y_i \left(\textstyle\sum_{\ell=1}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}_i) + b \right) \qquad \forall i = 1, \ldots, n. \end{aligned}$$

We define again

$$\boldsymbol{\gamma} = \begin{bmatrix} \boldsymbol{\alpha} \\ b \\ \boldsymbol{\xi} \end{bmatrix}.$$

The quadratic term is

$$H = \begin{bmatrix} \lambda I_{m\times m} & O_{m\times 1} & O_{n\times n} \\ O_{1\times n} & 0 & O_{1\times n} \\ O_{n\times n} & O_{n\times 1} & O_{n\times n} \end{bmatrix}.$$

The linear term is

$$\boldsymbol{f} = \begin{bmatrix} -\frac{1}{n}\bar{\Phi}^\top \mathbf{1} \\ -\frac{1}{n}\sum_{i=1}^{n} w_i a_i y_i \\ \frac{1}{2n}\boldsymbol{w} \end{bmatrix}$$

The lower-bound is

$$lb = \begin{bmatrix} -\infty_{m\times 1} \\ -\infty \\ \mathbf{0}_{n\times 1} \end{bmatrix}.$$

The linear term is

$$-\Phi\boldsymbol{\alpha} - b\boldsymbol{y} - \boldsymbol{\xi} \preceq -\mathbf{1}_{n\times 1}.$$

This gives a matrix of

$$L = \begin{bmatrix} -\Phi & -\boldsymbol{y} & -I_{n\times n} \end{bmatrix},$$

and $\boldsymbol{k}$ is

$$\boldsymbol{k} = [-\mathbf{1}_{n\times 1}].$$

B.4.3. Calculation of the Fenchel dual of $H_\epsilon(z)$

In this section, we briefly give the derivation of the Fenchel dual of $H_\epsilon(z)$

$$\begin{aligned} H_\epsilon^*(t) &= \sup_v tv - H_\epsilon(v) \\ &= \sup_v tv - \frac{1}{2}\max(0, \epsilon - v). \end{aligned}$$

To make the above easier, we split the domain of the v:

$$\begin{aligned} H_\epsilon^*(t) &= \max\left(\sup_{v\le\epsilon} tv - \frac{1}{2}\max(0,\epsilon - v), \sup_{v>\epsilon} tv - \frac{1}{2}\max(0,\epsilon - v)\right), \\ &= \max\left(\sup_{v\le\epsilon} tv - \frac{1}{2}(\epsilon - v), \sup_{v>\epsilon} tv\right). \end{aligned}$$

For the first part:

$$\begin{aligned} \sup_{v\le\epsilon} tv - \frac{1}{2}(\epsilon - v) &= \sup_{v\le\epsilon} v\left(t + \frac{1}{2}\right) - \frac{1}{2}\epsilon, \\ &= \begin{cases} \epsilon t & t \ge -\frac{1}{2}, \\ \infty & t < \frac{1}{2} \end{cases} \end{aligned}$$

The second part is

$$\sup_{t>\epsilon} tv = \begin{cases} \epsilon v & t \le 0, \\ \infty & t > 0. \end{cases}$$

Putting these two together gives:

$$H_\epsilon^*(t) = \begin{cases} \epsilon t & -\frac{1}{2} \le t \le 0, \\ \infty & \text{otherwise.} \end{cases}$$