Supplementary Material: Guess-Averse Loss Functions For Cost-Sensitive Multiclass Boosting

Oscar Beijbom OBEIJBOM@UCSD.EDU
Mohammad Saberian SABERIAN@UCSD.EDU
David Kriegman KRIEGMAN@UCSD.EDU
Nuno Vasconcelos NVASCONCELOS@UCSD.EDU
University of California, San Diego, 9500 Gilman Drive, 92093 La Jolla, CA

A. Properties of $L_t(C, z, S(x))$ loss function

Lemma 1. *The loss function*

$$L_t(C, z, S(x)) = \sum_{j,k=1}^{M} C_{z,j} e^{S_j(x) - S_k(x)} \tag{1}$$

is Classification Calibrated.

Proof. Using (1), risk of classification is

$$\begin{aligned} \mathcal{R}_{L_t}[S] &= E_{X,Z}\{L_t[C, z, S(x)] && (2)\\ &= \sum_{z=1}^{M} \eta_z(x) L_t(C, z, S(x)) && (3)\\ &= \sum_{z=1}^{M} \eta_z(x) \sum_{j,k=1}^{M} C_{z,j} e^{S_j(x) - S_k(x)} && (4)\\ &= \sum_{z=1}^{M} \sum_{j,k=1}^{M} \eta_z(x) C_{z,j} e^{S_j(x) - S_k(x)}. && (5) \end{aligned}$$

To find the optimal scores, $S^*(x)$, we start by setting derivatives to zero, where

$$\begin{aligned} 0 = \frac{\partial \mathcal{R}_{L_t}[S]}{\partial S_i(x)} &= \sum_{z=1}^{M} \eta_z(x) C_{z,i} e^{S_i(x)} \sum_{k=1}^{M} e^{-S_k(x)} && (6)\\ &\quad - e^{-S_i(x)} \sum_{z=1}^{M} \sum_{j=1}^{M} \eta_z(x) C_{z,j} e^{S_j(x)}, && (7) \end{aligned}$$

results in

$$e^{-2S_i(x)} = \sum_{z=1}^{M} \eta_z(x) C_{z,i} \frac{\sum_{k=1}^{M} e^{-S_k(x)}}{\sum_{z,j=1}^{M} \eta_z(x) C_{z,j} e^{S_j(x)}}. \tag{8}$$

Assuming $\sum_{i=1}^{M} S_i(x) = 0$, and defining

$$\psi(i) = \sum_{z=1}^{M} \eta_z(x) C_{z,i} \tag{9}$$

results in

$$S^*{}_i(x) = -\frac{1}{2} \log\left(\psi(i)\right) + \frac{1}{2M} \sum_{j=1}^{M} \log(\psi(j)). \tag{10}$$

Therefore $S^*{}_i(x)$ will be inversely proportional to Bayes cost of i^{th} class and thus (1) will be classification calibrated. □

Lemma 2. *$L_t(C, z, S(x))$ is not guess-averse.*

Proof. The proof is based on a counter example. Assume a cost insensitive problem i.e $C_{i,j} = 1 \quad i \neq j$, with $M = 3$ and $S(x) = [3, 0, -3]$ where example x belongs to the first class. In this case $S(x)$ results in the correct prediction but its loss is greater than random guessing since $L(C, 1, 0) = 6 < L(C, 1, S(x)) \approx 22.19$. Therefore (1) is not guess-averse. □

B. Comparing Guess-aversion and c-calibration

We start with following lemma that shows that c-calibration (**?**) implies guess-aversion.

Lemma 3. *If a loss function $L(C, z, S(x))$ is c-calibrated, then it will be guess-averse.*

Proof. If $L(C, z, S(x))$ is c-calibrated, then according to c-calibration definition

$$\forall s_1 \in \mathcal{S}_z \quad \forall s_2 \notin \mathcal{S}_z, L(C, z, s1) < L(C, z, s2). \tag{11}$$

In addition note that, $\mathbf{0} \notin \mathcal{S}_z$ and using (11)

$$\forall s_1 \in \mathcal{S}_z \quad L(C, z, s1) < L(C, z, \mathbf{0}). \tag{12}$$

which is definition of guess-aversion. □

We next show that guess-aversion does not guarantee c-calibration.

Lemma 4. *If a loss function $L(C, z, S(x))$ is guess-averse, then it may not be c-calibrated.*

Proof. The proof is based on counter example. Consider the cost-insensitive GLL-loss of Figure 3-b. Since it satisfies Lemma 1 in the paper this loss is guess-averse. However, for sufficiently small $\epsilon > 0$, the set $\mathcal{A}_\epsilon = \{S | S \in \mathcal{S}_2, L(C, 1, S) < L(C, 1, 0) - \epsilon\}$ is non-empty. Similarly, since the loss surface is continuous and smooth in Figure 3-b, there exists a point $p_0 \in \mathcal{S}_1$ such that $L(C, 1, p_0) > L(C, 1, 0) - \epsilon$. Therefore for any $q_0 \in A_\epsilon$, $L(C, 1, p_0) > L(C, 1, q_0)$ which is contradictory to c-calibration, since p_0 results in correct classification and q_0 does not. □

C. Properties of the Generalized Exponential Loss

Lemma 5. *If $C_{i,j} \geq 0 \; \forall i, j = 1 \dots M$ and $\exists i, j : C_{i,j} > 0$ then*

$$\begin{aligned} \mathcal{R}_{L^{id,\exp}} &= E_{X,Z}\{L^{id,\exp}(C, z, S(x))\} \\ &= \sum_{z,j=1}^{M} \eta_z(x) C_{z,j} e^{S_j(x) - S_z(x)}, \end{aligned} \quad (13)$$

is strictly convex with respect to $S(x) \in \mathbb{R}^M$.

Proof. Denoting $\beta_{i,j} = \mathbf{1}_i - \mathbf{1}_j$, we start by computing first and second order derivatives,

$$\begin{aligned} \frac{\partial \mathcal{R}_{L^{\text{id},\exp}}}{\partial S} &= \frac{\partial}{\partial S} \sum_{z,j=1}^{M} \eta_z C_{z,j} e^{\langle S, \beta_{j,z} \rangle} \\ &= \sum_{z,j=1}^{M} \eta_z C_{z,j} \beta_{j,z} e^{\langle S, \beta_{j,z} \rangle} \end{aligned} \quad (14)$$

$$\begin{aligned} \frac{\partial^2 \mathcal{R}_{L^{\text{id},\exp}}}{\partial S^2} &= \frac{\partial}{\partial S} \sum_{z,j=1}^{M} \eta_z C_{z,j} \beta_{j,z} e^{\langle S, \beta_{j,z} \rangle} \\ &= \sum_{z,j=1}^{M} \eta_z C_{z,j} [\beta_{j,z} \beta_{j,z}^T] e^{\langle S, \beta_{j,z} \rangle}. \end{aligned} \quad (15)$$

where we omitted x for simplicity. Note that $[\beta_{j,z} \beta_{j,z}^T]$ is positive definite for all z, j, moreover $C_{i,j} \geq 0 \quad \forall i, j = 1 \dots M$ and $\exists i, j : C_{i,j} > 0$. Therefore the hessian is a sum of positive definite matrices, and is a positive definite matrix. Therefore $\mathcal{R}_{L^{\text{id},\exp}}$ is strictly convex. □

Lemma 6. *If the cost matrix, C, is symmetric then the minimizer of $\mathcal{R}_{L^{id,\exp}(C,z,S(x))}$, (13), is independent of C.*

Proof. We start by setting (14) to zero, therefore

$$\sum_{z,j=1}^{M} \eta_z C_{z,j} \mathbf{1}_j e^{\langle S, \beta_{j,z} \rangle} = \sum_{z,j=1}^{M} \eta_z C_{z,j} \mathbf{1}_z e^{\langle S, \beta_{j,z} \rangle} \quad (16)$$

Table 1. Cost Matrix for MLC (**?**).

	CCA	*Turf*	*Macro*	*Sand*	*Acro.*	*Pav.*	*Mon.*	*Pocil.*	*Porit*
CCA	0	1	1	2	4	4	4	4	4
Turf	1	0	1	2	4	4	4	4	4
Macro	1	1	0	2	4	4	4	4	4
Sand	2	2	2	0	4	4	4	4	4
Acropora	4	4	4	4	0	1	1	1	1
Pavona	4	4	4	4	1	0	1	1	1
Monti	4	4	4	4	1	1	0	1	1
Pocill	4	4	4	4	1	1	1	0	1
Porit	4	4	4	4	1	1	1	1	0

and thus

$$\sum_{j=1}^{M} \eta_k C_{k,j} e^{S_j - S_k} = \sum_{z=1}^{M} \eta_z C_{z,k} e^{S_k - S_z}. \quad (17)$$

However note that when C is symmetric,

$$S_k(x) = \frac{1}{2} \log(\eta_k(x)) - \frac{1}{2M} \sum_{j}^{M} \log(\eta_j(x)) \quad (18)$$

satisfies (17). This is because $e^{S_j - S_k} = \frac{\sqrt{\eta_j}}{\sqrt{\eta_k}}$ and thus left and right sides of (17)

$$\begin{aligned} \sum_{j=1}^{M} \eta_k C_{k,j} e^{S_j - S_k} &= \sum_{j=1}^{M} C_{k,j} \sqrt{\eta_j \eta_k} \quad (19) \\ \sum_{z=1}^{M} \eta_z C_{z,k} e^{S_k - S_z} &= \sum_{z=1}^{M} C_{z,k} \sqrt{\eta_z \eta_k}. \quad (20) \end{aligned}$$

become equal. Therefore (18) is a minimizer of $\mathcal{R}_{L^{\text{id},\exp}(C,z,S(x))}$. In addition according to lemma (5), $\mathcal{R}_{L^{\text{id},\exp}}$ is strictly convex and thus (18) will be the unique minimizer. □

D. MLC - Cost Matrix

The cost matrix for the Moorea Labelled Corals dataset is shown in Table 1. The costs are set with an coral ecology application in mind. There, the most important goal is a binary estimate of the amount of corals versus everything else. Thus, the cost of confusion between the coral genera (classes 5-9) and the non-corals (classes 1-4) is set to a high value, 4. Cost of confusion among corals is low, 1, and similarly for cost of confusion among algae (classes 1-3). Finally, confusion between any algae and the sand class is worse than confusion within algae, but not as bad as confusion to (or from) corals. These values are set to 2.

E. Structured SVMs are guess-averse

Let $\mathcal{Y} = \{Y_1, \dots Y_M\}$ be a set of structured outputs. For a training set $\mathcal{D} = \{(x_i, Y_{z_i})\}_1^n$, where $z_i \in \{1 \dots M\}$, a

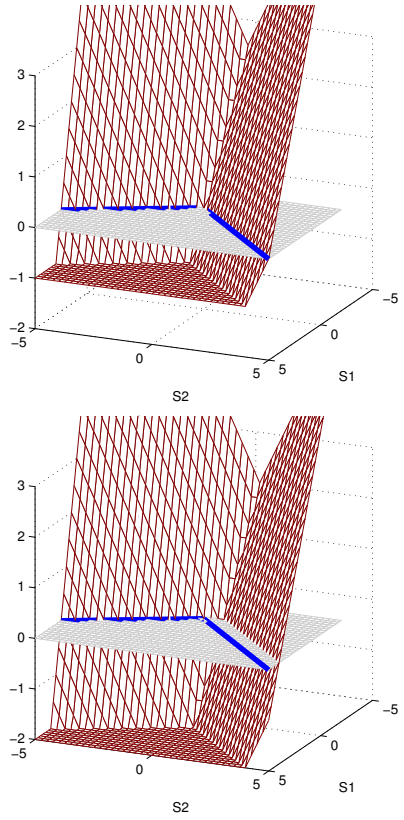

Figure 1. **Structured SVM loss functions:** Cost-insensitive (top), and cost-sensitive, with $C_{1,2} = 1, C_{1,3} = 2$, (bottom).

structured SVM (**?**) solves

$$\begin{cases} \min_{\mathbf{w},\epsilon} & \frac{\lambda}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{n} \epsilon_i \\ \text{s.t.,} & \forall_{i,z\in\mathcal{Z}},\ \ S_z(x_i) + C_{z_i,z} \leq S_{z_i}(x_i) + \epsilon_i \\ & \epsilon_i \geq 0\ \forall i, \end{cases} \tag{21}$$

where

$$S_z(x) = \langle \mathbf{w}, \Psi(x, z) \rangle \tag{22}$$

is the score of structure Y_z for the example x, $\Psi(x, z)$ is a feature vector extracted with respect to structure Y_z, and $C_{z_i,z} \geq 0$ is the cost of assigning structure Y_z instead of the true structure Y_{z_i}.

An equivalent way of writing (21) is

$$\min_{\mathbf{w}} \frac{\lambda}{2}\|\mathbf{w}\|^2 + \sum_i L_H(C, z_i, S(x_i)) \tag{23}$$

where

$$L_H(C, z_i, S(x_i)) = \max_k (S_k(x_i) + C_{z_i,k} - S_{z_i}(x_i)), \tag{24}$$

is the loss function for structured SVM. Similar to Figure 3 of the paper, loss surfaces for L_H, (24), are shown in Figure 1. Note how, in the bottom figure, the surface shifts away the boundary between $\mathcal{S}_1$ and $\mathcal{S}_2$, in a similar manner as the cost-sensitive $L^{\log,\exp}$ did. This is not surprising as the logistic function approximates the hinge loss. Finally, the following lemma shows that L_H is guess-averse.

Lemma 7. *The loss function for structured SVM,(24), is guess-averse.*

Proof. Let x be a sample corresponding to a structure Y_z, $S \in \mathbb{R}^M$ the classifier score vector and C a non-negative cost function. First note that using (24) if $S = \mathbf{0} \in \mathbb{R}^M$,

$$L_H(C, z, \mathbf{0}) = \max_k C(z, k). \tag{25}$$

Second, if x is correctly classified, i.e. $S(x) \in \mathcal{S}_z$, then $S_z(x) > S_k(x)\ \forall k \neq z$ and thus using (24), (25)

$$\begin{aligned} L_H(C, z, S(x)) &= \max_k \left[C_{z,k} + (S_k(x) - S_z(x))\right] \\ &< \max_k (C_{z,k}) \\ &= L_H(C, z, \mathbf{0}). \end{aligned}$$

Therefore if $S(x) \in \mathcal{S}_z$, then $L_H(C, z, S(x)) < L_H(C, z, \mathbf{0})$ and thus L_H is guess-averse. □