Supplementary Material for
“On Loss Functions and Regret Bounds for Multi-category Classification”
Zhiqiang Tan and Xinwei Zhang

The Supplementary Material provides proofs and additional discussions, which are organized by the sections in the main paper.

I. TECHNICAL DETAILS

A. Preparation
For a convex function $\psi$ defined on a convex domain $\Omega$, the Bregman divergence is defined as

$$B_\psi(x, y) = \psi(x) - \psi(y) - (x - y)^T \partial \psi(y),$$

where $\partial \psi$ is a sub-gradient of $\psi$. The symmetrized Bregman divergence is

$$B_\psi(x, y) + B_\psi(y, x) = (y - x)^T \{\partial \psi(y) - \partial \psi(x)\}.$$

The following lemma shows that the Bregman divergence is nondecreasing as the first (or second) argument, $x$ or $y$, moves away from the other argument, or $x$ or $y$, along a straight line, while the second (or respectively first) argument remains fixed.

**Lemma S1.** For any $x, y \in \Omega$ and $w \in [0, 1]$, we have

$$B_\psi(x, y) \geq B_\psi(x^w, y), \tag{S1}$$

$$B_\psi(x, y) \geq B_\psi(x, x^w). \tag{S2}$$

where $x^w = (1 - w)x + wy$.

**Proof.** If $w = 0$ or 1, then (S1) and (S2) hold trivially. In the following, assume $w \in (0, 1)$. To show (S1), direct calculation yields

$$B_\psi(x, y) - B_\psi(x^w, y) = \psi(x) - \psi(x^w) - (x - x^w)^T \partial \psi(y)$$

$$= B_\psi(x, x^w) + (x - x^w)^T \{\partial \psi(x^w) - \partial \psi(y)\}$$

$$= B_\psi(x, x^w) + \frac{w}{1 - w} (x^w - y)^T \{\partial \psi(x^w) - \partial \psi(y)\}.$$ 

Hence (S1) follows because $(x^w - y)^T \{\partial \psi(x^w) - \partial \psi(y)\}$ is the symmetrized Bregman divergence between $x^w$ and $y$. From the preceding equations, we see

$$B_\psi(x, y) - B_\psi(x, x^w) = B_\psi(x^w, y) + \frac{w}{1 - w} (x^w - y)^T \{\partial \psi(x^w) - \partial \psi(y)\}.$$ 

Hence (S2) follows because $(x^w - y)^T \{\partial \psi(x^w) - \partial \psi(y)\} \geq 0$ again.

B. Proofs of results in Section III

**Proof of Proposition 2.** Denote by $\partial^1 f$ the set of all sub-gradients of $f$. For any $u \in \mathbb{R}_+^{m-1}$ and $s = \partial f(u) \in \partial^1 f(u)$, Fenchel’s conjugacy property implies that $f^*(s) = u^T s - f(u)$ and hence $s \in \text{dom}(f^*)$. Moreover, we have

$$\sum_{j=1}^m \eta_j L_{f^2}(j, u) = \sum_{j=1}^{m-1} \eta_j (-\partial_j f(u)) + \eta_m (u^T \partial f(u) - f(u))$$

$$= \sum_{j=1}^{m-1} (-\eta_j s_j) + \eta_m f^*(s) = \sum_{j=1}^m \eta_j L_f(j, s).$$

Therefore,

$$\inf_{u \in \mathbb{R}_+^{m-1}} \left\{ \sum_{j=1}^m \eta_j L_{f^2}(j, u) \right\} \geq \inf_{s \in \text{dom}(f^*)} \left\{ \sum_{j=1}^m \eta_j L_f(j, s) \right\} = H_f(\eta).$$
Next, we show the reverse inequality. For any \( \eta \in \Delta_m \), denote \( u^\eta = (\eta_1/\eta_m, \ldots, \eta_{m-1}/\eta_m)^T \) and \( s^\eta = \partial f(u^\eta) \in \partial f(u^\eta) \). Then
\[
H_f(\eta) = -\eta_m f(u^\eta) = -\eta_m \left\{ \sum_{j=1}^{m-1} u_j^\eta s_j^\eta - f^*(s^\eta) \right\} = \sum_{j=1}^{m-1} (-\eta_j s_j^\eta) + \eta_m f^*(s^\eta) = \sum_{j=1}^{m-1} \eta_j (-\partial_j f(u^\eta)) + \eta_m (u^{\eta^T} \partial f(u^\eta) - f(u^\eta)),
\]
where Fenchel’s conjugacy property, \( u^{\eta^T} s^\eta = f(u^\eta) + f^*(s^\eta) \), is used in the last equalities on the first and second lines. Hence \( H_f(\eta) \geq \inf_{u \in \mathbb{R}^m} \{ \sum_{j=1}^{m} \eta_j L_{f^2}(j, u) \} \).

**Proof of equation (19)**. By definition, \( H_f(q) = -q_m f(q_1/q_m, \ldots, q_{m-1}/q_m) \). The sub-gradient of \(-H_f\) can be calculated as
\[
-\partial_j H_f(q_1, \ldots, q_m) = \begin{cases} \partial_j f(u^q), & \text{if } j \in [m-1], \\ f(u^q) - \sum_{j=1}^{m-1} \frac{q_j}{q_m} \partial_j f(u^q), & \text{if } j = m, \end{cases}
\]
where \( u^q = (q_1/q_m, \ldots, q_{m-1}/q_m)^T \). Substituting these expressions into \( H_f(q) = -\sum_{j=1}^{m} (q_j - \eta_j) \partial_j H_f(q) \) yields the second equality in Eq. (19):
\[
H_f(q) - \sum_{j=1}^{m} (q_j - \eta_j) \partial_j H_f(q)
= -q_m f(u^q) + \sum_{j=1}^{m-1} (q_j - \eta_j) \partial_j f(u^q) + (q_m - \eta_m) \left\{ f(u^q) - \sum_{j=1}^{m-1} \frac{q_j}{q_m} \partial_j f(u^q) \right\}
= -\sum_{j=1}^{m-1} \eta_j \partial_j f(u^q) + \eta_m \left\{ -f(u^q) + \sum_{j=1}^{m-1} \frac{q_j}{q_m} \partial_j f(u^q) \right\}.
\]

**C. Proofs of Lemmas 2–3 in Section IV-A**

**Proof of Lemma 2**. Note that \( f^{cw}(t) = \max_{k \in [m]} (-C_k^T \hat{t}) \), that is, the maximum of \( m \) functions \(-C_1^T \hat{t}, \ldots, -C_m^T \hat{t}\). By a direct extension of Eq. (1) in [28] to allow multiple functions, we have
\[
f^{cw*}(s) = \min_{\lambda \in \Delta_m} f^*_\lambda(s),
\]
where \( f^*_\lambda = -(C \lambda)^T \hat{t} \). For each \( \lambda \in \Delta_m \), direct calculation yields
\[
f^*_\lambda(s) = \sup_{t \in \mathbb{R}^{m-1}} \{ st + (C \lambda)^T \hat{t} \} = \begin{cases} (C \lambda)_m, & \text{if } s_j \leq -(C \lambda)_j, j \in [m-1], \\ \infty, & \text{otherwise}. \end{cases}
\]
The desired result then follows.

**Proof of Lemma 3**. We need to show that for \( \eta \in \Delta_m \),
\[
H^{cw}(\eta) = \inf_{\lambda \in \Delta_m} \left\{ \sum_{j=1}^{m-1} \eta_j (C \lambda)_j + \eta_m (C \lambda)_m \right\}.
\]
Although this can be directly established, we give a proof based on Proposition 1. In fact, applying Proposition 1 with \( f = f^{cw} \) yields
\[
H^{cw}(\eta) = \inf_{s \in \text{dom}(f^{cw})} \left\{ \sum_{j=1}^{m-1} \eta_j (-s_j) + \eta_m f^{cw*}(s) \right\}.
\]
For each $s \in \text{dom}(f^{cw^*})$, there exists some $\lambda^s \in \Delta_m$ such that $s_j \leq -(C\lambda^s)_j$, $j \in [m-1]$ and hence by Lemma 2,

$$\sum_{j=1}^{m-1} \eta_j(-s_j) + \eta_m f^{cw^*}(s) \geq \sum_{j=1}^{m-1} \eta_j(C\lambda^s)_j + \eta_m(C\lambda^s)_m.$$

Therefore,

$$H^{cw}(\eta) \geq \inf_{\lambda \in \Delta_m} \left\{ \sum_{j=1}^{m-1} \eta_j(C\lambda)_j + \eta_m(C\lambda)_m \right\}.$$

The reverse inequality can be obtained by using the fact that for each $\lambda \in \Delta_m$, the vector $s^\lambda$ is contained in $\text{dom}(f^{cw^*})$ with $s^\lambda_j = -(C\lambda)_j$.

\end{proof}

D. Proofs of results related to $L^{cw^3}$ in Sections IV-A–IV-B

\begin{proof} [Proof of Proposition 4] We need to show that for $\eta \in \Delta_m$,

$$H^{cw}(\eta) = \inf_{\tau \in \mathbb{R}^{m-1}} \left\{ \sum_{j=1}^{m} \eta_j L^{cw^3}(j, \tau) \right\}.$$  \hfill (S3)

In fact, Lemma 3 implies that for $\eta \in \Delta_m$,

$$H^{cw}(\eta) = \inf_{\lambda \in \Delta_m} \left\{ \sum_{j=1}^{m} \eta_j L^{cw^3}(j, \lambda) \right\},$$

where by definition

$$L^{cw^3}(j, \lambda) = \begin{cases} (C\lambda)_j = c_{jm}\lambda_m + \sum_{k \in [m-1], k \neq j} c_{jk}\lambda_k, & \text{if } j \in [m-1], \\ (C\lambda)_m = \sum_{k \in [m-1]} c_{mk}\lambda_k, & \text{if } j = m, \end{cases}$$  \hfill (S4)

It suffices to show that

(i) $L^{cw^3}$ is an extension of $L^{cw^2}$ from $\Delta_m$ to $\mathbb{R}^{m-1}$, and

(ii) the minimum in (S3) is achieved at $\tau \in \mathbb{R}^{m-1}$ such that $\bar{\tau} = (\tau_1, \ldots, \tau_{m-1}, 1 - \sum_{k=1}^{m-1} \tau_k)^\tau$.

For the extension in (i), $L^{cw^2}(j, \lambda)$ is considered a function of $j$ and $(\lambda_1, \ldots, \lambda_{m-1})^\tau$, with $\lambda_m = 1 - \sum_{j=1}^{m-1} \lambda_j$, such that $\lambda \in \Delta_m$.

Result (i) is immediate by comparison of (23) with (S4). For any $\tau \in \mathbb{R}^{m-1}$ such that $\bar{\tau} \in \Delta_m$, we have $\tau_{k+} = \tau_k$ for $k \in [m-1]$, $\tau_{m+} = \sum_{k \in [m-1]} \tau_k$ for any $j \in [m-1]$, and hence $L^{cw^3}(j, \tau) = L^{cw^3}(j, \bar{\tau})$ for $j \in [m-1]$ or $j = m$.

For result (ii), we distinguish two cases. First, we show that for any $\tau \in \mathbb{R}^{m-1}$ with one or more negative components and $j \in [m]$,

$$L^{cw^3}(j, \tau') \leq L^{cw^3}(j, \tau),$$  \hfill (S5)

where $\tau'$ is obtained from $\tau$ by setting all negative components of $\tau$ to 0. In fact, by examining (23), we have $L^{cw^3}(m, \tau') = L^{cw^3}(m, \tau)$, because $\tau_{k+} = \tau_k$ for each $k \in [m-1]$. Moreover, $L^{cw^3}(j, \tau') \leq L^{cw^3}(j, \tau)$, by noting that $\tau_{j+} \leq \tau_j$ and $(\tau')_{m+} \leq \tau_m$ for $j \in [m-1]$, where $(\tau')_{m+}$ is defined by (24) with $\tau$ replaced by $\tau'$.

Second, we show that for any $\tau \in \mathbb{R}^{m-1}$ (i.e., all components of $\tau$ are nonnegative) with $\sum_{k=1}^{m-1} \tau_k > 1$ and $j \in [m]$,

$$L^{cw^3}(j, \tau'') \leq L^{cw^3}(j, \tau),$$  \hfill (S6)

where $\tau'' = (\tau''_1, \ldots, \tau''_{m-1})^\tau \in \mathbb{R}^{m-1}$ with $\tau''_{k+} = (\tau_k - b)_+$ and $b > 0$ chosen such that $\sum_{k=1}^{m-1} \tau''_k = 1$. This choice of $b$ exists, because $\sum_{k=1}^{m-1} (\tau_k - b)_+$ is continuous in $b$, attaining a value $> 1$ at $b = 0$ but a value $< 1$ at a sufficiently large $b$. By examining (23), we have $L^{cw^3}(m, \tau'') \leq L^{cw^3}(m, \tau)$ because $\tau''_k \leq \tau_k$ for each $k \in [m-1]$. Moreover, $L^{cw^3}(j, \tau'') \leq L^{cw^3}(j, \tau)$ for $j \in [m-1]$, by noting that $(\tau'')_m = 1 - \sum_{k \in [m-1]} \tau''_k = 0$, $(\tau'')_j = 1 - \sum_{k \in [m-1]} \tau''_k < 0$, and hence $(\tau''_m)_{m+} = (\tau''_j)_{m+} = 0$.

By combining the preceding two steps, the minimum in (S3) is achieved at some $\tau \in \mathbb{R}^{m-1}$ with $\sum_{k=1}^{m-1} \tau_k \leq 1$, that is, satisfying $\bar{\tau} \in \Delta_m$.

\end{proof}
Proof of Proposition 6(i). Note that $H_{L_{\omega}}(\eta) = H_{L_{\omega}^{\ast}}(\eta)$ by Proposition 4. Then inequality (29) is equivalent to

$$\frac{1}{m} R_{L_{\omega}}(\eta, \tau^1) + \frac{m-1}{m} H_{L_{\omega}}(\eta) \leq R_{L_{\omega}^{\ast}}(\eta, \tau),$$

(S7)

where $\tau^1 = (\tau_1, \ldots, \tau_{m-1}, 1 - \sum_{k=1}^{m-1} \tau_k)^T$. We distinguish three cases.

In the first case, suppose that $\tau \in \mathbb{R}^{m-1}$ with one or more negative components. We show that for any $\eta \in \Delta_m$,

$$R_{L_{\omega}}(\eta, \tau^1) = R_{L_{\omega}}(\eta, \tau), \quad R_{L_{\omega}^{\ast}}(\eta, \tau^1) \leq R_{L_{\omega}^{\ast}}(\eta, \tau),$$

where $\tau'$ is obtained from $\tau$ by setting all negative components of $\tau$ to 0. The second inequality follows from (S5) directly. To see the first inequality, note that a maximum component among $\tau^1 = (\tau_1, \ldots, \tau_{m-1}, \tau_m^1)^T$ must be positive; otherwise, $\tau_j \leq 0$ for each $j \in [m-1]$ and hence $\tau_{m-1}^1 = 1$, a contradiction. A maximum component among $\tau'' = (\tau_1', \ldots, \tau_{m-1}', \tau_m''^T)$ must also be positive. But for $j \in [m-1]$, we have $j' \tau_j' = j$ whenever $\tau_j'$ or $\tau_j$ is positive. Moreover, we have $\tau_m' = \tau_m''$, regardless of the signs of $\tau_m'$ and $\tau_m''$, because $\tau_k' = \tau_k''$ for each $k \in [m-1]$. Therefore, $\argmax_{j \in [m]} (\tau_j')$ and $\argmax_{j \in [m]} (\tau_j''$) can be set to be same, and the first inequality above holds.

In the second case, suppose that $\tau \in \mathbb{R}^{m-1}$ (i.e., all components of $\tau$ are nonnegative) with $\sum_{k=1}^{m-1} \tau_k > 1$. We show that for any $\eta \in \Delta_m$,

$$R_{L_{\omega}}(\eta, \tau^1) = R_{L_{\omega}}(\eta, \tau^1), \quad R_{L_{\omega}^{\ast}}(\eta, \tau^1) \leq R_{L_{\omega}^{\ast}}(\eta, \tau),$$

where $\tau'' = (\tau_1', \ldots, \tau_{m-1}', 1 - \sum_{j \in [m]} \tau_j')^T \in \mathbb{R}^{m-1}$ are defined as in Proof of Proposition 4. The second inequality follows from (S6) directly. To see the first equality, note that argmax$_{j \in [m]} (\tau_j)$, and argmax$_{j \in [m]} (\tau_j'')$ must lie in the set $\{m-1\}$ because $(\tau_j)''_{m-1} = 1 - \sum_{k \in [m-1]} \tau_k' = 0$, where $\tau_{m-1}^1 = 1 - \sum_{k \in [m-1]} \tau_k < 0$. But the first $m-1$ components of $\tau'$, $(\tau')_j' = (\tau_j - b)_+$ for $j \in [m-1]$, are ordered in the same way as those of $\tau$. Hence argmax$_{j \in [m]} (\tau_j')$ and argmax$_{j \in [m]} (\tau_j''$) can be set to be same, and the desired equality holds.

From the preceding discussion, it suffices to show (S7) in the third case where $\tau \in \mathbb{R}_+^{m-1}$ with $\sum_{k=1}^{m-1} \tau_k^1 \leq 1$, and hence $\tau^1 = (\tau_1, \ldots, \tau_{m-1}, 1 - \sum_{j \in [m-1]} \tau_j^1)^T \in \Delta_m$. Let $k = \argmin_{j \in [m]} \eta^T \tau_j$ and $l = \argmax_{j \in [m]} \eta^T \tau_j$. Then $\tau^1_1 \geq m^{1-1}$ and

$$R_{L_{\omega}}(\eta, \tau^1) = \eta^T C_l, \quad H_{L_{\omega}}(\eta) = \eta^T C_k.$$

Moreover, direct calculation yields

$$R_{L_{\omega}^{\ast}}(\eta, \tau) = \sum_{j=1}^{m} \eta^T C_j \tau_j^1 \geq \tau_1 \eta^T C_l + (1 - \tau_1) \eta^T C_k.$$  

The right-hand side above is non-decreasing in $\tau_1$ because $\eta^T C_l \geq \eta^T C_k$, and hence is no smaller than its value at $\tau_1^1 = m^{-1}$, that is, the left-hand side of (S7).

Proof of equivalence between $L_{\text{LLW}}^{\ast}$ and $L_{\text{LLW}}^{\ast 2}$. Suppose that $\tau_k = (1 + \gamma_k)/m$ for $k \in [m-1]$. Then it is immediate $L_{\text{LLW}}^{\ast 2}(m, \tau) = L_{\text{LLW}}(m, \gamma)/m$. Moreover, because $0 = \sum_{k=1}^{m} \gamma_k = \gamma_m + \sum_{k=1}^{m-1} (m \tau_k - 1)$, we have

$$1 + \gamma_m = m - m \sum_{k=1}^{m-1} \tau_k.$$  

Substituting this into the definition of $L_{\text{LLW}}$ and using $1 + \gamma_k = m \tau_k$ for $k \in [m-1]$ yields $L_{\text{LLW}}^{\ast 2}(j, \tau) = L_{\text{LLW}}(j, \tau)/m$ for $j \in [m-1]$.

Comparison between $L^{\text{zo}}$ and $L_{\text{LLW}}^{\ast 2}$. On one hand, the two losses $L^{\text{zo}}$ and $L_{\text{LLW}}^{\ast 2}$ share some similar properties. It can be verified that, similarly to $L^{\text{zo}}$, $L_{\text{LLW}}^{\ast 2}$ is a convex extension of $L^{\text{zo}}$ in (22), considered a function of $j$ and $(\lambda_1, \ldots, \lambda_{m-1})^T$ with $\lambda_m = 1 - \sum_{k=1}^{m-1} \lambda_k$. Moreover, by Proposition 4 and [12, Example 5], the losses $L^{\text{zo}}$ and $L_{\text{LLW}}^{\ast 2}$ lead to the same generalized entropy $H^{\text{zo}}$. Our result, Proposition 6, also yields a classification regret bound for $L^{\text{zo}}$, similar to that for $L_{\text{LLW}}^{\ast 2}$ in [12, Supplement Lemma 7.9]. On the other hand, there are interesting differences between $L^{\text{zo}}$ and $L_{\text{LLW}}^{\ast 2}$. While $L^{\text{zo}}(j, \tau)$ and $L_{\text{LLW}}^{\ast 2}(j, \tau)$ are aligned with $L^{\text{zo}}(j, \bar{\tau})$ for $\bar{\tau} = (\tau_1, \ldots, \tau_{m-1}, 1 - \sum_{k=1}^{m-1} \tau_k)^T \in \Delta_m$, the loss $L^{\text{zo}}$ stays uniformly lower than $L_{\text{LLW}}^{\ast 2}$,

$$0 \leq L^{\text{zo}}(j, \tau) \leq L_{\text{LLW}}^{\ast 2}(j, \tau), \quad j \in [m], \tau \in \mathbb{R}^{m-1},$$  

because $L^{\text{zo}}(j, \tau)$ can be written as $\sum_{k \in [m-1]} \tau_k + (1 - \tau_j - \sum_{k \in [m-1], k \neq j} \tau_k^+) + \sum_{j \in [m-1]} \tau_k^+$ for $j \in [m-1]$. Hence the loss $L^{\text{zo}}$ is a tighter convex extension than $L_{\text{LLW}}^{\ast 2}$. Another remarkable difference is that $L^{\text{zo}}(j, \tau)$ appears to be geometrically simpler with fewer non-differentiable ridges than $L_{\text{LLW}}^{\ast 2}(j, \tau)$ for $j \in [m-1]$. See Figure 3 for an illustration in the three-class setting. Further research is needed on whether the aforementioned differences can be translated into advantages in classification performance.
E. Proofs of results related to $L^{oz4}$ in Sections IV-A–IV-B

Proof of Proposition 5. We need to show that for $\eta \in \Delta_m$,

$$H^{oz}(\eta) = \inf_{\tau \in \mathbb{R}^{m-1}} \left\{ \sum_{j=1}^{m} \eta_j L^{oz4}(j, \tau) \right\}. \quad (S8)$$

Similarly as in the proof of Proposition 4, it suffices to show that

(i) $L^{oz4}$ is an extension of $L^{oz2}$ from $\Delta_m$ to $\mathbb{R}^{m-1}$, and

(ii) the minimum in (S8) is achieved at $\tau \in \mathbb{R}^{m-1}$ such that $\hat{\tau} \in \Delta_m$, where $\hat{\tau} = (\tau_1, \ldots, \tau_{m-1}, 1 - \sum_{k=1}^{m-1} \tau_k)^T$.

We use the following equivalent expressions for $S^{j^*}_{\tau}$:

$$S^{j^*}_{\tau} = \max \left\{ 0, \hat{\tau}_j - 1, \frac{-\hat{\tau}_j(m-1)}{m-1}, \frac{-\hat{\tau}_j(m-1) - \hat{\tau}_j(m-2)}{m-2}, \ldots, \frac{-\hat{\tau}_j(m-1) - \cdots - \hat{\tau}_j(2)}{2} \right\}. \quad (S9)$$

and, if $m^j_{\tau} \geq 1$,

$$S^{j^*}_{\tau} = \max \left\{ \hat{\tau}_j - 1, \max_{m^{j^*}_{\tau} \leq l \leq m-2} \frac{-\sum_{k=1}^{l} \hat{\tau}_j(m-k)}{m-l} \right\}, \quad (S10)$$

where $m^j_{\tau} = \# \{ k \in [m] : k \neq j, \hat{\tau}_k < 0 \}$. The first expression is immediate because $\sum_{k=1}^{m} \hat{\tau}_k = 1$. The second expression follows because $\{ \hat{\tau}_j(m-k) : 1 \leq k \leq m^j_{\tau} \}$, the smallest $m^j_{\tau}$ components among $\hat{\tau}$ excluding $\hat{\tau}_j$, are $\{ \hat{\tau}_k : \hat{\tau}_k \leq 0, k \neq j, k \in [m] \}$, and $-\sum_{k=1}^{l} \hat{\tau}_j(m-k)/(m-l)$ is nonnegative and nondecreasing in $1 \leq l \leq m^j_{\tau}$.

Result (i) can be directly verified. For any $\tau \in \mathbb{R}^{m-1}$ such that $\hat{\tau} \in \Delta_m$, we have $S^{j^*}_{\tau} = 0$ for $j \in [m]$ by examining the expression (S9), and hence $L^{oz4}(j, \tau) = L^{oz2}(j, \hat{\tau}) = 1 - \hat{\tau}_j$ for $j \in [m]$ by the definitions (25) and (22).

For result (ii), we show that for any $\tau \in \mathbb{R}^{m-1}$ with one or more negative components in $\hat{\tau}$, there exists $\tau' = (\tau'_1, \ldots, \tau'_{m-1})^T \in \mathbb{R}^{m-1}$ such that $\tau'' = (\tau''_1, \ldots, \tau''_{m-1}, 1 - \sum_{k=1}^{m-1} \tau'_k)^T \in \Delta_m$ and for $j \in [m]$,

$$L^{oz4}(j, \tau') \leq L^{oz4}(j, \tau). \quad (S11)$$

Then the minimum in (S8) is achieved at some $\tau \in \mathbb{R}^{m-1}$ with $\hat{\tau} \in \Delta_m$.

First, let $\tau'' = (\tau''_1, \ldots, \tau''_{m-1})^T$ and $\tau'' = (\tau''_1, \ldots, \tau''_{m-1})^T$ with

$$\hat{\tau}_j = \begin{cases} \hat{\tau}_j - b, & \text{if } \hat{\tau}_j \geq 0, \\ \hat{\tau}_j + \frac{m^+_{\tau}}{m^-_{\tau}} b, & \text{if } \hat{\tau}_j < 0, \end{cases}$$

for $j \in [m]$, where $m^-_{\tau} = \# \{ k \in [m] : \hat{\tau}_k < 0 \} \geq 1$, $m^+_{\tau} = \# \{ k \in [m] : \hat{\tau}_k \geq 0 \} = m - m^-_{\tau}$, and $b > 0$ is determined such that $\max \{ \hat{\tau}_k + (m^+_{\tau}/m^-_{\tau}) b : k \in [m], \hat{\tau}_k < 0 \}$ equals $\min \{ 0, \min \{ \hat{\tau}_k - b : k \in [m], \hat{\tau}_k \geq 0 \} \}$. Then the following properties hold:

(a) $\sum_{k=1}^{m} \hat{\tau}_k = \sum_{k=1}^{m} \hat{\tau}_k = 1$.

(b) The ordering among components of $\tau''$ remains the same as that among $\hat{\tau}$.

(c) If $\hat{\tau}_k \leq 0$ then $\hat{\tau}'_k \leq 0$ for $k \in [m]$.

It can be shown that $L^{oz4}(j, \tau'') \leq L^{oz4}(j, \tau)$ for $j \in [m]$, depending on the sign of $\hat{\tau}_j$.

- Suppose $\hat{\tau}_j \geq 0$. Then $m^-_{\tau} \leq m^j_{\tau}$ by definition, and $m^j_{\tau} \leq m^j_{\tau''}$ by property (c). For $m^-_{\tau} \leq l \leq m - 2$, by property (b),

$$1 - \hat{\tau}_j'' = \frac{\sum_{k=1}^{l} \hat{\tau}_j(m-k)}{m-l} = \frac{1 - \hat{\tau}_j - \sum_{k=1}^{l} \hat{\tau}_j(m-k)}{m-l} = \frac{-\sum_{k=1}^{l} \hat{\tau}_j(m-k)}{m-l} = \frac{-\sum_{k=1}^{l} \hat{\tau}_j(m-k)}{m-l} = \frac{b - \frac{m^+_{\tau}}{m^-_{\tau}} b (l - m^-_{\tau})}{m-l} = 0.$$

By combining the preceding properties with (S10),

$$L^{oz4}(j, \tau') = \max \left\{ 0, 1 - \hat{\tau}_j + \max_{m^j_{\tau} \leq l \leq m-2} \frac{-\sum_{k=1}^{l} \hat{\tau}_j(m-k)}{m-l} \right\}, \quad (S12)$$

$$L^{oz4}(j, \tau'') = \max \left\{ 0, 1 - \hat{\tau}_j'' + \max_{m^j_{\tau} \leq l \leq m-2} \frac{-\sum_{k=1}^{l} \hat{\tau}_j(m-k)}{m-l} \right\}. \quad (S13)$$
we see that \( L^{\text{out}}(j, \tau''') = L^{\text{out}}(j, \tau) \).

- Suppose \( \tilde{\tau}_j < 0 \) and \( m_{\tilde{\tau}} \geq 2 \). Then \( 1 \leq m_{\tilde{\tau}} - 1 \leq m_{\tau}^{(j)} \) by definition, and \( m_{\tau}^{(j)} \leq m_{\tilde{\tau}}^{(j)} \) by property (c). For \( m_{\tilde{\tau}} - 1 \leq l \leq m - 2 \), by property (b),

\[
1 - \tilde{\tau}_j'' = \sum_{k=1}^{l} \frac{\tilde{\tau}_j''(m-k)}{m-l} - \sum_{k=1}^{l} \frac{\tilde{\tau}_j(m-k)}{m-l} = -(\tilde{\tau}_j'' - \tilde{\tau}_j) = -\frac{\sum_{k=1}^{l} \tilde{\tau}_j''(m-k) - \tilde{\tau}_j(m-k)}{m-l} = -\frac{m^+ b - (m^+ b - (l - m_{\tilde{\tau}} + 1)b}{m-l} = -\left( \frac{m^+ b}{m_{\tilde{\tau}}} + 1 \right) \frac{b}{m-l} < 0.
\]

Hence \( L^{\text{out}}(j, \tau''') \leq L^{\text{out}}(j, \tau) \) by the expressions (S12)–(S13).

- Suppose \( \tilde{\tau}_j < 0 \) and \( m_{\tilde{\tau}} = 1 \). Then \( \tilde{\tau}_k \geq 0 \) for \( k \in [m] \) and \( k \neq j \), and hence \( L^{\text{out}}(j, \tau) = 1 - \tilde{\tau}_j \) by (S9). Moreover, \( \tilde{\tau}_j'' \leq 0 \) and \( \tilde{\tau}_k'' \geq -b \) for \( k \in [m] \) and \( k \neq j \), and hence by (S9) applied to \( \tau''' \),

\[
L^{\text{out}}(j, \tau''') \leq 1 - \tilde{\tau}_j'' + (m - 2)b/2 = 1 - \tilde{\tau}_j - (m - 1)b + (m - 2)b/2 \leq L^{\text{out}}(j, \tau).
\]

If \( \tilde{\tau}'' \) has no negative components, then \( \tilde{\tau}'' \in \Delta_m \) and (S11) holds with \( \tau' = \tau'' \). Otherwise, the preceding mapping from \( \tilde{\tau} \) to \( \tilde{\tau}'' \), denoted as \( \mathcal{F}(\cdot) \), can be iteratively applied. Let \( \tilde{\tau}^{(0)} = \tilde{\tau} \) and for \( i = 1, 2, \ldots \), if \( \tilde{\tau}^{(i-1)} \) has one or more negative components, then let \( \tilde{\tau}^{(i)} = \mathcal{F}(\tilde{\tau}^{(i-1)}) \). It suffices to show that this process necessarily terminates after finite steps. The final iteration \( \tilde{\tau}^{(i)} \) has no negative components and hence \( \tilde{\tau}^{(i)} \in \Delta_m \). The first \( m - 1 \) components of \( \tilde{\tau}^{(i)} \) can be taken as the desired \( \tau'' \) in (S11).

Denote the set of \( m_{\tilde{\tau}} \) negative components of \( \tilde{\tau} \) (or equivalently the \( m_{\tilde{\tau}} \) smallest components of \( \tilde{\tau} \)) as \( 0 > \tilde{\tau}_{j1} \geq \cdots \geq \tilde{\tau}_{jm_{\tilde{\tau}}} \). By property (b), \( \tilde{\tau}_{j1}^{(i)} \geq \cdots \geq \tilde{\tau}_{jm_{\tilde{\tau}}}^{(i)} \) remain the smallest \( m_{\tilde{\tau}} \) components of \( \tilde{\tau}^{(i)} \) for each \( i \geq 1 \). It suffices to show that \( \tilde{\tau}_{j1}^{(i)} \) becomes 0 for a certain finite \( i \geq 1 \). Then the number of negative components of \( \tilde{\tau}^{(i)} \) decreases to \( m_{\tilde{\tau}} - 1 \) or smaller. Applying this argument repeatedly shows that \( \tilde{\tau}^{(i)}_{j1} \) necessarily becomes 0 (or equivalently the number of negative components of \( \tilde{\tau}^{(i)} \) decreases to 0) for a certain finite \( i \), hence proving the finite-termination of the iterations.

Return to the mapping from \( \tilde{\tau} \) to \( \tilde{\tau}'' = \tilde{\tau}^{(i)} \). By the choice of \( b \), \( \tilde{\tau}_{j1}^{(i)} \) either equals 0 or \( \tilde{\tau}_{j1}^{(i)} \) for some \( k \in [m] \) such that \( \tilde{\tau}_k \geq 0 \) but \( \tilde{\tau}_k'' < 0 \). In the latter case, the number of negative components of \( \tilde{\tau}'' \) increases to at least \( m_{\tilde{\tau}} - 1 \). Applying this argument repeatedly shows that \( \tilde{\tau}_{j1}^{(i)} \) necessarily equals 0 for some \( i \leq m - m_{\tilde{\tau}} \). Otherwise, the number of negative components of \( \tilde{\tau}_{j1}^{(i)} \) would be \( m \), which contradicts the fact that all the components of \( \tilde{\tau}^{(i)} \) sum up to 1, by property (a).

**Proof of Proposition 6(ii).** Because \( L^{\text{out}} \) induces the same generalized entropy as the zero-one loss by Proposition 5, the result can be obtained from Proposition 8. Alternately, the following gives a direct proof, building on the proof of Proposition 5.

The main steps of the proof are similar as in the proof of Proposition 6(i). First, note that \( H_{L=\eta}(\eta) = H_{L=\eta,\tau}(\eta) \) by Proposition 5. Then inequality (30) is equivalent to

\[
\frac{1}{m} R_{L=\eta}(\eta, \tilde{\tau}) + \frac{m-1}{m} H_{L=\eta}(\eta) \leq R_{L=\eta,\tau}(\eta, \tau).
\]

Second, for any \( \tau \in \mathbb{R}^{m-1} \) with one or more negative components in \( \tilde{\tau} \), there exists \( \tau' = (\tau'_1, \ldots, \tau'_{m-1}) \in \mathbb{R}^{m-1} \) such that \( \tilde{\tau}' = (\tau'_1, \ldots, \tau'_{m-1}, 1 - \sum_{k=1}^{m-1} \tau'_k) \in \Delta_m \) and for any \( \eta \in \Delta_m \),

\[
R_{L=\eta}(\eta, \tilde{\tau}) = R_{L=\eta}(\eta, \tilde{\tau}') = R_{L=\eta,\tau}(\eta, \tau'), \quad R_{L=\eta,\tau}(\eta, \tau') \leq R_{L=\eta,\tau}(\eta, \tau).
\]

The second equality follows from (S11) directly. Moreover, in the proof of (S11), \( \tilde{\tau}' \) is obtained from \( \tilde{\tau} \) by iteratively applying the mapping \( \mathcal{F}(\cdot) \) from \( \tilde{\tau} \) to \( \tilde{\tau}'' \). By property (b), the ordering among components of \( \tilde{\tau} \) is preserved (although not strictly preserved) under the mapping. Hence argmax\( j \in [m] \tilde{\tau}'_j \) and, through iterations, argmax\( j \in [m] \tilde{\tau}_j \) can all be set to be same as argmax\( j \in [m] \tilde{\tau}_j \). The first equality holds.

Finally, it suffices to show (S14) for \( \tau \in \mathbb{R}^{m-1} \) with \( \tilde{\tau} \in \Delta_m \). Let \( k = \text{argmax}_{j \in [m]} \eta_j \) and \( l = \text{argmax}_{j \in [m]} \tilde{\tau}_j \). Then \( \tilde{\tau}_1 \geq m^{-1} \). Direct calculation yields

\[
R_{L=\eta,\tau}(\eta, \tau) = \sum_{j \in [m]} \eta_j (1 - \tau_j) = 1 - \sum_{j \in [m]} \eta_j \tilde{\tau}_j.
\]
and
\[
\frac{1}{m} R_{L_n^m}(\eta, \tilde{\tau}) + \frac{m-1}{m} H_{L_n^m}(\eta) = \frac{1}{m}(1 - \eta) + \frac{m-1}{m}(1 - \eta_k) = 1 - \left(\frac{1}{m} \eta + \frac{m-1}{m} \eta_k\right).
\]

Inequality (S14) can be obtained by comparing the above two expressions: \(\sum_{j \in [m]} \eta_j \tilde{\tau}_j\) is upper-bounded by \(\eta_1 \tilde{\tau}_1 + \eta_k (1 - \tilde{\tau}_1)\), which is nonincreasing in \(\tilde{\tau}_1\) with \(\eta_1 \leq \eta_k\), and hence is no greater than its value at \(\tilde{\tau}_1 = m^{-1}\).

**Comparison between \(L^{\text{iso}}\) and \(L^{\text{DKR2}}\).** The comparison is similar to that between \(L^{\text{iso}}\) and \(L^{\text{LLW2}}\). On one hand, it can be verified that \(L^{\text{DKR2}}\) is a convex extension of \(L^{\text{iso}}\), similarly to \(L^{\text{iso}}\), and by Proposition 5 and [12, Example 3], both \(L^{\text{iso}}\) and \(L^{\text{DKR2}}\) lead to the same generalized entropy \(H^{\text{iso}}\). Our result, Proposition 6, also gives a classification regret bound for \(L^{\text{iso}}\), similar to that for \(L^{\text{DKR2}}\) in [12, Proposition 5]. On the other hand, there are interesting differences between \(L^{\text{iso}}\) and \(L^{\text{DKR2}}\).

While \(L^{\text{iso}}(j, \tau)\) and \(L^{\text{DKR2}}(j, \tau)\) coincide with \(L^{\text{iso}}(j, \tilde{\tau})\) provided \(\tilde{\tau} \in \Delta_m\), the loss \(L^{\text{iso}}\) gives a tighter convex extension than \(L^{\text{DKR2}}\):
\[
0 \leq L^{\text{iso}}(j, \tau) \leq L^{\text{DKR2}}(j, \tau), \quad j \in [m], \tau \in \mathbb{R}^{m-1},
\]

because \(S^{(j)}_{\tilde{\tau}} \leq S^\tau\) for \(j \in [m]\), with \(S^{(j)}\) being the maximum of \(m\) numbers which are respectively no greater than those in the definition of \(S^\tau\). Moreover, \(L^{\text{iso}}(j, \tau)\) appears to be geometrically simpler with fewer non-differentiable ridges than \(L^{\text{DKR2}}(j, \tau)\) for \(j \in [m]\). See Figure 4 for an illustration in the three-class setting.

**F. Proofs of results in Section IV-C**

**Proof of Proposition 7.** Denote by \(v_1, \ldots, v_m\) the vertices of \(S^{\text{iso}}\), where \(v_j \in \mathbb{R}^m\) has \(j\)th component 0 and the remaining components 1. For two vectors \(x, y \in \mathbb{R}^m\), write \(x \preceq y\) if \(x_j \leq y_j\) for \(j \in [m]\).

First, suppose that the inclusion property (34) holds. Then by (33), we have for \(\eta \in \Delta_m\),
\[
H_{L}(\eta) = \inf_{z \in S_L} \eta^T z = \inf_{z \in S^{\text{iso}}} \eta^T z = \inf_{\lambda \in \Delta_m} \sum_{j \in [m]} \lambda_j \eta_j v_j = \inf_{\lambda \in \Delta_m} \sum_{j \in [m]} \lambda_j (1 - \eta_j) = 1 - \max_{j \in [m]} \eta_j.
\]

The equality \(\inf_{z \in S_L} \eta^T z = \inf_{z \in S^{\text{iso}}} \eta^T z\) appears to be geometrically simpler with fewer non-differentiable ridges than \(L^{\text{DKR2}}(j, \tau)\) for \(j \in [m]\). See Figure 4 for an illustration in the three-class setting.

The second case, \(S^{\text{iso}} \subset S_L\), is similar to the first case, except that \(S^{\text{iso}}\) is closed and \(S_L\) is not. Then there exists a point \(x \in S_L\) but \(x \not\in S^{\text{iso}}\). The set \(S^{\text{iso}}\) is easily seen to be closed and convex. By the support hyperplane theorem, there exists a hyperplane which strictly separates \(x\) and \(S^{\text{iso}}\), that is, there exists some \(\eta \in \mathbb{R}^m\) and \(b \in \mathbb{R}\) such that \(\eta^T x < b\), but \(\eta^T z > b\) for all \(z \in S^{\text{iso}}\). The coefficient vector \(\eta\) must be nonzero, \(\eta \neq 0\), and have all nonnegative components, \(\eta \in \mathbb{R}^m_+\). Otherwise, suppose that, for example, \(\eta_1 < 0\) and fix some point \(\hat{z} \in S^{\text{iso}}\). Define \(\tilde{z} = \hat{z} + ke_1\), where \(e_1 = (1, 0, \ldots, 0)^T\). Then \(\tilde{z} \in S^{\text{iso}}\) for all \(k\), but \(\eta^T \tilde{z} = \eta^T \hat{z} + k\eta_1 \rightarrow -\infty\) as \(k \rightarrow \infty\), which contradicts the fact that \(\eta^T z > b\) for all \(z \in S^{\text{iso}}\). Hence \(\eta\) can be normalized such that \(\eta \in \Delta_m\). But then
\[
\inf_{z \in S_L} \eta^T z \leq \eta^T \tilde{x} < \inf_{z \in S^{\text{iso}}} \eta^T z = \inf_{z \in S^{\text{iso}}} \eta^T z = 1 - \max_{k \in [m]} \eta_k,
\]
a contradiction to the assumption that \(H_{L}(\eta) = 1 - \max_{j \in [m]} \eta_j\).

In the second case, \(S^{\text{iso}} \subset S_L\). Then there exists a vertex of \(S^{\text{iso}}\) which is not contained in \(S_L\); otherwise \(S^{\text{iso}} \subset S_L\) by the convexity of \(S_L\). Without loss of generality, assume that \(v_1 \not\in S_L\). Then \(v_1 \not\in S_L + \mathbb{R}_{+}^m\). Otherwise, there exist some \(x \in S_L\) and \(y (\neq 0) \in \mathbb{R}_+^m\) such that \(v_1 = x + y\). Then \(\sum_{j \in [m]} (x_j + y_j) = m - 1\), which contradicts the fact that \(\sum_{j \in [m]} (x_j + y_j) \geq \sum_{j \in [m]} x_j = m - 1\). The second equality holds because \(\inf_{z \in S_L} \sum_{j \in [m]} z_j = m - 1\) by the assumption that \(H_{L}(1/m) = \inf_{z \in S_L}(1/m)z = 1 - 1/m\) for \(1/m = (1/m, \ldots, 1/m)^T \in \Delta_m\). The set \(S_L + \mathbb{R}_{+}^m\) is closed and convex. By the support hyperplane theorem, there exists a hyperplane which strictly separates \(v_1\) and \(S_L + \mathbb{R}_{+}^m\), that is, there exists some \(\eta \in \mathbb{R}^m\) and \(b \in \mathbb{R}\) such that \(\eta^T v_1 < b\), but \(\eta^T z > b\) for all \(z \in S_L + \mathbb{R}_{+}^m\). Similarly as in the first case, \(\eta\) must be nonzero, \(\eta \neq 0\), and have all nonnegative components, \(\eta \in \mathbb{R}^m_+\). Hence \(\eta\) can be normalized such that \(\eta \in \Delta_m\). But then
\[
\inf_{z \in S_L} \eta^T z \geq \inf_{z \in S_L + \mathbb{R}_{+}^m} \eta^T z \geq \inf_{z \in S^{\text{iso}}} \eta^T z = 1 - \max_{k \in [m]} \eta_k,
\]
again a contradiction to the assumption that \( H_L(\eta) = 1 - \max_{j \in [m]} \eta_j \).

Combining the preceding two cases shows that (34) holds as desired. \( \blacksquare \)

**Proof of Proposition 8.** Note that \( H_L(\eta) = H_{L^\tau}(\eta) \) by assumption. Inequality (35) reduces to

\[
\frac{1}{m} R_{L^\tau}(\eta, \sigma_L(\gamma)) + \frac{1}{m} H_{L^\tau}(\eta, \gamma) \leq R_L(\eta, \gamma).
\]

By definition, \( \sigma_L(\gamma) = (-L(1, \gamma), \ldots, -L(m, \gamma))^T \). The preceding inequality can be stated such that for \( \eta \in \Delta_m \) and 
\( z = (L(1, \gamma), \ldots, L(m, \gamma))^T \in \mathcal{R}_L \),

\[
\frac{1}{m} (1 - \eta_k) + \frac{m - 1}{m} (1 - \eta_k) \leq \eta^T z, \quad \text{(S15)}
\]

\( l = \arg\min_{j \in [m]} z_j \), and \( k = \arg\max_{j \in [m]} \eta_j \). In the following, we show that (S15) holds for \( \eta \in \Delta_m \) and \( z \in \mathcal{S}_L \). The notation \( \preceq \) is used as in the proof of Proposition 7.

First, we show that for any \( z \in \mathcal{S}_L \), there exists some \( \tilde{z} \in \mathcal{S}^{zo} \) such that

\[
\tilde{z} \preceq z, \quad \arg\min_{j \in [m]} \tilde{z}_j = \arg\min_{j \in [m]} z_j.
\]

which means that \arg\min_{j \in [m]} \tilde{z}_j \) can be set to be same as \arg\min_{j \in [m]} z_j \). Because \( \mathcal{S}_L \subset \mathcal{S}^{zo} \) by Proposition 7, it suffices to show that for any \( z \in \mathcal{S}^{zo} \), there exists \( \tilde{z} \in \mathcal{S}^{zo} \) such that (S16) holds. Without loss of generality, assume that \( z_1 \geq z_2 \geq \cdots \geq z_m \). Let \( b = \sup \{ b' \geq 0 : z - b' e_m \in \mathcal{S}^{zo} \} \), such that \( z' = z - b e_m \in \partial \mathcal{S}^{zo} \). Then \( z' \preceq z \) and the \( m \)th component of \( z, z'_m \), remains a minimum component of \( z' \). By the definition of \( \mathcal{S}^{zo} \), there exists some \( \tilde{z} \in \mathcal{S}^{zo} \) satisfying \( \tilde{z} \preceq z' \). For any such point \( \tilde{z} \), we have

(i) \( \tilde{z}_m = z'_m \)

(ii) \( \tilde{z}_j \geq z'_m \) for \( j \in [m - 1] \),

which then imply that (S16) is satisfied. Property (i) follows because if \( \tilde{z}_m < z'_m \), then by the definition of \( \mathcal{S}^{zo} \), \( z' - (z'_m - \tilde{z}_m) e_m = \tilde{z} + (z'_m - \tilde{z}_m) e_m \in \mathcal{S}^{zo} \), but this contradicts the definition of \( b \). To show property (ii), suppose that there exists \( \tilde{z} \in \mathcal{S}^{zo} \) such that \( \tilde{z} \preceq z \) and \( \tilde{z}_j < z'_m \) for some \( j \in [m - 1] \). Let \( \tilde{z} = (\tilde{z}_1, \ldots, \tilde{z}_{j-1}, \tilde{z}_j, \tilde{z}_{j+1}, \ldots, \tilde{z}_m)^T \) by exchanging the \( j \)th and \( m \)th components of \( \tilde{z} \). Then \( \tilde{z} \in \mathcal{S}^{zo} \) by symmetry of \( \mathcal{S}^o \). Moreover, \( \tilde{z} \preceq z' \), because \( \tilde{z}_j = \tilde{z}_m = z'_m \leq z'_j \) by property (i) and \( \tilde{z}_m = z'_m \) for \( j \in [m - 1] \). Then \( \tilde{z} \) must also satisfy property (i), i.e., \( \tilde{z}_m = z'_m \), a contradiction.

By the preceding result, it suffices to show that (S15) holds for any \( \eta \in \Delta_m \) and \( \tilde{z} \in \mathcal{S}^{zo} \). This can be obtained as follows:

\[
\eta^T \tilde{z} = 1 - \sum_{j \in [m]} \eta_j (1 - \tilde{z}_j) = 1 - \eta_l (1 - \tilde{z}_l) - \sum_{j \neq l} \eta_j (1 - \tilde{z}_j) \\
\geq 1 - \eta_l (1 - \tilde{z}_l) - \sum_{j \neq l} \eta_j (1 - \tilde{z}_j) = 1 - \eta_l + (\eta_l - \eta_k) \tilde{z}_l, \\
\geq 1 - \eta_l + \frac{m - 1}{m} (\eta_l - \eta_k).
\]

The second line above uses the fact that \( \eta_k = \max_{j \in [m]} \eta_j, \) \( 0 \leq \tilde{z}_j \leq 1 \) for \( j \in [m] \), and \( \sum_{j \in [m]} \tilde{z}_j = m = 1.1 \). The last line holds because \( 1 - \eta_l + (\eta_l - \eta_k) \tilde{z}_l \) is non-increasing in \( \tilde{z}_l \) with \( \eta_l \leq \eta_k \), and hence is no smaller than its value at \( \tilde{z}_l = \frac{m - 1}{m} \), where \( \tilde{z}_l = \min_{j \in [m]} \tilde{z}_j \leq \frac{m - 1}{m} \) with \( \sum_{j \in [m]} \tilde{z}_j = m = 1.1 \).

**Simplification of prediction mapping \( \sigma_L \).** We show that for each of the four losses, \( L^{LLW^2}, L^{DKR^2}, L^{zo3}, \) and \( L^{zo4} \), the prediction mapping \( \sigma_L \) in Proposition 8 is monotonically related to that in the corresponding regret bound discussed in Section IV-B.

The loss \( L^{LLW^2} \) can be written as

\[
L^{LLW^2}(j, \tau) = \sum_{k \in [m], k \neq j} \tilde{\tau}_{k+} = -\tilde{\tau}_j + \sum_{k \in [m]} \tilde{\tau}_{k+}, \quad j \in [m],
\]

where \( \tilde{\tau} = (\tau_1, \ldots, \tau_{m-1}, 1 - \sum_{k=1}^{m-1} \tau_k)^T \). Hence if \( \tilde{\tau}_j \leq \tilde{\tau}_k \), then \( L^{LLW^2}(j, \tau) \geq L^{LLW^2}(k, \tau) \). Similarly, it is easily seen that if \( \tilde{\tau}_j \leq \tilde{\tau}_k \), then \( L^{DKR^2}(j, \tau) \geq L^{DKR^2}(k, \tau) \).

The loss \( L^{zo3} \) can be written as

\[
L^{zo3}(j, \tau) = \begin{cases} 
\max \{1 - \tau_j, 1 - \tau_j^L - \tau_j^+ \}, & \text{if } j \in [m - 1], \\
1 - \tau_j^+, & \text{if } j = m,
\end{cases}
\]

where \( \tau^L = (\tau_1, \ldots, \tau_{m-1}, 1 - \sum_{k=1}^{m-1} \tau_k)^T \). For \( j, k \in [m - 1] \), if \( \tau_j \leq \tau_k \), then \( L^{zo3}(j, \tau) \geq L^{zo3}(k, \tau) \) trivially. For \( j \in [m - 1] \), if \( \tau_j \leq \tau_m \), then \( L^{zo3}(j, \tau) \geq 1 - \tau_j \geq L^{zo3}(m, \tau) \), and if \( \tau_j \geq \tau_m \), then \( L^{zo3}(j, \tau) \leq 1 - \tau_m = L^{zo3}(m, \tau) \).
The loss $L^{z4}$ can be written as

$$L^{x4}(j, \tau) = \max \left\{ 0, \frac{1 + \tilde{r}_j(j) - \tilde{r}_j}{2}, \ldots, m - 2 + (\tilde{r}_j(j) - \tilde{r}_j) + \ldots + (\tilde{r}_j(m-2) - \tilde{r}_j), \ldots, m - 1 + (\tilde{r}_j(j) - \tilde{r}_j) + \ldots + (\tilde{r}_j(m-1) - \tilde{r}_j) \right\}, \quad j \in [m],$$

where $\tilde{r}_j(1) \geq \ldots \geq \tilde{r}_j(m-1)$ are the sorted components of $\tilde{r}$ excluding $\tilde{r}_j$. Without loss of generality, assume that $\tilde{r}_1 \geq \ldots \geq \tilde{r}_m$. Then

$$L^{x4}(j, \tau) = \max \left\{ 0, \frac{1 + \tilde{r}_j(j) - \tilde{r}_j}{2}, \ldots, j - 1 + (\tilde{r}_j(j) - \tilde{r}_j) + \ldots + (\tilde{r}_j(j-1) - \tilde{r}_j), \ldots, m - 1 + (\tilde{r}_j(j) - \tilde{r}_j) + \ldots + (\tilde{r}_j(m-1) - \tilde{r}_j) \right\}.$$

Denote the $i$th term in the curly brackets above as $\ell_i(j, \tau)$ for $i = 1, \ldots, m$, that is, $\ell_1(j, \tau) = 0$ and for $i = 2, \ldots, m$,

$$\ell_i(j, \tau) = \begin{cases} i - 1 + \sum_{h=1}^{i-1} (\tilde{r}_h - \tilde{r}_j), & \text{if } i < j, \\ i - 1 + \sum_{h=1}^{i} (\tilde{r}_h - \tilde{r}_j), & \text{if } i \geq j. \end{cases}$$

For $j > k$ with $\tilde{r}_j \leq \tilde{r}_k$, if $i < k$ or $i \geq j$, then $\ell_i(j, \tau) \geq \ell_i(k, \tau)$, and if $k \leq i < j$, then

$$\ell_i(j, \tau) = \frac{i - 1 + \sum_{h=1}^{i-1} (\tilde{r}_h - \tilde{r}_j)}{i} \geq \frac{i - 1 + \sum_{h=1}^{i} (\tilde{r}_h - \tilde{r}_k)}{i} = \ell_i(k, \tau),$$

where the second inequality follows because $\tilde{r}_i \leq \tilde{r}_k$. In summary, if $\tilde{r}_j \leq \tilde{r}_k$, then $\ell_i(j, \tau) \geq \ell_i(k, \tau)$ for $i = 1, \ldots, m$, and hence $L^{x4}(j, \tau) \geq L^{x4}(k, \tau)$.

**G. Proofs of results in Section V-A**

**Proof of equation (39).** By manipulating the summation, we have

$$L^{\text{loss}}_{f_0}(j, q) = \sum_{l \in [m], k \neq l} \left[ -\mathbb{1}_k(j) \partial f_0 \left( \frac{q_k}{q_l} \right) + \mathbb{1}_l(j) \left\{ \frac{q_k}{q_l} \partial f_0 \left( \frac{q_k}{q_l} \right) - f_0 \left( \frac{q_k}{q_l} \right) \right\} \right]$$

$$= \sum_{l \in [m]} \sum_{k \in [m], k \neq l} \left[ -\mathbb{1}_k(j) \partial f_0 \left( \frac{q_k}{q_l} \right) \right] + \sum_{k \in [m]} \sum_{l \in [m], l \neq k} \mathbb{1}_l(j) \left\{ \frac{q_k}{q_l} \partial f_0 \left( \frac{q_k}{q_l} \right) - f_0 \left( \frac{q_k}{q_l} \right) \right\}$$

$$= \sum_{l \in [m], j \neq l} \left\{ -\partial f_0 \left( \frac{q_j}{q_l} \right) \right\} + \sum_{k \in [m], j \neq k} \left\{ \frac{q_k}{q_j} \partial f_0 \left( \frac{q_k}{q_j} \right) - f_0 \left( \frac{q_k}{q_j} \right) \right\},$$

which yields the desired result.

**Convexity of two-class composite losses.** Consider a logistic link $q^{h_0} = (q_1^{h_0}, q_2^{h_0})^T$, where $q_1^{h_0} = \{1 + \exp(-h_0)\}^{-1}$ or equivalently $q_1^{h_0}/q_2^{h_0} = \exp(h_0)$. Then it can be easily shown that the three composite losses, $L_e(j, q^{h_0})$, $L_c(j, q^{h_0})$, and $L_e(j, q^{h_0})$, are convex in $h_0$, with the following gradients:

$$\frac{d}{dh_0} L_e(j, q^{h_0}) = - \left\{ \mathbb{1}_1(j) - q_1^{h_0} \right\},$$

$$\frac{d}{dh_0} L_c(j, q^{h_0}) = - \left\{ \mathbb{1}_1(j) - q_1^{h_0} \right\} (q_2^{h_0} q_1^{h_0})^{-1/2},$$

$$\frac{d}{dh_0} L_e(j, q^{h_0}) = - \left\{ \mathbb{1}_1(j)/q_1^{h_0} - 1 \right\}/2.$$
Proof of Proposition 9. The scoring rules are obtained directly from Proposition 3. First, we show the three limits of $H_\beta$ for $\beta = 0, 1, \infty$.

(i) Rewrite $H_\beta(q)$ as
\[
H_\beta(q) = \frac{\exp\left(\frac{1}{\beta} \log(1 + \frac{\sum_{j=1}^m (q_j^\beta - 1)}{m})\right) - m^{-\frac{1}{\beta}}}{m^{-1} - m^{-\frac{1}{\beta}}}.
\]
Using $\log(1 + x)/x \to 1$ as $x \to 0$, we have
\[
\lim_{\beta \to 0^+} H_\beta(q) = \lim_{\beta \to 0^+} \frac{\exp\left(\frac{1}{\beta} \log(\sum_{j=1}^m q_j^\beta) + \frac{1}{\beta} \log m\right)}{m} = m \left(\prod_{j=1}^m q_j\right)^{\frac{1}{\beta}},
\]
where the last step holds because $\lim_{\beta \to 0^+} q_j^\beta - 1/\beta = \log q_j$ by L'Hopital's rule.

(iii) Rewrite $H_\beta(q)$ as
\[
H_\beta(q) = \frac{\exp\left(\frac{1}{\beta} \log(\sum_{j=1}^m q_j^\beta)\right) - 1}{\exp\left(\frac{1}{\beta} - 1\right) \log m}.
\]
Using $(e^x - 1)/x \to 1$ as $x \to 0$, we obtain
\[
\lim_{\beta \to 1} H_\beta(q) = \lim_{\beta \to 1} \frac{\log(\sum_{j=1}^m q_j^\beta)}{(1 - \beta) \log m}.
\]
Applying L'Hopital's rule yields
\[
\lim_{\beta \to 1} H_\beta(q) = \lim_{\beta \to 1} \frac{-\sum_{j=1}^m q_j^\beta \log q_j}{(\log m)(q_1^\beta + \cdots + q_m^\beta) - \log m} = \frac{-\sum_{j=1}^m q_j \log q_j}{\log m}.
\]

(iv) The result follows from the standard limit of $L^p$-norm, $\lim_{p \to \infty} \|x\|_p = \|x\|_{\infty}$, where $\|x\|_p = (\sum_{j=1}^m |x_j|^p)^{1/p}$ and $\|x\|_{\infty} = \max_{j \in [m]} |x_j|$ for $x \in \mathbb{R}^m$.

Finally, we show that the composite loss $L_\beta^r(j, q^h)$ is convex in $h$ for $\beta \in [0, 1]$. The case $\beta = 0$ or 1 can be verified directly, corresponding to the simultaneous exponential or likelihood composite loss. For $\beta \in (0, 1)$, the unscaled composite loss $L_\beta(j, q^h)$ is
\[
L_\beta(j, q^h) = \left\{ 1 + \sum_{i \neq j} \exp(\beta(h_i - h_j)) \right\}^{\frac{1}{\beta} - 1}.
\]
It suffices to show that for $\beta \in (0, 1)$, the function
\[
g(x) = \left\{ 1 + \sum_{i=1}^{m-1} \exp(\beta x_i) \right\}^{\frac{1}{\beta} - 1}
\]
is convex in $x \in \mathbb{R}^{m-1}$. Rewrite $g(x)$ as
\[
g(x) = \exp\left[\left(\frac{1}{\beta} - 1\right) \log \left\{ 1 + \sum_{i=1}^{m-1} \exp(\beta x_i) \right\}\right].
\]
Note that $\log\{1 + \sum_{i=1}^{m-1} \exp(\beta x_i)\}$ is convex in $x$ [27, Example 3.14]. The convexity of $g(x)$ follows by the scalar composition rule in [27, Section 3.2.4].

\[\]
Identity (S17) follows from a second-order Taylor expansion with an integral remainder for the univariate function $H_L(q + t(\eta - q))$ with $t \in [0,1]$. 

(i) By definition (5), the generalized entropy corresponding to the pairwise (symmetrized) loss $L$ in (39) is $H_L(q) = -\sum_{i=1}^m \sum_{j \neq i} q_i f_0(q_j / q_i)$. See also Supplement Table S1. The first-order and second-order derivatives of $H(q)$ are

$$
\frac{\partial H}{\partial q_i} = -\sum_{j \neq i} \left\{ f_0(q_j / q_i) - q_j f_0'(q_j / q_i) + f_0'(q_i / q_j) \right\},
$$

$$
\frac{\partial^2 H}{\partial q_i^2} = -\sum_{j \neq i} \left\{ \frac{q_i}{q_j} f_0''(q_j / q_i) + \frac{1}{q_i} f_0''(q_i / q_j) \right\},
$$

$$
\frac{\partial^2 H}{\partial q_i \partial q_j} = \frac{q_i}{q_j} f_0''(q_j / q_i) + \frac{q_j}{q_i} f_0''(q_i / q_j), \quad j \neq i.
$$

By the relationship $w(q_1) = f_0''(w^2)/q_2^2$, we obtain $f_0''(w^2) = 2\nu q_1^{-1} q_2^{-1} + 2\nu$ from $w(q_1) = 2\nu q_1^{-1} q_2^{-1}$. Then the quadratic form $-x^T \nabla^2 H_L(\eta)x$ with $x = \eta - q$ and $\eta = q + ts(\eta - q) \in \Delta_m$ can be written as

$$
-x^T \nabla^2 H_L(\eta)x = \sum_{i=1}^m \sum_{j=1}^m \left\{ \left( \frac{\eta_i^2}{q_i} f_0''(\frac{\eta_i}{q_i}) + \frac{1}{q_i} f_0''(\frac{\eta_i}{q_i}) \right) x_i^2 \right\} - \left\{ \frac{\eta_j}{q_i} f_0''(\frac{\eta_i}{q_i}) + \frac{\eta_i}{q_j} f_0''(\frac{\eta_i}{q_j}) \right\} x_i x_j.
$$

With $\nu \leq 0$, note that $(\eta_i \eta_j)^{\nu}/(\eta_i + \eta_j)^{2\nu+1} \geq 2^{-2\nu}$ because $\eta_i \eta_j \leq 2^{-2}(\eta_i + \eta_j)^2$ and $\eta_i + \eta_j \leq 1$ for each pair $(i,j)$. Hence we have

$$
-x^T \nabla^2 H_L(\eta)x \geq \sum_{i=1}^m \sum_{j=1}^m (\eta_i^2 - \eta_i^2 + 2 \eta_i \eta_j - \eta_i^2 \eta_j^2 x_j^2)
$$

(S18)

$$
= 2 \left\{ \left( \sum_{i=1}^m \eta_i^{-1} x_i^2 \right) \left( \sum_{j=1}^m \eta_j \right) - \left( \sum_{i=1}^m x_i^2 \right) \right\} \geq 2\|x\|^2,
$$

(S19)

where the last inequality follows because $(\sum_{i=1}^m \eta_i^{-1} x_i^2)/(\sum_{i=1}^m \eta_i) \geq (\sum_{i=1}^m |x_i|)^2$ by the Cauchy–Schwarz inequality and $\sum_i x_i = \sum_i p_i - \sum_i \eta_i = 0$. Combining this lower bound with (S17) and integrating over $s$ and $t$ yield $\kappa_L = 2$.

We show that the constant $\kappa_L = 2$ cannot be improved as stated. For $m = 2$, we take $q_1 = \eta_2 = \frac{1}{2} - \delta$ and $q_2 = \eta_1 = \frac{1}{2} + \delta$, so that $x_1 = -2\delta$ and $x_2 = 2\delta$. For $m \geq 3$, we take $q_1 = \frac{1}{2} - \delta_2$, $q_2 = \frac{1}{2} + \delta_2$, $q_3 = \cdots = q_{m-1} = \delta/(m-2)$, $\eta_1 = \frac{1}{2} + \delta$, $\eta_2 = \frac{1}{2} - \delta$, $\eta_3 = \cdots = \eta_m = \delta/(m-2)$, so that $x_1 = -3\delta$, $x_2 = 3\delta$, $x_3 = \cdots = x_m = 0$. Then it can be verified that as $\delta \to 0$, each of inequalities (S18) and (S19) used above divided by $\|x\|^2$ on both sides becomes equality. In fact, for inequality (S19), we have

$$
\frac{(\sum_{i=1}^m \eta_i^{-1} x_i^2)(\sum_{i=1}^m \eta_i)}{\|x\|^2} \to 1,
$$

because $\eta_1, \eta_2 \to 1/2$ and $\eta_3, \ldots, \eta_m \to 0$ as $\delta \to 0$. For inequality (S18), we distinguish the following cases of $(i,j)$. If $i \neq j \in \{1,2\}$, then $2^{2\nu}(\eta_i \eta_j)^{\nu}/(\eta_i + \eta_j)^{2\nu+1} \to 1$. If $i \in \{1,2\}$ and $j \notin \{1,2\}$, then $x_j = 0$ and for $m \geq 3$ and $\nu \in (-1, 0]$,

$$
2^{2\nu}(\eta_i \eta_j)^{\nu}/(\eta_i + \eta_j)^{2\nu+1} = O(\eta_i^{1+\nu}) \to 0,
$$

$$
2^{2\nu}(\eta_i \eta_j)^{\nu}/(\eta_i + \eta_j)^{2\nu+1} = O(\eta_j) \to 0.
$$

If $i \neq j \in \{3, \ldots, m\}$, then $x_i = x_j = 0$ and

$$
2^{2\nu}(\eta_i \eta_j)^{\nu}/(\eta_i + \eta_j)^{2\nu+1} = 0,
$$

$$
2^{2\nu}(\eta_i \eta_j)^{\nu}/(\eta_i + \eta_j)^{2\nu+1} = 0.
$$

Combining the three cases shows the desired reduction of inequality (S18) for $m = 2$ or for $m \geq 3$ and $\nu \in (-1, 0]$. 
(ii a) Suppose $\beta \in [1/2, 1]$. The generalized entropy corresponding to the simultaneous loss $L$ in (41) is $H(q) = \|q\|_\beta$. The first-order and second-order derivatives are

$$
\frac{\partial H}{\partial q_i} = q_i^{\beta - 1}\|q\|^{1-\beta}_\beta, \\
\frac{\partial^2 H}{\partial q_i^2} = - (1 - \beta) (\sum_{j \neq i} q_j q_i)^{\beta - 2}\|q\|^{1-2\beta}_\beta, \\
\frac{\partial^2 H}{\partial q_i \partial q_j} = (1 - \beta) (q_j q_i)^{\beta - 1}\|q\|^{1-2\beta}_\beta, \quad j \neq i.
$$

The quadratic form $-x^T \nabla^2 H_L(\hat{q}) x$ with $x = \eta - q$ and $\hat{q} = q + ts(\eta - q) \in \Delta_m$ can be written as

$$
-x^T \nabla^2 H_L(\hat{q}) x = \frac{1 - \beta}{2} \left\{ \sum_{i=1}^m \sum_{j=1}^m \|\hat{q}\|^{1-2\beta}_\beta (\hat{q}_i \hat{q}_j)^{\beta - 1} (\hat{q}_i^{\frac{1}{2}} \hat{q}_j^{\frac{1}{2}} x_i - \hat{q}_i^{\frac{1}{2}} \hat{q}_j^{\frac{1}{2}} x_j)^2 \right\}.
$$

With $\beta \in [1/2, 1)$ and hence $\beta - 1 < 0$, it holds that $(\hat{q}_i \hat{q}_j)^{\beta - 1} \geq 2^{2-2\beta}$ by inverting the inequality $\hat{q}_i \hat{q}_j \leq 2^{-2}(\hat{q}_i + \hat{q}_j)^2 \leq 2^{-2}$. In addition, $\|\hat{q}\|_\beta^2$ is concave and attains the maximum $m^{1-1/\beta}$ over $\Delta_m$ when $\hat{q}_i = 1/m$ for $i \in [m]$. Because $1 - 2\beta \leq 0$, it follows that the minimum of $\|\hat{q}\|^{1-2\beta}_\beta$ over $\Delta_m$ is $m^{(1-1/\beta)(2\beta - 1)}$. Then the quadratic form is lower bounded by

$$
-x^T \nabla^2 H_L(\hat{q}) x \geq \frac{1 - \beta}{2} m \frac{(\beta - 1)(2\beta - 1)}{2^{2-2\beta}} \left\{ \sum_{i=1}^m \sum_{j=1}^m (\hat{q}_i^{\frac{1}{2}} \hat{q}_j^{\frac{1}{2}} x_i - \hat{q}_i^{\frac{1}{2}} \hat{q}_j^{\frac{1}{2}} x_j)^2 \right\} = (1 - \beta)m \frac{(\beta - 1)(2\beta - 1)}{2^{2-2\beta}} \left\{ \left( \sum_{i=1}^m \hat{q}_i^{-1} x_i^2 \right) \left( \sum_{j=1}^m \hat{q}_j - \sum_{i=1}^m x_i \right)^2 \right\} \geq \frac{1 - \beta}{2} m \frac{(\beta - 1)(2\beta - 1)}{2^{2-2\beta}} \|x\|^2_1,
$$

where the last inequality follows similarly as in the proof of (i), by the Cauchy–Schwarz inequality and $\sum_i x_i = 0$. Integration of (S17) with the preceding lower bound yields $\kappa_L = (1 - \beta)m \frac{(\beta - 1)(2\beta - 1)}{2^{2-2\beta}}$.

(ii b) Suppose $\beta \in (0, 1/2]$. The generalized entropy, derivatives and quadratic form remain the same as in (ii a). With $\beta \in (0, 1/2]$ and hence $1 - 2\beta > 0$, we have

$$
\|\hat{q}\|^{1-2\beta}_\beta (\hat{q}_i \hat{q}_j)^{\beta - 1} \geq (\hat{q}_i^{\beta} + \hat{q}_j^{\beta})^{1-2\beta} (\hat{q}_i \hat{q}_j)^{\beta - 1} = \left\{ \frac{(\hat{q}_i \hat{q}_j)^{\beta}}{(\hat{q}_i^{\beta} + \hat{q}_j^{\beta})} \right\}^{1-\frac{1}{2}} (\hat{q}_i^{\beta} + \hat{q}_j^{\beta})^{-\frac{1}{2}} \geq 2^{\frac{1}{2} - 1}.
$$

The first inequality holds trivially. The second inequality holds because $1 - 1/\beta < 0$, $(\hat{q}_i \hat{q}_j)^{\beta} \leq 2^{-2}(\hat{q}_i^{\beta} + \hat{q}_j^{\beta})^2$, and $(\hat{q}_i^{\beta} + \hat{q}_j^{\beta})^{-1/\beta}$ is lower bounded by $2^{1-1/\beta}$. Similarly as in (ii a), the quadratic form is lower bounded by

$$
-x^T \nabla^2 H_L(\hat{q}) x \geq (1 - \beta) 2^{\frac{1}{2} - 1} \left\{ \sum_{i=1}^m \hat{q}_i^{-1} x_i^2 \left( \sum_{j=1}^m \hat{q}_j - \sum_{i=1}^m x_i \right)^2 \right\} \geq (1 - \beta) 2^{\frac{1}{2} - 1} \|x\|^2_1, \quad (S20)
$$

where the last inequality follows from the Cauchy–Schwarz inequality and $\sum_i x_i = 0$. Integration of (S17) with the preceding lower bound yields $\kappa_L = (1 - \beta)2^{1/2-1}$.

To show that the constant $\kappa_L$ cannot be improved as stated, we take $q$ and $\eta$ the same as in the proof of (i). Similarly, it can be verified that as $\delta \to 0$, each of inequalities (S20) and (S21) used above divided by $\|x\|^2_1$ on both sides becomes equality. In fact, inequality (S21) reduces to equality for the same reason as inequality (S19). For inequality (S20), we distinguish the following cases of $(i, j)$. If $i \neq j \in \{1, 2\}$, then $\|\hat{q}\|^{1-2\beta}_\beta (\hat{q}_i \hat{q}_j)^{\beta - 1} \to 2^{\frac{1}{2} - 1}$. If $i \in \{1, 2\}$ and $j \not\in \{1, 2\}$, then $x_j = 0$ and for $m \geq 3$ and $\beta \in (0, 1/2]$,

$$
\|\hat{q}\|^{1-2\beta}_\beta (\hat{q}_i \hat{q}_j)^{\beta - 1} \left( \frac{\hat{q}_j^{\frac{1}{2}} \hat{q}_i^{\frac{1}{2}} x_i - \hat{q}_i^{\frac{1}{2}} \hat{q}_j^{\frac{1}{2}} x_j}{\|x\|^2_1} \right)^2 = O(\hat{q}_j^{\beta}) \to 0,
$$

$$
2^{\frac{1}{2} - 1} \left( \frac{\hat{q}_j^{\frac{1}{2}} \hat{q}_i^{\frac{1}{2}} x_i - \hat{q}_i^{\frac{1}{2}} \hat{q}_j^{\frac{1}{2}} x_j}{\|x\|^2_1} \right)^2 = O(\hat{q}_j^{\beta}) \to 0.
$$
If \( i \neq j \in \{3, \ldots, m\} \), then \( x_i = x_j = 0 \) and
\[
\|\tilde{\eta}\|^{1-2\beta}(\tilde{\eta}_j \tilde{\eta}_j)^{\beta-1} \left( \frac{\tilde{\eta}_j^2}{\|\tilde{\eta}\|^2} x_i - \frac{\tilde{\eta}_i}{\|\tilde{\eta}\|^2} \tilde{\eta}_j^2 x_j \right)^2 = 0,
\]
\[
2^{\frac{1}{\beta}-1} \frac{\tilde{\eta}_j^2}{\|\tilde{\eta}\|^2} x_i - \frac{\tilde{\eta}_i}{\|\tilde{\eta}\|^2} \tilde{\eta}_j^2 x_j \right)^2 = 0.
\]
Combining the three cases shows the desired reduction of inequality (S20) for \( m = 2 \) or for \( m \geq 3 \) and \( \beta \in (0, 1/2] \).

Discussion on the multinomial likelihood loss. By Proposition 9(iii), the standard likelihood loss \( L(j, q) = -\log q_j \) multiplied by \((\log m)^{-1}\) is equivalent to (41) in the limit of \( \beta \to 1 \). By Proposition 10(ii), inequality (47) can be shown to hold for the rescaled entropy \( H^2_\beta \) in (42) with \( \kappa_L = \kappa_\beta / (m^{1/\beta - 1} - 1) \), where \( \kappa_\beta = (1 - \beta) m^{(1-1/\beta)(2\beta-1)2^{-2\beta}} \) if \( \beta \in [1/2, 1] \). Then (49) can be recovered from (47) as \( \beta \to 1 \), because
\[
\lim_{\beta \to 1} \frac{\kappa_\beta}{m^{1/\beta - 1} - 1} = \lim_{\beta \to 1} \frac{1 - \beta}{m^{1/\beta - 1} - 1} = (\log m)^{-1}.
\]

Discussion on simultaneous exponential loss. The simultaneous exponential loss \( L_0' \) as used in [6] can be obtained from (41) in the limit of \( \beta \to 0^+ \) after properly rescaled, by Proposition 9(i). However, for \( m \geq 3 \), the corresponding modulus \( \kappa_L \) from Proposition 10(ii) for \( L_\beta' \) as \( \beta \to 0^+ \) gives 0:
\[
\lim_{\beta \to 0^+} \frac{(1 - \beta) 2^{1/\beta - 1}}{m^{1/\beta - 1} - 1} = 0.
\]
The limit above gives 1 for \( m = 2 \), in agreement with the relationship \( L_0' = L_{1/2} - 1 \) with \( m = 2 \). Our further calculation (not shown) suggests that a uniform bound in the form of (47) might not be feasible on the associated Bregman divergence. Hence an alternative approach would be needed to analyze \( \psi \) and deduce a concrete meaningful implication from regret bound (45) for the simultaneous exponential loss \( L_0' \).

I. Proofs of results in Sections V-B2–V-B3

Proof of Lemma 4. By definition,
\[
\begin{align*}
R_L(\eta, \gamma) &= \sum_{j \in [m]} \eta_j c_{jM} L(j, \gamma) + \sum_{j \in [m]} \eta_j \sum_{k \in [m], k \neq j} (c_{jM} - c_{jk}) \{ L(k, \gamma) - 1 \} \\
&= \sum_{j \in [m]} \eta_j c_{jM} L(j, \gamma) + \sum_{j \in [m]} \eta_j \sum_{k \in [m], k \neq j} (c_{jM} - c_{jk}) L(k, \gamma) - D(\eta),
\end{align*}
\]
where \( D(\eta) = \sum_{j \in [m]} \sum_{k \in [m], k \neq j} \eta_j (c_{jM} - c_{jk}) \). By an exchange of indices \( j \) and \( k \), the second term above is 0. Substituting this into the preceding expression for \( R_L(\eta, \gamma) \) yields
\[
R_L(\eta, \gamma) = \sum_{j \in [m]} L(j, \gamma) \tilde{\eta}_j - D(\eta) = (1_\eta^T \tilde{\eta}) R_L(\tilde{\eta}, \gamma) - D(\eta).
\]
The generalized entropy from \( \tilde{\eta} \) is
\[
H_L(\eta) = \inf_\gamma R_L(\eta, \gamma) = (1_\eta^T \tilde{\eta}) \inf_\gamma R_L(\tilde{\eta}, \gamma) - D(\eta) = (1_\eta^T \tilde{\eta}) H_L(\tilde{\eta}) - D(\eta).
\]
The desired result on \( B_{\tilde{L}}(\eta, \gamma) \) then follows.

Proof of Lemmas 5 and 6. The bound in Lemma 5 is a special case of Lemma 6 with \( C = 1_m 1_m^T - I_m \). If \( \eta = (\eta_1, \eta_2, 0, \ldots, 0)^T \) and \( q = (1/2, 1/2, 0, \ldots, 0)^T \), then the bound becomes exact: \( B^{\text{cw}}(\eta, q) = |2\eta_1 - 1| = |\eta_1 - q_1| + |\eta_2 - q_2| \) with \( \eta_1 = \eta_2 = 1 \).

For Lemma 6, let \( l = \arg\max_{j \in [m]} (C_j^T \eta) \) and \( k = \arg\max_{j \in [m]} (C_j^T \eta) \), where \( C = (C_1, \ldots, C_m) \) is a column representation of \( \tilde{C} \). By definition, \( C_j = C_j - C_M \) and \( C_j^T \eta = C_j^T \eta - C_M^T \eta = C_j^T \eta - C_j^T \eta \) for \( j \in [m] \). Direct calculation yields
\[
\begin{align*}
B^{\text{cw}}(\eta, C_j^T \eta) &= R_L(\eta, C_j^T \eta) - H^{\text{cw}}(\eta) \\
&= C_{jM}^T \eta - C_j^T \eta = C_j^T \eta - C_j^T \eta.
\end{align*}
\]
Then $B^{cw}(\eta, C^T q) = \overline{C}^i_\eta \eta - \overline{C}^i_k q \leq \overline{C}^i_\eta \eta - \overline{C}^i_l q + \overline{C}^i_k q - \overline{C}^i_k \eta$ because $\overline{C}^i_\eta q = \overline{C}^i_k q$ by definition. Hence $B^{cw}(\eta, C^T q) \leq |\overline{C}^i_\eta \eta - \overline{C}^i_l q| + |\overline{C}^i_k q - \overline{C}^i_k \eta| \leq \|C^T(\eta - q)\|_{\infty}$. ■

**Proof of Proposition 11.** Note that $L^{cw}(j, \gamma) = \bar{L}^{cw}(j, \gamma)$ by direct calculation. Applying Lemma 4 to $L$ and $\bar{L}^{cw}$ shows that for any $\eta, q \in \Delta_m$,

\[
B_L(\eta, q) = (1^T_m \eta)B_L(\tilde{\eta}, q),
\]

\[
B^{cw}(\eta, q) = (1^T_m \tilde{\eta})B^{cw}(\tilde{\eta}, q),
\]

where $\tilde{\eta}$ and $\hat{\eta}$ are defined as in Lemma 4. The desired result then follows because $B^{cw}(\tilde{\eta}, q) \leq \|\tilde{\eta} - q\|_{\infty}$ by Lemma 5, $\psi_q(||\tilde{\eta} - q||_{\infty}) \leq B_L(\tilde{\eta}, q)$ by definition, and $\psi_q(\cdot)$ is nondecreasing. ■

**Proof of Corollary 2.** Applying (60) with $C_0 = 1_m$ and $\eta$ replaced by $\tilde{\eta}$ yields

\[
\underline{\psi}(B^{cw}(\tilde{\eta}, q)) \leq B_L(\tilde{\eta}, q).
\]

Combining this with (S22) and (S23) gives the desired result. ■

**Proof of Proposition 12.** The desired result is obtained by combining the following observations: $B^{cw}(\eta, C^T q) \leq \|C^T(\eta - q)\|_{\infty}$ by Lemma 6, $\psi_q^C(\|C^T(\eta - q)\|_{\infty}) \leq B_L(\eta, q)$ by definition, and $\psi_q^C(\cdot)$ is nondecreasing. ■

**Proof of inequality (58).** For any $w \in \mathcal{W}_{\eta, q}$, $\arg\max_{j \in [m]}(C^T_j q^w)$ can be set to be same as $\arg\max_{j \in [m]}(C^T_j q)$ and hence $B^{cw}(\eta, C^T q^w) = B^{cw}(\eta, C^T q)$. Then inequality (57) with $q$ replaced by $q^w$ shows that $\psi_q^C(B^{cw}(\eta, C^T q)) \leq B_L(\eta, q)$. The desired result then follows because $B_L(\eta, q^w) \leq B_L(\eta, q)$ by the representation of $B_L(\eta, q)$ as the Bregman divergence (8) and inequality (S2) in Lemma S1. ■

**Proof of Corollary 3.** By the representation of $B_L(\eta, q)$ as the Bregman divergence (8) and inequality (S1) in Lemma S1, $\psi_q^C(t)$ in Proposition 12 can be equivalently defined with $\eta' \in \Delta_m$ restricted such that $\|C^T(\eta' - q)\|_{\infty} = t$.

We distinguish three cases. Let $k = \arg\max_{j \in [m]}(C^T_j q)$. First, if $\overline{C}_k^T q > 1_m^T \overline{C}^T q/2$ and $\overline{C}_\eta^T q > 1_m^T \overline{C}^T q/2$, then $k = \arg\max_{j \in [m]}(C^T_j \eta)$ and hence $B^{cw}(\eta, C^T q) = 0$ and (59) holds trivially. Second, if $\overline{C}_k^T q > 1_m^T \overline{C}^T q/2$ and $\overline{C}_\eta^T q \leq 1_m^T \overline{C}^T q/2$, then (58) holds with some $w \in \mathcal{W}_{\eta, q}$ such that $\overline{C}_k^T q^w = 1_m^T \overline{C}^T q^w/2$ and hence $\max_{j \in [m]}(C^T_j q^w) = 1_m^T \overline{C}^T q^w/2$, because $\overline{C}_k^T q^w/(1_m^T \overline{C}^T q^w)$ is continuous in $w \in [0, 1]$, while taking a value $\leq 1/2$ at $w = 0$ and $> 1/2$ at $w = 1$ by assumption. Third, if $\overline{C}_k^T q \leq 1_m^T \overline{C}^T q/2$, then (58) holds with $w = 1 \in \mathcal{W}_{\eta, q}$. In the latter two cases, inequality (59) can be shown as follows:

\[
\psi_C^C(B^{cw}(\eta, C^T q)) \leq \psi_q^C\left(B^{cw}(\eta, C^T q)\right) \leq B_L(\eta, q).
\]

The first inequality holds because $\psi_C^C(t) \leq \psi_q^C(t)$ with $q^w$ satisfying $\max_{j \in [m]}(C^T_j q^w) \leq 1_m^T \overline{C}^T q^w/2$. The second inequality holds by (58) with $w \in \mathcal{W}_{\eta, q}$. ■

**Proof of inequality (61).** For $\eta, q \in \Delta_2$, we have $\|C_0 \circ (\eta - q)\|_{\infty} = (c_{10} + c_{20})|\eta_1 - q_1|$, where $\eta = (\eta_1, \eta_2)^T$ and $q = (q_1, q_2)^T$. Moreover, $\max_{j=1,2}(c_j q_j)$ $\leq C_0^\circ q/2$ for $g \in \Delta_2$ leads to a single probability vector $q = (c_{20}, c_{10})^T/(c_{10} + c_{20})$. From these expressions, $\psi_C(t)$ can be simplified as $\psi_C^C(t) = \min\{\psi_{\text{RW}}(-t), \psi_{\text{RW}}(t)\}$. ■
### TABLE S1(a)
Examples of losses, dissimilarity functions, and generalized entropies

| Name | Loss $L(j, q)$ | Dissimilarity function $f(t)$ | Generalized Entropy $H(\eta)$ |
|------|-----------------|-------------------------------|-------------------------------|
| **TWO-CLASS LOSS** | | | |
| Likelihood | $-\mathbb{I}_{(j=1)} \log q_1 - \mathbb{I}_{(j=2)} \log q_2$ | $t \log t - (t + 1) \log (t + 1)$ | $-\eta_1 \log \eta_1 - \eta_2 \log \eta_2$ |
| Exponential | $\mathbb{I}_{(j=1)} \sqrt{\frac{q_1}{q_2}} + \mathbb{I}_{(j=2)} \sqrt{\frac{q_2}{q_1}}$ | $(\sqrt{t} - 1)^2$ | $-(\sqrt{\eta_1} - \sqrt{\eta_2})^2$ |
| Calibration, | $\mathbb{I}_{(j=1)} \frac{q_1}{q_2} + \mathbb{I}_{(j=2)} \frac{1}{2} (\log \frac{q_1}{q_2} - 1)$ | $-\frac{1}{2} \log t$ | $\frac{q_1}{2} \log \frac{q_1}{q_2}$ |
| Calibration, | $\mathbb{I}_{(j=1)} \frac{1}{2} (\log \frac{q_1}{q_2} + \frac{q_2}{q_1} - 1) + \mathbb{I}_{(j=2)} \frac{1}{2} (\log \frac{q_1}{q_2} + \frac{q_2}{q_1} - 1)$ | $\frac{1}{2} (t \log t - \log t)$ | $\frac{1}{2} (\eta_1 \log \frac{q_1}{q_2} + \eta_2 \log \frac{q_2}{q_1})$ |
| **MULTI-CLASS PAIRWISE ASYMMETRIC** | | | |
| Likelihood | $\mathbb{I}_{(j \in [m-1]}) \log(1 + \frac{q_j}{q_m}) + \mathbb{I}_{(j=m)} \sum_{i=1}^{m-1} \log(1 + \frac{q_i}{q_m})$ | $\sum_{i=1}^{m-1} \{t_i \log t_i - (1 + t_i) \log (1 + t_i)\}$ | $-\sum_{i=1}^{m-1} (\eta_m \log \frac{q_m}{q_m + q_i} + \eta_i \log \frac{q_i}{q_i + q_m})$ |
| Exponential | $\mathbb{I}_{(j \in [m-1]}) (\sqrt{\frac{q_j}{q_m}} - 1) + \mathbb{I}_{(j=m)} \sum_{i=1}^{m-1} (\sqrt{\frac{q_i}{q_m}} - 1)$ | $\sum_{i=1}^{m-1} (\sqrt{t_i} - 1)^2$ | $-\sum_{i=1}^{m-1} (\sqrt{\eta_i} - \sqrt{\eta_m})^2$ |
| Calibration | $\mathbb{I}_{(j \in [m-1])} \frac{q_j}{q_m} + \mathbb{I}_{(j=m)} \sum_{i=1}^{m-1} \frac{1}{2} (\log \frac{q_i}{q_m} - 1)$ | $-\sum_{i=1}^{m-1} \frac{1}{2} \log t_i$ | $\sum_{i=1}^{m-1} \frac{q_i}{2} \log \frac{q_i}{q_m}$ |
| **MULTI-CLASS PAIRWISE SYMMETRIC** | | | |
| Likelihood | $\sum_{i \neq j} 2 \log(1 + \frac{q_j}{q_i})$ | $-\sum_{i=1}^{m} \sum_{j \neq i} 2 t_i \log(1 + \frac{q_i}{t_i})$ | $\sum_{i=1}^{m} \sum_{j \neq i} \eta_i \log(1 + \frac{q_i}{q_j})$ |
| Exponential | $\sum_{i \neq j} 2 (\sqrt{\frac{q_i}{q_j}} - 1)$ | $\sum_{i=1}^{m} \sum_{j \neq i} (\sqrt{t_i} - \sqrt{t_j})^2$ | $-\sum_{i=1}^{m} \sum_{j \neq i} (\sqrt{\eta_i} - \sqrt{\eta_j})^2$ |
| Calibration | $\sum_{i \neq j} \frac{1}{2} (\log \frac{q_i}{q_j} + \frac{q_j}{q_i} - 1)$ | $-\sum_{i=1}^{m} \sum_{j \neq i} \frac{1}{2} \log \frac{q_i}{t_i}$ | $\sum_{i=1}^{m} \sum_{j \neq i} \frac{q_i}{2} \log \frac{q_i}{q_j}$ |
| **MULTI-CLASS SIMULTANEOUS** | | | |
| $L_\beta$ Family | $(m^{\frac{1}{\beta}} - 1)^{-1} \left[ \left( 1 + \sum_{i \neq j} \left( \frac{q_j}{q_i} \right)^{\frac{1}{\beta}} \right)^{\frac{1}{\beta}} - 1 \right]$ | $-m^{\frac{1}{\beta}} \left[ \left( 1 + \sum_{i=1}^{m} \eta_i \right)^{\frac{1}{\beta}} - 1 \right]$ | $(m^{\frac{1}{\beta}} - 1)^{-1} \left\{ \sum_{i=1}^{m} \eta_i \right\}^{\frac{1}{\beta}} - 1$ |
| Pairwise Exp ($\beta = 0$) | $(\prod_{i \neq j} \frac{q_j}{q_i})^{\frac{1}{\beta}}$ | $-m (\prod_{i=1}^{m} t_i)^{\frac{1}{\beta}}$ | $m (\prod_{i=1}^{m} \eta_i)^{\frac{1}{\beta}}$ |
| Simultaneous Exp ($\beta = \frac{1}{2}$) | $(m - 1)^{-1} \sum_{i \neq j} \sqrt{\frac{q_i}{q_j}}$ | $-(m - 1)^{-1} \left( 1 + \sum_{i=1}^{m-1} \sqrt{t_i} \right)^2 - t_i$ | $(m - 1)^{-1} \left( \sum_{i=1}^{m-1} \sqrt{\eta_i} \right)^2 - 1$ |
| Multinomial Lik ($\beta = 1$) | $-m (\log m)^{-1} \log(q_i)$ | $(\log m)^{-1} \sum_{i=1}^{m} t_i \log \frac{t_i}{t}$ | $-(\log m)^{-1} \sum_{i=1}^{m} \eta_i \log \eta_i$ |

Note: $t_m = 1$ and $t_* = \sum_{i=1}^{m} t_i$. Calibration, and Calibration, are the asymmetric and symmetric versions.
### Table S1(b)

**Examples of losses and gradients**

| Name                           | Gradients $\frac{\partial}{\partial q} L(j, q^h)$ |
|--------------------------------|-----------------------------------------------|
| **TWO CLASS LOSS**             |                                               |
| Likelihood                     |                                               |
| Exponential                    | $\frac{1}{2} (1_{(j=1)} \frac{q_2}{q_1} + 1_{(j=2)} \sqrt{\frac{q_2}{q_1}} \frac{1}{1\cdot (\cdot 1)}$ |
| Calibration$_1$                | $\frac{1}{2} (1_{(j=1)} \frac{q_2}{q_1} + 1_{(j=2)} \cdot 1) \cdot (1\cdot (\cdot 1)$ |
| Calibration$_2$                | $\frac{1}{2} (1_{(j=1)} (1 + \frac{q_2}{q_1}) + 1_{(j=2)} (1 + \frac{q_2}{q_1}) \cdot (1\cdot (\cdot 1)$ |
| **MULTI-CLASS PAIRWISE ASYMMETRIC** |                                               |
| Likelihood                     | $\begin{cases} -\frac{1}{l} (j=1) \frac{q_m}{q_l} + 1_{(j=m)} \frac{q_l}{q_m} & l \in [m-1] \\ 1_{(j\neq l)} \frac{q_l}{q_m} + 1_{(j=m)} \sum_{i=1}^{m-1} \frac{q_l}{q_i} & l = m \end{cases}$ |
| Exponential                    | $\begin{cases} \frac{1}{l} (1_{(j=1)} \frac{q_m}{q_l} + 1_{(j=m)} \sqrt{\frac{q_m}{q_l}} & l \in [m-1] \\ \frac{1}{l} (1_{(j\neq l)} \sqrt{\frac{q_m}{q_l}} - 1_{(j=m)} \sum_{i=1}^{m-1} \sqrt{\frac{q_m}{q_i}} & l = m \end{cases}$ |
| Calibration                    | $\begin{cases} \frac{1}{l} (1_{(j\neq l)} \frac{q_m}{q_l} + 1_{(j=m)} \cdot 1) & l \in [m-1] \\ \frac{1}{l} (1_{(j\neq l)} \frac{q_m}{q_l} - 1_{(j=m)} \cdot (m-1)) & l = m \end{cases}$ |
| **MULTI-CLASS PAIRWISE SYMMETRIC** |                                               |
| Likelihood                     | $2(1_{(j\neq l)} \frac{q_l q_m}{q_l + q_m} - 1_{(j=l)} \sum_{i\neq l} 1_{(q_i = q_l)})$ |
| Exponential                    | $2(1_{(j\neq l)} \sqrt{\frac{q_l}{q_m}} - 1_{(j=l)} \sum_{i\neq l} \sqrt{\frac{q_l}{q_i}}$ |
| Calibration                    | $\frac{1}{2} (1_{(j\neq l)} \frac{q_m q_l}{q_l + q_m} + 1_{(j=l)} \sum_{i\neq l} (q_m + 1))$ |
| **MULTI-CLASS SIMULTANEOUS**   |                                               |
| $L_\beta$ Family               | $\frac{1-\beta}{m} \left( \sum_{l=1}^{m} q_l^{\beta} \right)^{\frac{1}{\beta}} \cdot 2 (1_{(j\neq l)} q_j^{\beta-1} q_l q_j - 1_{(j=l)} \sum_{i\neq l} q_i^{\beta-1} q_l q_i)$ |
| Pairwise $\text{Exp}(\beta = 0)$| $\frac{1}{l} 1_{(j\neq l)} \prod_{j\neq l} 1_{(q_j = q_l)} + m \sum_{l=1}^{m} 1_{(j=l)} \prod_{j\neq l} 1_{(q_j = q_l)}$ |
| Simultaneous $\text{Exp}(\beta = \frac{1}{2})$ | $\frac{1}{2(m-1)} (1_{(j\neq l)} \sqrt{\frac{q_l}{q_j}} - 1_{(j=l)} \sum_{i\neq l} \sqrt{\frac{q_l}{q_i}}$ |
| Multinomial $\text{Lik}(\beta = 1)$ | $\frac{1}{\log m} (1_{(j\neq l)} q_j + 1_{(j=l)} (q_l - 1))$ |