9. Additional Proofs

In addition to providing proofs not in the main text in chronological order, we restate what is being proved for convenience.

Lemma For Theorem 2.2. Let $\psi : \mathbb{R}^M \times \mathcal{Y} \to [0, \infty)$ be a nonnegative loss function. $L^*_\psi : \Delta_M \to \mathbb{R}$ defined by $L^*_\psi(\eta) = \inf_{s\in\mathbb{R}^M} \sum_{i=1}^M \eta_i \psi(s, i)$ is continuous.

Proof. First, note that L^*_ψ is concave, because it is a pointwise infimum of affine functions of η. Also, it is finite valued, because ψ is lower bounded (thus $L^*_\psi(\eta) > -\infty$) and clearly $L^*_\psi(\eta) < \infty$.

By Theorem 10.2 of Rockafellar (1970), any concave function taking finite real values on a locally simplicial subset $S \subseteq \mathbb{R}^M$ is lower semicontinuous. That is, for all $x \in S$ and sequences $\{x^{(n)}\}$ converging to x, $f(x) \le \lim_{n\to\infty} f(x^{(n)})$ if the limit on the right exists.

Δ_M is locally simplicial (it is the probability simplex) and L^*_ψ satisfies the assumptions, so L^*_ψ is lower semicontinuous.

Now we just need to show upper semicontinuity, which can be stated as: for any $\epsilon > 0, \eta \in \Delta_M$, there exists $\delta > 0$ where for all $\eta' \in \Delta_M$, $\|\eta' - \eta\|_2 \le \delta$ implies $L^*_\psi(\eta') \le L^*_\psi(\eta) + \epsilon$.

Let $\eta \in \Delta_M, \epsilon > 0$. Choose s so that $L_\psi(s, \eta) \le L^*_\psi(\eta) + \epsilon/2$, which is possible by definition of L^*. Now set $\delta = \epsilon \left(2 \max \left\{ \sqrt{\sum_{i=1}^M \psi(s, i)^2}, 1 \right\} \right)^{-1}$ (taking the max with 1 is to avoid a zero in the denominator), and suppose $\eta' \in \Delta$, $\|\eta - \eta'\|_2 \le \delta$. We have,

$$\begin{aligned} L^*_\psi(\eta') &\le L_\psi(s, \eta') = \sum_{i=1}^M \eta'_i \psi(s, i) \\ &= \sum_{i=1}^M \eta_i \psi(s, i) + \sum_{i=1}^M (\eta'_i - \eta_i)\psi(s, i) \\ &\le L^*_\psi(\eta) + \epsilon/2 + \|\eta' - \eta\|_2 \sqrt{\sum_{i=1}^M \psi(s, i)^2} \\ &\le L^*_\psi(\eta) + \epsilon/2 + \epsilon/2 = L^*_\psi(\eta) + \epsilon. \end{aligned}$$

The first inequality is by definition of L^*, and the second inequality uses the Cauchy-Schwartz inequality. Therefore, L^* is upper semicontinuous. Since it is also lower semicontinuous, it is continuous. □

Theorem 2.2. Suppose ψ is a nonnegative top-k calibrated loss function. Then ψ is top-k consistent in the sense that for any sequence of measurable functions $f^{(n)} : \mathcal{X} \to \mathbb{R}^M$, we have

$$L_\psi(f^{(n)}) \to L^*_\psi \implies L_{\mathrm{err}_k}(f^{(n)}) \to L^*_{\mathrm{err}_k}.$$

Proof. We place top-k classification in the abstract decision model in Appendix A. of Zhang (2004a) with output-model space $\mathcal{Q} = \Delta_M$, decision space $\mathcal{D}$ equal to the set of subsets of $[M]$ of size k, and estimation-model space $\Omega = \mathbb{R}^M$. The risk function is the top-k error and the decision rule is equal to r_k, the top-k thresholding operator.

By Corollary 26 of Zhang (2004a) we just need to show that for any $\epsilon > 0$,

$$\Delta H(\epsilon) = \inf \left\{ \Delta L_\psi(s, \eta) \mid \Delta L^*_{\mathrm{err}_k}(s, \eta) \ge \epsilon \right\} > 0,$$

where $\Delta L(s, \eta) := L(s, \eta) - L^*(\eta)$. In other words, we need to show that given any $\epsilon > 0$, there is a $\delta > 0$ such that $\Delta L_{\mathrm{err}_k}(s, \eta) \ge \epsilon$ implies $\Delta L_\psi(s, \eta) \ge \delta$.

Proof by contradiction. Given $\epsilon > 0$, assume there does not exist $\delta > 0$ such that the above holds. Then, there is a sequence $\{s^{(n)}, \eta^{(n)}\}$ such that $\Delta L_{\mathrm{err}_k}(s^{(n)}, \eta^{(n)}) \ge \epsilon$ for all $n \in \mathbb{N}$ and yet $\Delta L_\psi(s^{(n)}, \eta^{(n)}) \to 0$. Since $\eta^{(n)}$ comes from a compact set Δ_M, we may assume that $\eta^{(n)} \to \eta$ without loss of generality, since otherwise we could take a convergent subsequence.

We will show that $\Delta L_\psi(s^{(n)}, \eta) \to 0$, which provides a contradiction in the following. Because ψ is top-k calibrated, $s^{(n)}$ is top-k preserving with respect to η for all n greater than some N. This means there exists N where $\Delta L_{\mathrm{err}_k}(s^{(n)}, \eta) = 0$ for all $n > N$, i.e. $L_{\mathrm{err}_k}(s^{(n)}, \eta) = L^*_{\mathrm{err}_k}(\eta)$. By continuity of $L^*_{\mathrm{err}_k}$, there exists N' such that $|L^*_{\mathrm{err}_k}(\eta^{(n)}) - L^*_{\mathrm{err}_k}(\eta)| < \frac{\epsilon}{2}$ for all $n > N'$. But this means $\Delta L^*_{\mathrm{err}_k}(s^{(n)}, \eta^{(n)}) < \frac{\epsilon}{2}$ for $n > \max\{N, N'\}$, a contradiction.

Since $\Delta L_\psi(s^{(n)}, \eta^{(n)}) \to 0$, for any $\epsilon' > 0$, there exists $N > 0$ such that for all $n > N$, we have

$$|L_\psi(s^{(n)}, \eta^{(n)}) - L^*_\psi(\eta^{(n)})| \le \epsilon'/2.$$

Moreover, since L^*_ψ is continuous by Lemma **??** and $\eta^{(n)} \to \eta$, there exists $N' > 0$ such that for all $n > N'$, we have

$$|L^*_\psi(\eta^{(n)}) - L^*_\psi(\eta)| \le \epsilon'/2.$$

Then, for all $n > \max\{N, N'\}$,

$$\begin{aligned} |L_\psi(s^{(n)}, \eta^{(n)}) - L^*_\psi(\eta)| &\le |L_\psi(s^{(n)}, \eta^{(n)}) - L^*_\psi(\eta^{(n)})| \\ &\quad + |L^*_\psi(\eta^{(n)}) - L^*_\psi(\eta)| \le \epsilon'. \end{aligned}$$

Since ϵ' was arbitrary, we have $L_\psi(s^{(n)}, \eta^{(n)}) \to L^*_\psi(\eta)$.

Now we extend to $L_\psi(s^{(n)}, \eta) \to L^*_\psi(\eta)$ by showing that $L_\psi(s^{(n)}, \eta^{(n)})$ is close to $L_\psi(s^{(n)}, \eta)$. Given any $\epsilon' > 0$, let N be such that for all $n > N$, $L_\psi(s^{(n)}, \eta^{(n)}) - L^*_\psi(\eta) \le \epsilon'$. Then we have for all $n > N$

$$L_\psi(s^{(n)}, \eta^{(n)}) - L_\psi(s^{(n)}, \eta) \le L_\psi(s^{(n)}, \eta^{(n)}) - L^*_\psi(\eta) \le \epsilon'.$$

Let I be the support of η. For every $i \in I$, $\{\psi(s^{(n)}, i)\}$ is bounded, since $\psi \ge 0$ and if it were unbounded above then

$L_\psi(s^{(n)}, \eta^{(n)}) \geq \frac{\eta_i}{2}\psi(s^{(n)}, i) \to \infty > L^*(\eta)$ eventually. Now suppose $C > 0$ upper bounds $\{\psi_i(s^{(n)})\}$ for every $i \in I$. Since $\eta^{(n)} \to \eta$, There exists N' such that $n > N'$ implies $\eta_i^{(n)} \geq \eta_i - \epsilon'/(MC)$ for every $i \in [M]$. Then,

$$\begin{aligned} L_\psi(s^{(n)}, \eta^{(n)}) - L_\psi(s^{(n)}, \eta) &= \sum_{i=1}^{M}(\eta_i^{(n)} - \eta_i)\psi(s^{(n)}, i) \\ &\geq \sum_{i\in I}(\eta_i^{(n)} - \eta_i)\psi(s^{(n)}, i) \\ &\geq M\left(\frac{-\epsilon'}{MC}C\right) = -\epsilon'. \end{aligned}$$

Therefore, for all $n > \max\{N, N'\}$, we have

$$|L_\psi(s^{(n)}, \eta^{(n)}) - L_\psi(s^{(n)}, \eta)| \leq \epsilon'.$$

Since $\epsilon' > 0$ was arbitrary, this implies that $\{L_\psi(s^{(n)}, \eta)\}$ converges to the same limit as $\{L_\psi(s^{(n)}, \eta^{(n)})\}$. Thus, $L_\psi(s^{(n)}, \eta) \to L_\psi^*(\eta)$. We have thus reached the contradiction laid out earlier. □

Proof of Theorem 3.1. To prove Theorem 3.1, we use the following two lemmas. The first establishes the openness of the set $\{s \in \mathbb{R}^M \mid \mathsf{P}_\mathsf{k}(s, \eta)\}$ for any $\eta \in \mathbb{R}^M$. The second says that a convex function with a unique minimizer has bounded sublevel sets.

Lemma 9.1. $\mathsf{P}_k(\eta) := \{s \in \mathbb{R}^M \mid \mathsf{P}_\mathsf{k}(s, \eta)\}$ *is open for any* $\eta \in \mathbb{R}^M$, $k \in \mathbb{Z}^+$.

Proof. Let $\eta \in \mathbb{R}^M$ and $s \in \mathsf{P}_k(\eta)$. Define

$$\begin{aligned} \delta_1 &= \min_{i\in[M]}\{s_i - s_{[k+1]} \mid s_i > s_{[k+1]}\} \\ \delta_2 &= \min_{i\in[M]}\{s_{[k]} - s_i \mid s_i < s_{[k]}\} \end{aligned}$$

Take $\delta = \min\{\delta_1, \delta_2\}$, and notice $\delta > 0$. Then, take $s' \in \mathbb{R}^M$ with $|s'_i - s_i| < \delta/2$ for all $i \in [M]$. If $s_i > s_{[k+1]}$, then

$$s'_i > s_i - \delta/2 > s_{[k+1]} + \delta/2 > s'_{[k+1]},$$

and similarly if $s_i < s_{[k]}$ then $s'_i < s'_{[k]}$. Therefore, $\mathsf{P}_\mathsf{k}(s', \eta)$. This holds for every s' in the neighborhood – thus $\mathsf{P}_k(\eta)$ is open. □

Lemma 9.2. *If* $f : \mathbb{R}^M \to \mathbb{R}$ *is convex and has a unique minimizer, the sublevel sets* $\{x \in \mathbb{R}^M \mid f(x) \leq \alpha\}$ *are bounded for every* $\alpha \in \mathbb{R}$.

Proof. Suppose $x_0 \in \mathbb{R}^M$ is the unique minimizer. We can assume $x_0 = 0$ by taking $f(x + x_0)$, which has the same sublevel sets just shifted by x_0, and a unique minimizer at $x = 0$.

Then, $f(x) > f(0)$ for all $x \in \mathbb{R}^M$. Consider the set $B = \{x \in \mathbb{R}^M \mid \|x\|_2 = 1\}$. B is compact. Therefore, the image of B under f, $f(B) \subset \mathbb{R}$, is compact and has a minimum. Since $f(x) > f(0)$ for all $x \in B$, we have

$$\delta := \min(f(B)) - f(0) > 0.$$

Now, suppose $x \in \mathbb{R}^M$ such that $\|x\|_2 = D \geq 1$. Since $D \geq 1$, we have $0 < 1/D \leq 1$. Note $\|x/D\|_2 = 1$. Now we apply convexity:

$$f\left(\frac{x}{D}\right) \leq \frac{1}{D}f(x) + \left(1 - \frac{1}{D}\right)f(0).$$

Rearranging,

$$\begin{aligned} f(x) &\geq Df\left(\frac{x}{D}\right) + (1 - D)f(0) \\ &= D(f(x/D) - f(0)) + f(0) \\ &\geq D\delta + f(0). \end{aligned}$$

Thus, if $D \geq 1$, we have $\|x\|_2 \geq D$ implies $f(x) > D\delta/2 + f(0)$. The contrapositive is, $f(x) \leq D\delta/2 + f(0)$ implies $\|x\|_2 < D$ for $D \geq 1$. Therefore, for all $x \in \mathbb{R}^M$

$$f(x) \leq \alpha \implies \|x\|_2 \leq \max\left\{\frac{2(\alpha - f(0))}{\delta}, 1\right\}.$$

This says that the sublevel sets are bounded. □

Now we prove the theorem.

Theorem 3.1. Suppose $\phi : \mathbb{R}^M \to \mathbb{R}^M$ is strictly convex and differentiable. If $g : \mathbb{R}^M \to \mathbb{R}^M$ is inverse top-k preserving, continuous, and $\Delta_M \subseteq \mathrm{range}(g)$, then $\psi : \mathbb{R}^M \times \mathcal{Y} \to \mathbb{R}$ defined by

$$\psi(s, y) = D_\phi(g(s), e_y)$$

is top-k calibrated.

Proof. Let $\eta \in \Delta_M$. By Theorem 1 from Banerjee et al. (2005),

$$\arg\min_{\bar{\eta}\in\mathbb{R}^M} \mathbb{E}_{Y\sim\eta}D_\phi(\bar{\eta}, Y) = \mathbb{E}[Y] = \eta.$$

We view the label Y as an indicator vector in $\{0, 1\}^M$ where the position of the one corresponds to the label. Therefore,

$$\begin{aligned} \arg\min_{s\in\mathbb{R}^M} L_\psi(s, \eta) &= \arg\min_{s\in\mathbb{R}^M} \mathbb{E}_{Y\sim\eta}D_\phi(g(s), Y) \\ &= \{s \in \mathbb{R}^M \mid g(s) = \eta\}, \end{aligned}$$

and since $\Delta_M \subseteq \mathrm{range}(g)$ the last set is nonempty. Let s^* be such that $g(s^*) = \eta$.

Since g is inverse top-k preserving, $\mathsf{P}_\mathsf{k}(s^*, \eta)$. This holds for any s^* in $O := \{s \in \mathbb{R}^M \mid g(s) = \eta\}$. Given any s for which $\neg\mathsf{P}_\mathsf{k}(s,\eta)$, $s \notin O$, and thus $g(s) \neq \eta$, $L_\psi(s,\eta) = \mathbb{E}_{Y\sim\eta} D_\phi(g(s), Y) > \mathbb{E}_{Y\sim\eta} D_\phi(\eta, Y)$. Therefore,

$$\inf_{s\in\mathbb{R}^M:\neg\mathsf{P}_\mathsf{k}(s,\eta)} L_\psi(s,\eta) > \min_{s'\in\mathbb{R}^M} L_\psi(s',\eta).$$

To see this, first note $\mathbb{E}_{y\sim\eta} D_\phi(g, e_y)$ is convex in g while attaining a unique minimum by Banerjee et al. (2005). Therefore, by Lemma 9.2 the sublevel sets $\{g \mid \mathbb{E}_{y\sim\eta} D_\phi(g, e_y) \leq \alpha\}$ are bounded for any $\alpha \in \mathbb{R}$. Then

$$\begin{aligned}\inf_{g\in\mathbb{R}^M:\neg\mathsf{P}_\mathsf{k}(g,\eta)} \mathbb{E}_{y\sim\eta} D_\phi(g, e_y) &= \min_{g\in\mathbb{R}^M:\neg\mathsf{P}_\mathsf{k}(g,\eta)} \mathbb{E}_{y\sim\eta} D_\phi(g, e_y) \\ &> \min_{s\in\mathbb{R}^M} L_\psi(s,\eta),\end{aligned}$$

as $\{g \in \mathbb{R}^M : \neg\mathsf{P}_\mathsf{k}(g,\eta)\}$ is closed by 9.1, and for the infimum we only have to consider its intersection with some bounded closed (i.e. compact) set, due to the boundedness of the sublevel sets. Then since continuous functions map compact sets to compact sets, we can switch the infimum to a minimum.

Because g is inverse top-k preserving, $\mathsf{P}_\mathsf{k}(s, g(s))$. Then, if $\mathsf{P}_\mathsf{k}(g(s),\eta)$, we see by transitivity of P_k that $\mathsf{P}_\mathsf{k}(s,\eta)$. Therefore, $\neg\mathsf{P}_\mathsf{k}(s,\eta) \implies \neg\mathsf{P}_\mathsf{k}(g(s),\eta)$. So, $A := \{L_\psi(s,\eta) \mid \neg\mathsf{P}_\mathsf{k}(s,\eta)\} \subseteq \{\mathbb{E}_{y\sim\eta} D_\phi(g,\eta) \mid \neg\mathsf{P}_\mathsf{k}(g,\eta)\} =: B$, and

$$\inf A \geq \min B > \min_{s\in\mathbb{R}^M} L_\psi(s,\eta).$$

Thus, ψ is top-k calibrated. □

Theorem 4.1. Say a permutation $\pi : [M] \to [M]$ *sorts* a vector $v \in \mathbb{R}^M$ if $v_{\pi_1} \geq v_{\pi_2} \geq \ldots \geq v_{\pi_M}$. Denote $S(v)$ as the set of permutations that sort v.

Let $\eta \in \Delta_M$, and suppose it has no zero entries. Then, for each of the following cases, the set of minimzers $\arg\min_s L_{\psi_1}(s,\eta)$ is precisely described by the conditions on s in the case.

1. $\eta_{[k]} > \sum_{i=k+1}^{M} \eta_{[i]} : \exists c \in \mathbb{R}, \pi \in S(\eta)$

   $s_{\pi_{k+1}} = \ldots = s_{\pi_M} = c, \quad s_{\pi_k} = c+1,$
   $\forall i \in \{1,\ldots,k-1\},\ s_{\pi_i} \in [c+1,\infty).$

2. $\eta_{[k]} < \sum_{i=k+1}^{M} \eta_{[i]} : \exists c \in \mathbb{R}, \pi \in S(\eta)$

   $s_{\pi_k} = \ldots = s_{\pi_M} = c,$
   $\forall i \in \{1,\ldots,k-1\},\ s_{\pi_i} \in [c+1,\infty).$

3. $\eta_{[k]} = \sum_{i=k+1}^{M} \eta_{[i]} : \exists c \in \mathbb{R}, \pi \in S(\eta)$

   $s_{\pi_{k+1}} = \ldots = s_{\pi_M} = c, \quad s_{\pi_k} \in [c, c+1],$
   $\forall i \in \{1,\ldots,k-1\},\ s_{\pi_i} \in [c+1,\infty).$

Proof. Suppose $\tau \in \Pi_M$ sorts s. Define $\delta := s_{\tau_k} - s_{\tau_{k+1}} = s_{[k]} - s_{[k+1]} \geq 0$. Since

$$\begin{aligned}&\max\{1 + s_{\tau_{k+1}} - s_{\tau_k}, 0\} \geq \max\{1-\delta, 0\} \\ &\max\{1 + s_{\tau_k} - s_{\tau_i}, 0\} \geq 1+\delta,\ \forall i \in \{k+1,\ldots,M\},\end{aligned}$$

$L_\psi(s,\eta)$ is lower bounded as follows:

$$\begin{aligned}L_\psi(s,\eta) &\geq \max\{1-\delta,0\}\eta_{\tau_k} + (1+\delta)\sum_{i=k+1}^{M} \eta_{\tau_i} \\ &\geq \max\{1-\delta,0\}\eta_{[k]} + (1+\delta)\sum_{i=k+1}^{M} \eta_{[i]} =: F(\delta).\end{aligned} \tag{5}$$

In the following, we discuss when equality in (5) is obtained in three cases. We may assume that $s_{\tau_{k+1}}$ is equal to an arbitrary $c \in \mathbb{R}$. Shifting each entry of s by a constant does not change the loss value. Before we begin, we note common requirements, regardless of case. Since η has no zero entries, the first line is an equality if and only if $s_{\tau_i} \geq s_{\tau_{k+1}} + 1 = c+1$ for all $i \in [k-1]$, and $s_{\tau_{k+1}} = s_{\tau_{k+2}} = \ldots = s_{\tau_M} = c$. And in any case where the second line is an equality, the sums on the right of both lines equal, which happens if and only if $\{\tau_{k+1},\ldots,\tau_M\} = \{\pi_{k+1},\ldots,\pi_M\}$ for some $\pi \in \Pi_M$ which sorts η.

Case 1: If $\eta_{[k]} > \sum_{i=k+1}^{M} \eta_{[i]}$, $F(\delta)$ is minimized uniquely at $\delta = 1$ in the interval $[0,1]$; by our assumption that η does not have 0 entries and $k < M$, $\delta > 1$ is suboptimal. Thus, $L^*_\psi(\eta) = 2\sum_{i=k+1}^{M} \eta_{[i]}$ (achieved by s described below).

The equality is achieved if and only if the common requirements hold and $\delta = 1$, giving $s_{\tau_k} = c+1$.

Case 2: If $\eta_{[k]} < \sum_{i=k+1}^{M} \eta_{[i]}$, then $F(\delta)$ is minimized by δ = 0, and $L^*_\psi(\eta) = \sum_{i=k}^{M} \eta_{[i]}$. Therefore, the equality holds if and only if $s_{\tau_k} = s_{\tau_{k+1}} = c$ and $\tau_k = \pi_k$ for some $\pi \in S_M$ which sorts η, along with the common requirements.

Case 3: If $\eta_{[k]} = \sum_{i=k+1}^{M} \eta_{[i]}$, then $L^*_\psi(\eta) = \sum_{i=k}^{M} \eta_{[i]} = 2\sum_{i=k+1}^{M} \eta_{[i]}$. Thus $F(\delta)$ is minimized by $\delta \in [0,1]$.

If $\delta \in (0,1)$, the inequality in (5) requires

$$\sum_{i=k}^{M} \eta_{\tau_i} = \sum_{i=k}^{M} \eta_{[i]} = 2\sum_{i=k+1}^{M} \eta_{\tau_i} = 2\sum_{i=k+1}^{M} \eta_{[i]}.$$

Thus, the equality holds if and only if in addition to the common requirements, $s_{\tau_k} \in (c, c+1)$, and for some $\pi \in S_M$ which sorts η, $\pi_k = \tau_k$.

If $\delta = 1$ or $\delta = 0$, we have the same iff conditions for the equality as in case 1 and case 2. □

Proposition 4.2. For any $\psi \in \{\psi_2, \psi_3, \psi_4\}$, if $\sum_{m=k+1}^{M} \eta_{[m]} > \frac{k}{k+1}$, we have $0 \in \arg\min_s L_\psi(s,\eta)$, and thus $L^*_\psi(\eta) = \min_s L_\psi(s,\eta) = L_\psi(0,\eta) = 1$.

Proof. We will show that $L^*_\psi(\eta) = 1$. WLOG, we can assume that $\eta_1 \geq \ldots \geq \eta_M$, $s_1 \geq s_2 \geq \ldots \geq s_M$, and $s_{k+1} = s_{k+2} = \ldots = s_M = 0$.

Suppose $s_i \geq 1$ for some $i \in [M]$. Then, for each $\psi \in \{\psi_2, \psi_3, \psi_4\}$, $\psi(s,i) \geq 1 + \frac{1}{k}$ for all $i \in \{k+1, \ldots, M\}$, and so $L_\psi(s,\eta) \geq \left(1 + \frac{1}{k}\right)(\eta_{k+1} + \ldots + \eta_m) > \frac{k+1}{k} \cdot \frac{k}{k+1} = 1$. This implies that s is suboptimal, since $L_\psi(0,\eta) = 1$.

Thus, at optimum $0 \leq s_i < 1$ for every i, under which $\psi_2(s,i) = \psi_3(s,i) = \psi_4(s,i)$ for every i. This is because in this regime, $\max\{1 + s_j - s_i, 0\} = 1 + s_j - s_i$, and the kth highest value of $\bar{\mathbf{1}}(i) + s$ coincides with the kth highest value of $1 + s$ excluding the ith index. Now for all $i \in [k]$, we have $s_i \in (0,1)$ and thus

$$\frac{\partial L_\psi(s,\eta)}{\partial s_i} = \frac{1}{k}\sum_{m\in[M], m\neq i} \eta_m - \eta_i = \frac{1}{k}(1-\eta_i) - \eta_i$$
$$> \frac{1}{k}\frac{k}{k+1} - \frac{1}{k+1} = 0.$$

The derivative is positive (and constant) in $(0,1)$, so the minimum value of s_i is achieved at 0, for every i. Therefore, $L^*_\psi(\eta) = 1$, achieved by a score vector of 0. This proves the desired statement. □

Proposition 4.3. $\psi_5 : \mathbb{R}^M \times \mathcal{Y}$ defined by $\psi_5(s,y) = \max\{1 + s_{[k+1]} - s_y, 0\}$ is top-k calibrated.

Proof. Let $\eta \in \Delta_M$. For any $s \in \mathbb{R}^M$, we have

$$L_{\psi_5}(s,\eta) = \sum_{i=1}^M \eta_i \psi_5(s,i) = \sum_{i=1}^M \eta_i \max\{1 + s_{[k+1]} - s_i, 0\}.$$

We may assume $\eta_1 \geq \eta_2 \geq \ldots \geq \eta_M$ WLOG. By inspection, setting $s_1 = \ldots = s_k = 1$ and $s_{k+1} = \ldots = s_M = 0$ gives $L_{\psi_5}(s,\eta) = \sum_{i=k+1}^M \eta_{[i]} =: C$.

We will show that any $s \in \mathbb{R}^M$ such that $\neg\mathsf{P}_\mathsf{k}(s,\eta)$ has $L_\psi(s,\eta) - L^*_\psi(\eta) \geq L_\psi(s,\eta) - C \geq \delta$ for some constant $\delta > 0$, which implies top-k calibration.

Suppose $\neg\mathsf{P}_\mathsf{k}(s,\eta)$. Define $\delta_1 = \min\{\eta_i - \eta_{[k+1]} \mid i \in [M], \eta_i > \eta_{[k+1]}\}$ and $\delta_2 = \min\{\eta_{[k]} - \eta_i \mid i \in [M], \eta_i < \eta_{[k]}\}$. If either set is empty, define its minimum to be ∞. Furthermore, define the set $I := \{i \in [M] \mid s_i \leq s_{[k+1]}\}$. Note by definition of $s_{[k+1]}$, $|I| \geq M - k$. We have $L_\psi(s,\eta) \geq \sum_{i\in I} \eta_i$. There are two cases.

If there exists $i \in [M]$ such that $\eta_i > \eta_{[k+1]}$ and $s_i \leq s_{[k+1]}$, then $i \in I$. But then $\sum_{j\in I} \eta_j \geq \sum_{j=k+1}^M \eta_{[j]} + \delta_1$.

If there exists $i \in [M]$ such that $\eta_i < \eta_{[k]}$, but $s_i \geq s_{[k]}$, then consider if $s_i > s_{[k+1]}$. Then, $i \notin I$. That is, η_i does not appear in the sum $\sum_{j\in I} \eta_j$. Since $|I| \geq M - k$, η_i must be replaced with a term $\eta_{i'} \geq \eta_{[k]}$. Thus, $\sum_{j\in I} \eta_j \geq \sum_{j=k+1}^M \eta_{[j]} + \delta_2$. If $s_i = s_{[k+1]}$, then since $s_i \geq s_{[k]} \geq s_{[k+1]}$, we have $s_i = s_{[k]}$. This implies $|I| > M - k$, and $\sum_{j\in I} \eta_j \geq \sum_{j=k}^M \eta_{[j]} \geq \sum_{j=k+1}^M \eta_{[j]} + \delta_2$.

Thus, for any s such that $\neg\mathsf{P}_\mathsf{k}(s,\eta)$, we have $L_\psi(s,\eta) \geq L^*_\psi(\eta) + \delta$ where $\delta = \min\{\delta_1, \delta_2\} > 0$. Therefore,

$$\inf_{s:\neg\mathsf{P}_\mathsf{k}(s,\eta)} L_\psi(s,\eta) \geq \inf_s L_\psi(s,\eta) + \delta > \inf_s L_\psi(s,\eta),$$

so $\psi = \psi_5$ is top-k calibrated. □

Proposition 5.1. Let $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{F} = \{x \mapsto Wx : W \in \mathbb{R}^{M\times d}\}$. Then if we consider *top-k separable* probability distributions over $\mathcal{X} \times \mathcal{Y}$, i.e. $L^*_{\text{err}_k}(\mathcal{F}) = 0 = L^*_{\text{err}_k}$, then:

1. If $k = 1$, Ent is $\mathcal{F}$-consistent.

2. If $d \geq 3$, $M \geq 3$, and $k = 2$, Ent is not $\mathcal{F}$-consistent.

3. ψ_1 and ψ_5 are $\mathcal{F}$-consistent.

Proof. **Proof of 2.** Let $\mathcal{X} = \mathbb{R}^3$, $M = 3$, and $k = 2$. It does not matter if we increase dimensions or M. Let the dataset S consist of the following 7 points, where e_i denotes the standard basis element with a 1 in the ith coordinate: $S = [2\times(e_1, 1), 2\times(e_2, 2), 2\times(e_3, 2), (-e_1, 1)] \subset \mathcal{X}\times\mathcal{Y}$. The intuition is that having e_1 and $-e_1$ both labeled 1 blatantly precludes linear separability.

Note that S is top-2 separable, since W_{sep} is a top-2 separator for S:

$$W_{\text{sep}} = \begin{bmatrix} 2 & 0 & 0 \\ 1 & 1 & 0 \\ 3 & 0 & 1 \end{bmatrix}.$$

The following score vectors are returned for each input:

$$W_{\text{sep}} e_1 = \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix} \quad W_{\text{sep}} e_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$
$$W_{\text{sep}} e_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \quad W_{\text{sep}}(-e_1) = \begin{bmatrix} -2 \\ -1 \\ -3 \end{bmatrix}.$$

$W_{\text{sep}} e_2$ and $W_{\text{sep}} e_3$ respectively have their second and third entries as their strictly greatest entries. Since they are respectively labeled 2 and 3, they are classified correctly. $W_{\text{sep}} e_1$ and $W_{\text{sep}}(-e_1)$ both have their first entry strictly greater than their third entry. This means they are classified correctly by a top-2 classifier, as their label is 1: $(W_{\text{sep}} e)_1 > (W_{\text{sep}} e)_{[3]}$ for $e \in \{e_1, -e_1\}$.

Now we show the solution returned by cross entropy minimization is not a top-k separator. $\text{Ent}(W, S)$ denotes the cross entropy loss incurred by W on the probability distribution defined by the dataset S, times the number of samples.

$$
\begin{aligned}
\operatorname{Ent}(W,S) &= \sum_{(x,y)\in S} \log\left(\sum_{m=1}^{3} e^{(Wx)_m-(Wx)_y}\right)\\
&= 2\sum_{i=1}^{3}\log\left(\sum_{m=1}^{3} e^{(We_i)_m-(We_i)_i}\right)\\
&\quad + \log\left(\sum_{m=1}^{3} e^{(We_1)_1-(We_1)_m}\right)\\
&= 2\sum_{i=1}^{3}\log\left(\sum_{m=1}^{3} e^{W_{mi}-W_{ii}}\right)\\
&\quad + \log\left(\sum_{m=1}^{3} e^{W_{11}-W_{m1}}\right).
\end{aligned}
$$

For each different i, the entries of W appearing in the ith term in the sum correspond to different columns of W – entries appearing in different terms are independent of each other. For $i \neq 1$, we see that $\log\left(\sum_{m=1}^{3} e^{W_{mi}-W_{ii}}\right) = \log\left(1+\sum_{m\neq i} e^{W_{mi}-W_{ii}}\right)$ can be taken to 0 by taking $W_{mi}-W_{ii} \to -\infty$ for each $m \neq i$. We cannot do the same for $i=1$ because of the appearance of both $(e_1,1)$ and $(-e_1,1)$. But at this point, we have gotten rid of terms with $i \neq 1$ and determined that the minimizer of Ent looks like the following:

$$
W_{CE} = \begin{bmatrix} ? & W_{22}-\infty & W_{33}-\infty \\ ? & W_{22} & W_{33}-\infty \\ ? & W_{22}-\infty & W_{33} \end{bmatrix}.
$$

The remainder of the loss function is

$$
\begin{aligned}
\operatorname{Ent}(W,S) =& 2\log\left(1+e^{W_{21}-W_{11}}+e^{W_{31}-W_{11}}\right)\\
&+\log\left(1+e^{W_{11}-W_{21}}+e^{W_{11}-W_{31}}\right).
\end{aligned}
$$

Denote $x_1 = W_{21}-W_{11}$ and $x_2 = W_{31}-W_{11}$, so we may write the loss as

$$
\operatorname{Ent}(W,S) = 2\log\left(1+e^{x_1}+e^{x_2}\right)+\log\left(1+e^{-x_1}+e^{-x_2}\right).
$$

We have

$$
\begin{aligned}
\frac{\partial \operatorname{Ent}}{\partial x_1} &= \frac{2e^{x_1}}{1+e^{x_1}+e^{x_2}} - \frac{e^{-x_1}}{1+e^{-x_1}+e^{-x_2}},\\
\frac{\partial \operatorname{Ent}}{\partial x_2} &= \frac{2e^{x_2}}{1+e^{x_1}+e^{x_2}} - \frac{e^{-x_2}}{1+e^{-x_1}+e^{-x_2}}.
\end{aligned}
$$

By the convexity of $\log(1+e^{x_1}+e^{x_2})$, we may minimize the function by setting the derivatives equal to 0. Note that if $x_1 \neq x_2$, this is not achievable – suppose it were the case that $\frac{\partial \operatorname{Ent}}{\partial x_1} = 0$. If $x_2 > x_1$, then $e^{x_2} > e^{x_1}$ and $e^{-x_2} < e^{-x_1}$, so $\frac{\partial L}{\partial x_2} > \frac{\partial L}{\partial x_1} = 0$. A similar argument holds if $x_2 < x_1$. Therefore, we may assume $x_1 = x_2$. Then we simply need

$$
\begin{aligned}
\frac{2e^{x}}{1+2e^{x}} - \frac{e^{-x}}{1+2e^{-x}} = 0 &\iff 2e^{x}+4-e^{-x}-2=0\\
&\iff 2e^{x}-e^{-x}=-2.
\end{aligned}
$$

If $x \geq 0$, then clearly $2e^{x}-e^{-x} > 0$. Thus, $x < 0$ (we can solve a quadratic, or note that there exists x where $2e^{x}-e^{-x}=-2$ because the LHS goes to $-\infty$ as $x \to -\infty$). Therefore, at minimum $W_{21}-W_{11} = W_{31}-W_{11} = x$ for some $x < 0$, so the cross entropy minimizer is the following:

$$
W_{CE} = \begin{bmatrix} W_{11} & W_{22}-\infty & W_{33}-\infty \\ W_{11}+x & W_{22} & W_{33}-\infty \\ W_{11}+x & W_{22}-\infty & W_{33} \end{bmatrix}.
$$

This is not a top-2 separator because $W_{CE}(-e_1) = \begin{bmatrix}-W_{11}, & -W_{11}-x, & -W_{11}-x,\end{bmatrix}^{\top}$, whose first entry is strictly the lowest entry since $x < 0$. Thus, $-e_1$ is not classified as its label, 1.

Proof of 3. Recall ψ_1, ψ_5:

$$
\begin{aligned}
\psi_1(s,y) &= \max\{1+(s_{\setminus y})_{[k]}-s_y, 0\},\\
\psi_5(s,y) &= \max\{1+s_{[k+1]}-s_y, 0\}.
\end{aligned}
$$

We will show these losses are linearly top-k consistent. Suppose $S = ((x_1,y_1),\ldots,(x_n,y_n))$ is top-k separable, that is, $\exists\, W \in \mathbb{R}^{M\times d}$ such that $\forall\, i \in [n],\ (Wx_i)_{y_i} > (Wx_i)_{[k+1]}$. In other words, there is a $\delta > 0$ such that for every $i \in [n],\ (Wx_i)_{y_i} - (Wx_i)_{[k+1]} \geq \delta$. Then, for $C \geq \frac{1}{\delta},\ (CWx_i)_{y_i} - (CWx_i)_{[k+1]} \geq C\delta \geq 1$ for every $i \in [n]$.

Now let $i \in [n]$ and denote $s = CWx_i$. Since $s_{y_i} > s_{[k+1]}$, we have $s_{[k+1]} = (s_{\setminus y})_{[k]}$. Thus,

$$
\psi_1(s,y_i) = \psi_5(s,y_i) = \max\{1+s_{[k+1]}-s_{y_i}, 0\} = 0.
$$

Therefore, CW achieves 0 loss on the dataset for both ψ_1 and ψ_5. This means their minimizers (over linear functions) achieve 0 loss. If 0 loss is achieved, it is clear that the resulting classifiers achieve 0 top-k error, since these losses upper bound the top-k error. Therefore, their minimizers are top-k separators.

We have shown that if a dataset is linearly top-k separable, then the minimizers of ψ_1 and ψ_5 are top-k linear separators for the dataset. This proves that ψ_1 and ψ_5 are linearly top-k □

10. Discussion of general hinge-like losses

Recall that the hinge loss for binary classification is defined by $\phi(x) = \max\{1-x, 0\}$. There are several extensions of the binary hinge loss to the setting of multiclass classification (often with multiclass error i.e. top-1 loss). We list them here because they serve as inspiration for designing hinge-like top-k losses, and the analysis of their consistency in the literature also informs the analysis of the top-k case.

The method of Crammer & Singer (2001) uses as its loss function $\psi : \mathbb{R}^M \times \mathcal{Y} \to \mathbb{R}$ where

$$\psi(s, y) = \max\{1 + (s_{\setminus y})_{[1]} - s_y, 0\} = \phi(s_y - \max_{y' \neq y} s_{y'}). \tag{6}$$

When $y \in \mathcal{Y}$ appears in a subscript it refers to the label as an index in $\{1, \ldots, M\}$. Furthermore, the notation $s_{\setminus y} = (s_1, \ldots, s_{y-1}, s_{y+1}, \ldots, s_M) \in \mathbb{R}^{M-1}$ denotes the vector s with the yth entry removed.

The method of Weston & Watkins (1999) solves a multiclass SVM problem for which the corresponding loss function is

$$\psi(s, y) = \sum_{y' \neq y} \phi(s_y - s_{y'}),$$

where ϕ is still the binary hinge loss. Furthermore, the one vs. all method Rifkin & Klautau (2004) solves M binary classification problems using the hinge loss for each class, using the instances of the class as positive examples and the rest of the instances as negative examples. The M scores returned by the M resulting classifiers are compiled into an M length vector, and the method proceeds like all the above methods by taking the argmax of the vector. Similarly, the method of Lee et al. (2004) minimizes the expectation of the loss function

$$\psi(s, y) = \sum_{y' \neq y} \phi(-s_{y'})$$

under the constraint that $\sum_{m=1}^{M} s_m = 0$. Interestingly, Zhang (2004a) showed the first three Crammer & Singer (2001); Weston & Watkins (1999); Rifkin & Klautau (2004) to be inconsistent, i.e. not top-1 calibrated, and the constrained Lee et al. (2004) to be consistent. These results were also found by Tewari & Bartlett (2005).

Table 5. Examples of predicted score vector $s = f(0)$ with the zero vector as input, where f is a neural net trained with the losses below.

	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8
ψ_1	0.87793601	-0.12823531	-0.12382337	-0.12676451	-0.12382337	-0.12235278	-0.12529394	-0.12764691
ψ_2	0.00176411	0.00044059	-0.00058873	-0.00176518	-0.00220636	0.0002936	0.00073477	0.00132302
ψ_3	0.00117588	0.00191117	0.00102892	-0.0010299	-0.0020593	-0.00029462	0.00073478	-0.00147108
ψ_4	0.00073472	0.00161706	0.00029361	-0.00264753	0.00117595	0.00088184	-0.00191224	-0.00014757
ψ_5	0.75734961	0.75734961	-0.25529474	-0.24823636	-0.2523534	-0.24823636	-0.25529483	-0.25529486