A Proof of Lemma 2

Proof: Since the support of LL distributions is $\mathbb{R}^d$, two such distributions are equivalent (absolutely continuous with respect to each other) and the divergence is well-defined.

We start by calculating the following integral, assuming $\mu_1 \leq \mu_2$:

\[
I = \int_{\mathbb{R}} \frac{\omega - \mu_2}{\sigma_2} \cdot \exp \left\{ \frac{\omega - \mu_1}{\sigma_1} \right\} d\omega \\
= \frac{\sigma_1}{\sigma_2} \left[ \int_{-\infty}^{\mu_1} \frac{\omega - \mu_2}{\sigma_2} \cdot \exp \left\{ - \frac{\omega - \mu_1}{\sigma_1} \right\} d\omega \\
+ \int_{\mu_1}^{\mu_2} \frac{\omega - \mu_2}{\sigma_2} \cdot \exp \left\{ \frac{\omega - \mu_1}{\sigma_1} \right\} d\omega \\
+ \int_{\mu_2}^{\infty} \frac{\omega - \mu_2}{\sigma_2} \cdot \exp \left\{ \frac{\omega - \mu_1}{\sigma_1} \right\} d\omega \right] .
\]

Changing variables $y = \frac{\omega - \mu_1}{\sigma_1}$ yields,

\[
I = \frac{\sigma^2_1}{\sigma_2} \left[ \int_{-\infty}^{\mu_2 - \mu_1} (-y + \frac{\mu_2 - \mu_1}{\sigma_1}) \cdot \exp \left\{ y \right\} dy \\
- \int_{0}^{\frac{\mu_2 - \mu_1}{\sigma_1}} (-y + \frac{\mu_2 - \mu_1}{\sigma_1}) \cdot \exp \left\{ -y \right\} dy \\
- \int_{\frac{\mu_2 - \mu_1}{\sigma_1}}^{\infty} (-y + \frac{\mu_2 - \mu_1}{\sigma_1}) \cdot \exp \left\{ -y \right\} dy \right] \\
= \frac{2\sigma_1^2}{\sigma_2} \left[ \frac{\mu_2 - \mu_1}{\sigma_1} \cdot \frac{\sigma_1}{\sigma_2} + \exp \left\{ - \frac{\mu_2 - \mu_1}{\sigma} \right\} \right] .
\]

We thus conclude for the general case,

\[
I = \frac{2\sigma_1^2}{\sigma_2} \left[ \frac{\mu_2 - \mu_1}{\sigma_1} + \exp \left\{ - \frac{\mu_2 - \mu_1}{\sigma} \right\} \right] . \tag{15}
\]

As for the Kulback-Leibler Divergence, we use the chain formula for independent random variables,

\[
KL(Q\|P) = \sum_{k=1}^{d} D_{KL}(Q_k\|P_k) = \sum_{k=1}^{d} \int_{\mathbb{R}} \log \left( \frac{Q_k}{P_k} \right) dQ_k \\
= \sum_{k=1}^{d} \int_{\mathbb{R}} \log \left( \frac{\sigma_{P,k}}{\sigma_{Q,k}} \right) + \int_{\mathbb{R}} (2\sigma_{Q,k})^{-1} \times e^{-\frac{\omega_k - \mu_{Q,k}}{\sigma_{Q,k}}} \left[ \frac{\omega_k - \mu_{P,k}}{\sigma_{P,k}} - \frac{\omega_k - \mu_{Q,k}}{\sigma_{Q,k}} \right] d\omega_k
\]

The first term of the integral is given in (15), and the second term is exactly the 1-dimensional $\sigma$-weighted $\ell_1$-norm, therefore, $(2\sigma_{Q,k})^{-1} EQ \left[ \frac{\omega_k - \mu_{Q,k}}{\sigma_{Q,k}} \right] = 1$, which completes the proof.

B Proof of Lemma 3

Proof: We prove that,$ \Pr_{\omega \sim Q} (y(\omega \cdot x) < 0) = \Pr_{\omega \sim Q} [y(\omega \cdot \mu) \cdot x) < -y(\mu \cdot x)] \\
\quad = \mathcal{E}(x, y, \mu_Q, \sigma_Q) .
\]

The random variable

\[
Z = y(\omega - \mu) \cdot x,
\]

is a sum of $d$ independent zero-mean laplace distributed random variables,

\[
Z_k \sim \text{Laplace}(0, \sigma_Q |x_k|) ,
\]

each is equal in distribution to a difference between two i.i.d. exponential random variables. Therefore,

\[
\Pr_{\omega \sim Q} (y(\omega \cdot x) < 0) = \Pr \left( \sum_{k=1}^{d} A_k - \sum_{k=1}^{d} B_k < -y(\mu \cdot x) \right) ,
\]

where $A_k, B_k \sim \text{Exp}(\lambda_k)$ and,

\[
\lambda_k = \lambda_k(x) = (\sigma_Q |x_k|)^{-1} \quad k = 1, \ldots, d.
\]

Without the loss of generality we assume that the coordinates of $x$ are sorted, i.e $\lambda_1 < \lambda_2 \cdots < \lambda_d$. Calculating the convolution for $x_j \neq x_k$ and $z \geq 0$,

\[
f_{A_j + \lambda_k}(z) = \int_{0}^{z} \lambda_j \lambda_k e^{-\lambda_j(t)z} e^{-\lambda_k(t)} dt
\]

\[
= \frac{\lambda_j \lambda_k}{\lambda_j - \lambda_k} [e^{-\lambda_k z} - e^{-\lambda_j z}] .
\]

Exploiting the structure of the resulting convolution, we convolve it with the $\text{th}$ density and get,

\[
f_{A_j + A_k + A_l}(z) = \lambda_j \lambda_k \lambda_l \times \left[ (\lambda_m - \lambda_j) e^{-\lambda_j z} - (\lambda_m - \lambda_k) e^{-\lambda_k z} + (\lambda_j - \lambda_k) e^{-\lambda_m z} \right] .
\]

Performing convolution for all $d$ densities yields,

\[
f_{\sum_{k=1}^{d} A_k}(z) = \sum_{k=1}^{d} \xi_k e^{-\lambda_k z} \quad z \geq 0 ,
\]

where we define $\xi_k = \xi_k(x) = \frac{(-1)^{k-1} \prod_{j=1}^{d} \lambda_j}{\prod_{n=1, n \neq k}^{d} |\lambda_n - \lambda_k|}$.

Similarly, we get the same result for $f_{-\sum_{k=1}^{d} B_k}(z)$, yet it is defined for $z \leq 0$. From (16) we convolute the

\[
\text{Notice that if } x_k = 0 \text{ the random variable } \omega_k x_k \text{ equals zero too, therefore we assume without loss of generality that } x_k \neq 0 .
\]
difference and get,
\[
\begin{align*}
    f_{\sum_{k=1}^d A_k - B_k}(z) &= \left( f_{\sum_{k=1}^d A_k} * f_{-\sum_{k=1}^d B_k} \right)(z) \\
    &= \int_{-\infty}^{\min(z,0)} \left( \sum_{m=1}^d \xi_m e^{\lambda_m t} \right) \left( \sum_{k=1}^d \xi_k e^{\lambda_k(z-t)} \right) dt \\
    &= \sum_{m,n=1}^d \xi_m \xi_k e^{-\lambda_k z} e^{(\lambda_m + \lambda_k) t} \bigg|_{-\infty}^{\min(z,0)} \\
    &= \sum_{m,n=1}^d \xi_m \xi_k e^{-\lambda_k |z|} e^{(\lambda_m + \lambda_k) |z|} \\
    \text{for } \psi_k &= \psi_k(x) = \sum_{m=1}^d \xi_m \xi_k \cdot \frac{x}{\lambda_m + \lambda_k}.
\end{align*}
\]

We integrate to get the CDF,
\[
\ell_{\text{off}}(y(\omega \cdot x)) = \int_{z=-\infty}^{\infty} \sum_{k=1}^d \psi_k e^{-\lambda_k |z|} \, dz
\]
\[
= \left\{ \begin{array}{ll}
\sum_{k=1}^d \frac{\psi_k}{\lambda_k} e^{-\lambda_k y(\mu \cdot x)} & y(\mu \cdot x) \geq 0 \\
1 - \sum_{k=1}^d \frac{\psi_k}{\lambda_k} e^{-\lambda_k y(\mu \cdot x)} & y(\mu \cdot x) < 0
\end{array} \right.
\]

Finally, we define \(\alpha_k(x) = \frac{\psi_k(x)}{\lambda_k} \) and obtain for \(\xi = \text{sort}(|x|)(3)\),
\[
\alpha_k(x) = \xi_k \left( \prod_{j=1}^d \xi_j \right)^{-2} \prod_{j=1,j\neq k}^d \left| \xi_j^{-1} - \xi_k^{-1} \right|^{-1} \\
\times \sum_{m=1}^d (-1)^{m+k} \left( \xi_k^{-1} + \xi_m^{-1} \right)^{-1} \prod_{j=1,j\neq m}^d \left| \xi_j^{-1} - \xi_m^{-1} \right|^{-1}.
\]

In particular, from the symmetry of \(f_{\sum_{k=1}^d A_k - B_k}(z)\), we have for \(\mu = 0\), that
\[
\frac{1}{2} = \Pr_{\omega \sim Q}(y(\omega \cdot x) < 0) = \sum_{k=1}^d \alpha_k(x)
\]
which concludes the proof.

## C Proof of Theorem 4

**Proof:** From the assumption that the data is linearly separable we conclude that the set \(\{\mu_Q \mid y_i x_i \cdot \mu_Q \geq 0, i = 1, \ldots, m\}\) is not empty. Additionally, the set is defined via linear constraints and thus convex. The objective (7) is convex in \(\sigma\) as its second derivative with respect to \(\sigma\) is \(d\sigma^2 > 0\).

The regularization term of (7) is convex in \(\mu\) as the second derivative of \(|z| + \exp(-|z|)\) is always positive and well defined for all values of \(z\) (see also Remark 1 for a discussion of this function for values \(z \approx 0\)).

## D Proof of Lemma 10

**Proof:** Assume \(f, g \in S\) and denote by \(h = f * g\). The derivative of a convolution between two differentiable functions always exists, and equals to, \(\frac{d}{dz} (f * g) =

Figure 6: Illustration of the cumulative sums, \(\sum_{i=1}^{k} \alpha_i(x)\), for five 10-dimensional vectors.

As for the loss term \(\ell(y_i x_i \cdot \mu)\), we use the following auxiliary lemma.

**Lemma 10** The following set of probability density functions over the reals
\[
\mathcal{S} = \left\{ f_{\text{pdf}} \mid f \in \mathcal{C}, f(z) = f(-z), \quad \text{and } \forall z_1, z_2, |z_2| > |z_1| \Rightarrow f(z_2) < f(z_1) \right\}
\]
is closed under convolution, i.e \(f, g \in \mathcal{S} \Rightarrow f * g \in \mathcal{S}\).

Since the random variables \(\omega_1, \ldots, \omega_d\) are independent, the density \(f_{Z_i}(z)\) of the margin \(Z_i = y_i (\omega - \mu_Q) \cdot x_i\), is obtained by convoluting \(d\) independent zero-mean Laplace distributed random variables \(y_i (\omega_k - \mu_Q) \cdot x_{i,k}\). Since the 1-dimensional Laplace pdf is in \(\mathcal{S}\), it follows from Lemma 10 by induction that so is \(f_{Z_i}\). As a member of \(\mathcal{S}\), the positivity of the derivative \(f'_{Z_i}(z)\) for \(z \leq 0\) is concluded from Lemma 10. Finally, we note that the integral of the density is \(\ell_{\text{off}}\), the cumulative density function, \(\mathcal{E}(x_i, y_i, \mu_Q, \sigma_Q) = \int_{-\infty}^{y_i \mu_Q \cdot x_i} f_{Z_i}(z) \, dz\).

Thus, the second derivative of \(\mathcal{E}(x_i, y_i, \mu_Q, \sigma_Q)\) for positive values of the margin, equals to \(f''_{Z_i}(z)\) for \(z \leq 0\), and hence positive. Changing variables according to (6) completes the proof.
We compute for the convolution derivative,

\[ h'(z) = \int_{-\infty}^{\infty} f(z-t) \cdot \left( \frac{dg(t)}{dt} \right) dt \]

\[ = \int_{-\infty}^{0} f(z-t) \cdot \left( \frac{dg(t)}{dt} \right) dt + \int_{0}^{\infty} f(z-t) \cdot \left( \frac{dg(t)}{dt} \right) dt \]

\[ = \int_{-\infty}^{0} f(z-t) \cdot \left( \frac{dg(t)}{dt} \right) dt + \int_{0}^{\infty} f(z+t) \cdot \left( \frac{dg(t)}{dt} \right) dt \]

\[ = \int_{0}^{\infty} \left[ f(z-t) - f(z+t) \right] \left( \frac{dg(t)}{dt} \right) dt , \]

where the last equality follows the fact \( \frac{dg(t)}{dt} \) is an odd function as a derivative of an even function. Since \( f, g \in \mathbb{S}, h(z) \in C_1 \) (i.e continuously differentiable almost everywhere), and since \( h'(z) \) is odd, we have that \( h(z) \) is even. Using the monotonicity property of \( f, g \), i.e \( |z_2| > |z_1| \Rightarrow f(z_2) < f(z_1) \), we get,

\[ \int_{-\infty}^{0} \left[ f(z-t) - f(z+t) \right] \left( \frac{dg(t)}{dt} \right) dt \]

\[ = - \text{sign}(z) \int_{-\infty}^{0} \left| f(z-t) - f(z+t) \right| \left| \frac{dg(t)}{dt} \right| dt . \]

Since \( f, g \) are pdfs, the integral is always defined, and thus the sign of the derivative of \( h \) depends on the sign of its argument, and in particular it is an increasing function for \( z < 0 \) and decreasing for \( z > 0 \), yielding the third property for \( h \). Thus, \( h \in \mathbb{S} \), as desired.

### E Proof of Lemma 5

**Proof:** Setting \( \mu = 0 \) and \( \sigma = 1 \) the objective becomes \( 0 + cm\eta \). Since the loss is non-negative we get that the minimizers satisfy,

\[ cm\eta \geq \]

\[ -d \log \sigma^* e + \sigma^* \sum_{k=1}^{d} |\mu_k^*| + e^{-|\mu_k^*|} \]

\[ + e \sum_{i} \ell(y_i x_i \cdot \mu^*) \geq \]

\[ -d \log \sigma^* e + \sigma^* \sum_{k=1}^{d} |\mu_k^*| + e^{-|\mu_k^*|} . \]

Substituting the optimal value of \( \sigma^* \) from (8) we get,

\[ cm\eta \geq -d \log \frac{ed}{\sum_{k=1}^{d} |\mu_k^*| + e^{-|\mu_k^*|}} \]

\[ = d \log \frac{\sum_{k=1}^{d} |\mu_k^*| + e^{-|\mu_k^*|}}{d} . \]

Rearranging, we get,

\[ d\exp \left( \frac{cm\eta}{d} \right) \geq \sum_{k=1}^{d} |\mu_k^*| + e^{-|\mu_k^*|} \geq \|\mu^*\|_1 , \]

and we can conclude,

\[ \sigma^* \geq \exp \left( - \frac{cm\eta}{d} \right) . \]

### F Proof of Theorem 6

**Proof:** While the empirical loss term depends only on \( \mu \), and was proved to be strictly convex for example that satisfies \( y_i x_i \cdot \mu \geq 0 \) in theorem 4, the regularization term is optimized over both \( \mu, \sigma \). Incorporating the optimal value for sigma from (8) into the objective yields the following:

\[ F(\mu, \sigma^*(\mu)) = d \log \left( \sum_{k=1}^{d} |\mu_k^*| + e^{-|\mu_k^*|} \right) \]

\[ + e \sum_{i} \ell(y_i x_i \cdot \mu) . \]

Differentiating the regularization term twice with respect to \( \mu \) results in the following Hessian matrix,

\[ H(\mu) = \frac{d}{\sum_{k=1}^{d} |\mu_k^*| + e^{-|\mu_k^*|}} \]

\[ \left\{ \text{diag}(\exp [-\mu]) - \frac{v \cdot v^T}{\sum_{k=1}^{d} |\mu_k^*| + e^{-|\mu_k^*|}} \right\} , \]

for the \( d \)-dimensional vector \( v_k = \text{sign}(\mu_k) \left( 1 - \exp[-|\mu_k|] \right) \), and \( \text{diag}(\exp [-\mu]) \) is a diagonal vector for which its \( i \)-th elements equals \( \exp(-\mu_i) \). The Hessian \( H(\mu) \) is a difference of two positive semi-definite matrices. We upper bound the maximal eigenvalues of the second term by its trace, indeed,

\[ \max_{j} \lambda_j \left( \frac{d}{\sum_{k=1}^{d} |\mu_k^*| + e^{-|\mu_k^*|}} \right)^2 \]

\[ \leq \frac{dv^T v}{\left( \sum_{k=1}^{d} |\mu_k^*| + e^{-|\mu_k^*|} \right)^2} \]

\[ = \frac{d \sum_{k=1}^{d} (1 - e^{-|\mu_k|})^2}{\left( \sum_{k=1}^{d} |\mu_k| + e^{-|\mu_k|} \right)^2} \]

\[ < \frac{d \times d}{d^2} = 1 . \]
Thus, the minimal eigenvalue of $H(\mu)$ is bounded from below by $-1$, and the Hessian of the sum of the objective and $\frac{1}{2}\|\mu\|^2$ has positive eigenvalues, therefore strictly convex.

For the second part, we use [17, Corollary 7.2.3] stating the a diagonally-dominated matrix with non-negative diagonal values is PSD. We next show that indeed $\|\mu\|_\infty \leq 1$ is a sufficient condition for the Hessian to be diagonally dominated. It is straightforward to verify that both conditions follows from the following set of inequalities, for all $k = 1, \ldots, d$,

$$e^{-|\mu_k|} \sum_{j=1}^d (|\mu_j| + e^{-|\mu_j|}) - (1 - e^{-|\mu_k|}) \sum_{j=1}^d (1 - e^{-|\mu_j|}) > 0$$

or equivalently,

$$e^{-|\mu_k|} + e^{-|\mu_j|} \frac{1}{d} \sum_{j=1}^d |\mu_j| + \frac{1}{d} \sum_{j=1}^d e^{-|\mu_j|} - 1 > 0$$

$$\Leftrightarrow e^{-|\mu_k|} \left( \frac{d+1}{d} + \frac{1}{d} |\mu_k| \right) + e^{-|\mu_k|} \left( \frac{1}{d} \sum_{j=1,j\neq k}^d |\mu_j| \right) + \frac{1}{d} \sum_{j=1,j\neq k}^d e^{-|\mu_j|} - 1 > 0.$$  \tag{17}

Fixing $\mu_k$ the left-hand-side is decomposed to a sum of one variable convex functions $\mu_j$. We minimize it for each $\mu_j$ by taking the derivative and setting it to zero, yielding,

$$\frac{1}{d} \left( \text{sign}(\mu_j) \left[ e^{-|\mu_k|} - e^{-|\mu_j|} \right] \right) = 0 \Rightarrow \mu_j = \mu_k. \tag{18}$$

From here we conclude that (17) is satisfied if $\|\mu\|_\infty \leq a$ for a scalar $a > 0$ that satisfy,

$$g(a) = 2e^{-a} + ae^{-a} - 1 > 0.$$  

The function $g(a)$ is monotonically decreasing and continuous, with $g(1) = 3/e - 1 > 0$, which completes the proof. In fact, one can compute numerically and find that $a^* \approx 1.146$ satisfy $g(a^*) \approx 0$, which leads to a slightly better constant than stated in the theorem.

\section{Proof of Lemma 8}

\textbf{Proof:} Denote the change of the loss term of (12) by,

$$\Delta_t = \sum_{i=1}^m \log \left(1 + D_i e^{-y_i x_i \cdot \mu^{(t)}}\right) - \sum_{i=1}^m \log \left(1 + D_i e^{-y_i x_i \cdot \left[\mu^{(t)} + \delta^{(t)}\right]}\right).$$

We start by bounding $\Delta_t$ from below, then add to it the difference of the regularization term, before and after the update. Bounding the improvement for a
single example, we get,
\[ \frac{\Delta_{t,i}}{c} = - \log \left( \frac{1 + D_t e^{-y_i x_i \mu_Q^{(t+1)}}}{1 + D_t e^{-y_i x_i \mu_Q^{(t)}}} \right) \]
\[ = - \log \left( 1 - \frac{D_t e^{-y_i x_i \mu_Q^{(t)}}}{1 + D_t e^{-y_i x_i \mu_Q^{(t+1)}}} \right) \]
\[ = - \log \left( 1 - \frac{D_t}{1 + e^{y_i x_i \mu_Q^{(t+1)}}} \right) \]
\[ = - \log \left( 1 - q_t(i) \left[ 1 - e^{-y_i x_i, k \delta_k^{(t)}} \right] \right). \]

By using \( -\log(1 - z) \geq z \) for \( z < 1 \) we get,
\[ - \log \left( 1 - q_t(i) \left[ 1 - e^{-y_i x_i, k \delta_k^{(t)}} \right] \right) \geq q_t(i) \left[ 1 - e^{-y_i x_i, k \delta_k^{(t)}} \right]. \]

Convexity of the exponent, for every \( \sigma_{Q,k} \in (0, 1) \), yields,
\[ e^{-y_i x_i, k \delta_k^{(t)}} \leq \sigma_{Q,k} \left| x_{i,k} \right| e^{-\text{sign}(y_i x_i) \frac{\delta_k^{(t)}}{\sigma_{Q,k}}} + (1 - \sigma_{Q,k} \left| x_{i,k} \right|) e^{0}. \]

Summing over the examples,
\[ \Delta_t \geq c \sum_{i=1}^{m} q_t(i) \sigma_{Q,k} \left| x_{i,k} \right| \left( 1 - e^{-\text{sign}(y_i x_i) \frac{\delta_k^{(t)}}{\sigma_{Q,k}}} \right) \]
\[ = e \sum_{i=1, y_i x_i \geq 0} q_t(i) \sigma_{Q,k} \left| x_{i,k} \right| \left( 1 - e^{-\frac{\delta_k^{(t)}}{\sigma_{Q,k}}} \right) \]
\[ + e \sum_{i=1, y_i x_i < 0} q_t(i) \sigma_{Q,k} \left| x_{i,k} \right| \left( 1 - e^{-\frac{\delta_k^{(t)}}{\sigma_{Q,k}}} \right) \]
\[ = e \sigma_{Q,k} \left( \gamma_k^+ \left[ 1 - e^{-\frac{\delta_k^{(t)}}{\sigma_{Q,k}}} \right] + \gamma_k^- \left[ 1 - e^{-\frac{\delta_k^{(t)}}{\sigma_{Q,k}}} \right] \right), \]

adding the regularization terms completes the proof.

\[ \square \]

**K Experiments- Data Details:**

**Synthetic data:** We generated 4,000 vectors \( x_i \in \mathbb{R}^8 \) sampled from a zero mean isotropic normal distribution \( x_i \sim \mathcal{N}(0, 1) \). Labels were assigned by generating once per run \( \omega \in \mathbb{R}^8 \) at random and using:
\[ y_i = \text{sign}(\omega \cdot x_i). \]
Each input \( x_i \) training data was then corrupted with probability \( p \) by adding to it a random vector sampled from a zero mean isotropic Gaussian, \( \epsilon_i \sim \mathcal{N}(0, \sigma \mathbf{I}) \), with some positive standard-deviation \( \sigma \). Each run was repeated 20 times, and results are average test-error over the 20 runs. All boosting algorithms were run for 1,000 iterations, except for the RobuCoP algorithm which was executed until a convergence criterion was met, which often was about 20 rounds.

**Vocal Joystick:** For each problem, we picked three sets of size 2,000 each, for training, parameter tuning and testing. Each example is a frame of spoken value described with 13 MFCC coefficients transformed into 27 features. In order to examine the robustness of different algorithms, we contaminate 10% of the data with an additive zero-mean i.i.d Gaussian noise, for different values of the standard-deviation \( \sigma \).