A Margin upper bound

From (15) and (10)

$$
\begin{aligned}
L_M[y, f(x)] &= \sum_{k=1}^{M} e^{-\frac{1}{2}[<f(x),y>-<f(x),y^k>]} && (39)\\
&= 1 + \sum_{y^k \neq y}^{M} e^{-\frac{1}{2}[<f(x),y>-<f(x),y^k>]} && (40)\\
&= 1 + e^{-\mathcal{M}(f(x),y)} \sum_{y^k \neq y} e^{-\frac{1}{2}[<f(x),y>-<f(x),y^k>]+\mathcal{M}(f(x),y)} && (41)\\
&= 1 + e^{-\mathcal{M}(f(x),y)} \sum_{y^k \neq y} e^{-\frac{1}{2}[\max_{y^l \neq y}<f(x),y^l>-<f(x),y^k>]} && (42)\\
&= 1 + e^{-\mathcal{M}(f(x),y)} \left[1 + \sum_{y^k \neq y, y^{l^*}} e^{-\frac{1}{2}[<f(x),y^{l^*}>-<f(x),y^k>]} \right] && (43)\\
&\geq 1 + e^{-\mathcal{M}(f(x),y)} && (44)
\end{aligned}
$$

where $l^* = \arg\max_{l|y^l \neq y} < f(x), y^l >$.

B Convexity

Defining $\beta_k = P_{Y|X}(y^k|x)$ and using (13)

$$
\begin{aligned}
R(f|x) &= E_{Y|X}\{L_M[y, f(x)]|x\} && (45)\\
&= \sum_{k=1}^{M} \beta_k L_M[y^k, f(x)] && (46)\\
&= \sum_{k=1}^{M} \beta_k \sum_{j=1}^{M} e^{-\frac{1}{2}<f(x),y^k-y^j>} = \sum_{k=1}^{M}\sum_{j=1}^{M} \beta_k e^{-\frac{1}{2}<f(x),y^k-y^j>}. && (47)
\end{aligned}
$$

Denoting $\eta_{i,j} = y^i - y^j$ the functional derivatives of first and second order, with respect to $f(x)$, are

$$
\begin{aligned}
\frac{\partial R(f|x)}{\partial f(x)} &= -\frac{1}{2} \sum_{k=1}^{M}\sum_{j=1}^{M} \beta_k \eta_{k,j} e^{-\frac{1}{2}<f(x),\eta_{k,j}>} && (48)\\
\frac{\partial^2 R(f|x)}{\partial f(x)^2} &= \frac{1}{4} \sum_{k=1}^{M}\sum_{j=1}^{M} \beta_k [\eta_{k,j}\eta_{k,j}^T] e^{-\frac{1}{2}<f(x),\eta_{k,j}>}. && (49)
\end{aligned}
$$

If all codewords are different, i.e. $\eta_{k,j} \neq 0 \; \forall k, j$, the matrices $[\eta_{k,j}\eta_{k,j}^T]$ are positive definite $\forall k, j$. Since $\beta_k \geq 0 \quad \forall k$ and $\sum_{k=1}^{M} \beta_k = 1$, $\beta_j > 0$ for at least one j and (49) is strictly positive definite. Hence, $R(f|x)$ is strictly convex, and has a *unique* global minimum.

C Bayes Consistency

Using $\beta_k = P_{Y|X}(y^k|x)$, setting

$$
e^{-\frac{1}{2}<f(x),\eta_{k,j}>} = \sqrt{\frac{\beta_j}{\beta_k}} \tag{50}
$$

and substituting in (48)

$$\frac{\partial R(f|x)}{\partial f(x)} = -\frac{1}{2}\sum_{k=1}^{M}\sum_{j=1}^{M}\beta_k\eta_{k,j}\sqrt{\frac{\beta_j}{\beta_k}} \quad (51)$$

$$= -\frac{1}{2}\sum_{k=1}^{M}\sum_{j=1}^{M}(y^k - y^j)\sqrt{\beta_j\beta_k} \quad (52)$$

$$= -\frac{1}{2}\sum_{k=1}^{M}y^k\sqrt{\beta_k}\sum_{j=1}^{M}\sqrt{\beta_j} + \frac{1}{2}\sum_{j=1}^{M}y^j\sqrt{\beta_j}\sum_{k=1}^{M}\sqrt{\beta_k} = 0 \quad (53)$$

Hence, when $f(x)$ is $f^*(x)$, the unique minimum of $R(f|x)$, (50) holds. It follows that

$$< f^*(x), y^i > - < f^*(x), y^k >= \log\frac{P_{Y|X}(y^i|x)}{P_{Y|X}(y^k|x)} \quad (54)$$

and

$$< f^*(x), y^i >= \log P_{Y|X}(y^i|x) + c.\forall i \quad (55)$$

for some constant c. This shows that (11) is equivalent to (18).

D Underdetermined predictor

Let $d > M-1$ and consider a set of M vectors $y^1, ...y^M \in \mathbb{R}^d$. There are three possibilities

1. If $d > M$ then $y^1, ...y^M$ belong to an at most M dimensional subspace $\mathcal{S}$ of $\mathbb{R}^d$. $\mathcal{S}'$, the orthogonal complement of $\mathcal{S}$, is nonempty and $\mathbb{R}^d = \mathcal{S}' \cup \mathcal{S}$. Since, by definition

$$\forall u \in \mathcal{S}, v \in \mathcal{S}', \quad < u, v >= 0, \quad (56)$$

any $v \in \mathcal{S}'$ satisfies (20).

2. If $d = M$ and $y^1, ...y^M$ are linearly dependent, then $y^1, ...y^M$ belong to an, at most, $M-1$ dimensional subspace $\mathcal{S}$ of $\mathbb{R}^M$. It follows that $\mathcal{S}'$, the orthogonal complement of $\mathcal{S}$, is nonempty. As in the previous case, this implies the existence of a $v \in \mathcal{S}'$ that satisfies (20).

3. If $d = M$ and $y^1, ...y^M \in \mathbb{R}^M$ are linearly independent, then, the matrix Y of rows $y^1, ...y^M$ is invertible and

$$Yv = \mathbf{1}, \quad (57)$$

with $\mathbf{1} = [1, ..1]^T \in \mathbb{R}^M$, has a unique solution. This solution satisfy (20) since

$$< y^i, v >= 1 \quad \forall i \quad (58)$$

E Solving optimization problem of codewords

Lemma 1. *Consider the set of distinct unit vectors $y^1, ...y^M \in \mathbb{R}^{M-1}$ of smallest pairwise distance*

$$d^2_{min} = \min_{i\neq j}||y^i - y^j||^2. \quad (59)$$

Then

$$d^2_{min} \leq \frac{2M}{(M-1)} \quad (60)$$

Proof. Since the minimum distance cannot be larger than the average distance between the vectors,

$$d^2_{min} \leq \frac{2}{M(M-1)}\sum_i\sum_{j\neq i}||y^i - y^k||^2. \quad (61)$$

To derive the bound of (60), we consider the following problem.

$$\begin{cases} \max y^1, ...y^M & \frac{1}{2}\sum_i \sum_{j\neq i} ||y^i - y^j||^2 \\ s.t & ||y^k|| = 1 \quad \forall k = 1..M. \end{cases} \tag{62}$$

This problem can be solved with the Lagrange multiplier method. The Lagrangian is

$$\mathcal{L} = \frac{1}{2}\sum_i \sum_{j\neq i} ||y^i - y^j||^2 - \sum_i \sigma_i(||y^i||^2 - 1) \tag{63}$$

for which

$$\frac{\partial \mathcal{L}}{\partial y^k} = -\sum_{i\neq k}(y^i - y^k) + \sum_{i\neq k}(y^k - y^i) - 2\sigma_k y^k = 2y^k(M - \sigma_k) - 2\sum_{i=1}^M y^i \tag{64}$$

$$\frac{\partial^2 \mathcal{L}}{\partial {y^k}^2} = 2(M - \sigma_k - 1). \tag{65}$$

This has a maximum when

$$y^k(M - \sigma_k) = \sum_{i=1}^M y^i \tag{66}$$

$$M - \sigma_k \leq 1 \tag{67}$$

We next consider two possibilities for $M - \sigma_k$.

1. $\exists k$ such that $M - \sigma_k = 0$. In this case, it follows from (66) that $\sum_{i=1}^M y^i = 0$. Since $||y^k|| = 1$, it follows that $M - \sigma_k = 0 \forall k$.

2. $M - \sigma_k \neq 0 \ \forall k$. Then, from (66)

$$y^k = \frac{\sum_{i=1}^M y^i}{M - \sigma_k}, \forall k. \tag{68}$$

Since the y^k have to be distinct, there can be no pair such that $M - \sigma_{k_1} = M - \sigma_{k_2}$. Using $||y^k|| = 1$, it follows from (68) that

$$(M - \sigma_k)^2 = \|\sum_{i=1}^M y^i\|^2 \tag{69}$$

and, for any k,

$$(M - \sigma_k) \in \{-||\sum_{i=1}^M y^i||, +||\sum_{i=1}^M y^i||\}. \tag{70}$$

Hence, there is a contradiction for $M > 2$. For $M = 2$ the contradiction can be avoided if $M - \sigma_{k_1} = -(M - \sigma_{k_2})$. In this case, using (68),

$$\sum_{i=1}^M y^i = y^1 + y^2 = \left[\frac{1}{M - \sigma_{k_1}} + \frac{1}{M - \sigma_{k_2}}\right]\sum_{i=1}^M y^i = 0 \tag{71}$$

Since $||y^k|| = 1$, it follows from (66) that $M - \sigma_k = 0 \forall k$, contradicting the initial hypothesis that $M - \sigma_k \neq 0 \ \forall k$.

In summary, the Lagrangian is maximum when

$$\sum_{i=1}^M y^i = 0 \tag{72}$$

$$\sigma_k = M, \ \forall k. \tag{73}$$

Taking the dot product of both sides of (72) with y^k and using the fact that $||y^k|| = 1$,

$$\sum_{j \neq k} < y^j, y^k > = -1. \tag{74}$$

Combining with the fact that $||y^i - y^j||^2 = 2 - 2 < y^i, y^j >$ it follows that

$$\begin{aligned} \sum_i \sum_{j \neq i} ||y^i - y^j||^2 &= 2M(M-1) - 2 \sum_i \sum_{j \neq i} < y^i, y^j > & (75) \\ &= 2M^2 & (76) \end{aligned}$$

Combining this with (61) leads to (60)

Theorem 2. *Any set of unit vectors $y^1, ...y^M \in \mathbb{R}^{M-1}$ which form a regular simplex in $\mathbb{R}^{M-1}$ is a solution of (19)*

Proof. From Lemma 1, if $d^2_{min} = \min_{i \neq j} ||y^i - y^j||^2$ then

$$d^2_{min} \leq \frac{2M}{(M-1)}. \tag{77}$$

Since the pairwise distances between the vertices of a regular unit simplex in $\mathbb{R}^{M-1}$ are all equal to $\sqrt{\frac{2M}{(M-1)}}$ [3], the set of these vertices achieves the upper bound of (77). Hence, this set is a solution to (19). Note that this solution is not unique, since any rotation of the simplex is an equally valid solution.

F Derivation of CD-MCBoost

From (13) and (22)

$$\begin{aligned} -\delta R[f^t; j, g] &= -\frac{\partial}{\partial \epsilon} \sum_{i=1}^n L_M[y_i, f^t(x_i) + \epsilon g(x_i) \mathbf{1}_j] \Big|_{\epsilon=0} & (78) \\ &= -\sum_{i=1}^n \frac{\partial L_M[y_i, f^t(x_i) + \epsilon g(x_i) \mathbf{1}_j]}{\partial \epsilon} \Big|_{\epsilon=0} & (79) \\ &= -\sum_{i=1}^n \frac{\partial}{\partial \epsilon} \sum_{k=1}^M e^{-\frac{1}{2} < f^t(x_i) + \epsilon g(x_i) \mathbf{1}_j, y_i - y^k >} \Big|_{\epsilon=0} & (80) \\ &= -\sum_{i=1}^n \sum_{k=1}^M \left[\frac{\partial}{\partial \epsilon} e^{-\frac{1}{2} \epsilon g(x_i) < \mathbf{1}_j, y_i - y^k >} \right] e^{-\frac{1}{2} < f^t(x_i), y_i - y^k >} \Big|_{\epsilon=0} & (81) \\ &= \frac{1}{2} \sum_{i=1}^n \sum_{k=1}^M g(x_i) < \mathbf{1}_j, y_i - y^k > e^{-\frac{1}{2} < f^t(x_i), y_i - y^k >} & (82) \\ &= \frac{1}{2} \sum_{i=1}^n g(x_i) \sum_{k=1}^M < \mathbf{1}_j, y_i - y^k > e^{-\frac{1}{2} < f^t(x_i), y_i - y^k >} & (83) \\ &= \sum_{i=1}^n g(x_i) w_i^j & (84) \end{aligned}$$

where

$$\begin{aligned} w_i^j &= \frac{1}{2} \sum_{k=1}^M < \mathbf{1}_j, y_i - y^k > e^{-\frac{1}{2} < f^t(x_i), y_i - y^k >} & (85) \\ &= \frac{1}{2} e^{-\frac{1}{2} < f^t(x_i), y_i >} \sum_{k=1}^M < \mathbf{1}_j, y_i - y^k > e^{\frac{1}{2} < f^t(x_i), y^k >}. & (86) \end{aligned}$$

G Derivation of GD-MCBoost

Using (13) and (29)

$$
\begin{aligned}
-\delta R[f^t; g] &= -\frac{\partial}{\partial \epsilon} \sum_{i=1}^{n} L_M[y_i, f^t(x_i) + \epsilon g(x_i)] \Big|_{\epsilon=0} && (87)\\
&= -\sum_{i=1}^{n} \frac{\partial L_M[y_i, f^t(x_i) + \epsilon g(x_i)]}{\partial \epsilon} \Big|_{\epsilon=0} && (88)\\
&= -\sum_{i=1}^{n} \frac{\partial}{\partial \epsilon} \sum_{k=1}^{M} e^{-\frac{1}{2} < f^t(x_i) + \epsilon g(x_i), y_i - y^k >} \Big|_{\epsilon=0} && (89)\\
&= -\sum_{i=1}^{n} \sum_{k=1}^{M} [\frac{\partial}{\partial \epsilon} e^{-\frac{1}{2}\epsilon < g(x_i), y_i - y^k >}] e^{-\frac{1}{2} < f^t(x_i), y_i - y^k >} \Big|_{\epsilon=0} && (90)\\
&= \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{M} < g(x_i), y_i - y^k > e^{-\frac{1}{2} < f^t(x_i), y_i - y^k >} && (91)\\
&= \frac{1}{2} \sum_{i=1}^{n} < g(x_i), \sum_{k=1}^{M} (y_i - y^k) e^{-\frac{1}{2} < f^t(x_i), y_i - y^k >} > && (92)\\
&= \sum_{i=1}^{n} < g(x_i), w_i > && (93)
\end{aligned}
$$

where

$$
\begin{aligned}
w_i &= \frac{1}{2} \sum_{k=1}^{M} (y_i - y^k) e^{-\frac{1}{2} < f^t(x_i), y_i - y^k >} && (94)\\
&= \frac{1}{2} e^{-\frac{1}{2} < f^t(x_i), y_i >} \sum_{k=1}^{M} (y_i - y^k) e^{\frac{1}{2} < f^t(x_i), y^k >} && (95)
\end{aligned}
$$