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 $\mu_1 \leq \mu_2$:

$$\begin{aligned} I = & \int_{\mathbb{R}} \frac{|\omega-\mu_2|}{\sigma_2}\cdot \exp\left\{\frac{|\omega-\mu_1|}{\sigma_1}\right\}d\omega \\ = & \frac{\sigma_1}{\sigma_2}\Bigg[\int_{-\infty}^{\mu_1} -\frac{\omega-\mu_2}{\sigma_1}\cdot\exp\left\{-\frac{\omega-\mu_1}{\sigma_1}\right\}d\omega \\ & + \int_{\mu_1}^{\mu_2} -\frac{\omega-\mu_2}{\sigma_1}\cdot\exp\left\{\frac{\omega-\mu_1}{\sigma_1}\right\}d\omega \\ & + \int_{\mu_2}^{\infty} \frac{\omega-\mu_2}{\sigma_1}\cdot\exp\left\{\frac{\omega-\mu_1}{\sigma_1}\right\}d\omega\Bigg] \ . \end{aligned}$$

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

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

We thus conclude for the general case,

$$I = \frac{2\sigma_1^2}{\sigma_2}\left[\frac{|\mu_2-\mu_1|}{\sigma_1} + \exp\left\{-\frac{|\mu_2-\mu_1|}{\sigma_1}\right\}\right] \ . \quad (15)$$

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

$$\begin{aligned} KL(Q\|P) = & \sum_{k=1}^{d} \mathrm{D_{KL}}(Q_k\|P_k) = \sum_{k=1}^{d}\int_{\mathbb{R}}\log\left(\frac{Q_i}{P_i}\right)dQ_i \\ = & \sum_{k=1}^{d}\Bigg[\log\left(\frac{\sigma_{P,k}}{\sigma_{Q,k}}\right) + \int_{\mathbb{R}}(2\sigma_{Q,k})^{-1}\times \\ & e^{-\frac{|\omega_k-\mu_{Q,k}|}{\sigma_{Q,k}}}\left[\frac{|\omega_k-\mu_{P,k}|}{\sigma_{P,k}} - \frac{|\omega_k-\mu_{Q,k}|}{\sigma_{Q,k}}\right]d\omega_k\Bigg] \ . \end{aligned}$$

The first term of the integral is given in (15), and the second term is exactly the 1-dimensional $\boldsymbol{\sigma}$-weighted ℓ_1-norm, therefore, $(2\sigma_{Q,k})^{-1}\,\mathrm{E}_Q\left[\frac{|\omega_k-\mu_{Q,k}|}{\sigma_{Q,k}}\right] = 1$, which completes the proof. ■

B Proof of Lemma 3

Proof: We prove that,

$$\begin{aligned} \Pr_{\omega\sim Q}(y(\boldsymbol{\omega}\cdot\boldsymbol{x})<0) = & \Pr_{\omega\sim Q}[y(\boldsymbol{\omega}-\boldsymbol{\mu})\cdot\boldsymbol{x}) < -y(\boldsymbol{\mu}\cdot\boldsymbol{x})] \\ = & \mathcal{E}\left(\boldsymbol{x}, y, \boldsymbol{\mu}_Q, \sigma_Q\right) \ . \end{aligned}$$

The random variable[4]

$$Z = y(\boldsymbol{\omega}-\boldsymbol{\mu})\cdot\boldsymbol{x} \ ,$$

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

$$Z_k \sim \mathrm{Laplace}(0, \sigma_Q\,|x_k|) \ ,$$

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

$$\Pr_{\omega\sim Q}(y(\boldsymbol{\omega}\cdot\boldsymbol{x})<0) = \Pr\left(\sum_{k=1}^{d}A_k - \sum_{k=1}^{d}B_k < -y(\boldsymbol{\mu}\cdot\boldsymbol{x})\right), \quad (16)$$

where $A_k, B_k \sim \mathrm{Exp}(\lambda_k)$ and,
$\lambda_k = \lambda_k(\boldsymbol{x}) = (\sigma_Q\,|x_k|)^{-1} \quad k = 1,\ldots,d.$

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

$$\begin{aligned} f_{A_j+A_k}(z) = & \int_0^z \lambda_j\lambda_j e^{-\lambda_j(-t)z}e^{-\lambda_k(t)}dt \\ = & \frac{\lambda_j\lambda_k}{\lambda_j-\lambda_k}\left[e^{-\lambda_k z} - e^{-\lambda_j z}\right] \ . \end{aligned}$$

Exploiting the structure of the resulting convolution, we convolve it with the lth density and get,

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

Performing convolution for all d densities yields,

$$f_{\sum_{k=1}^d A_k}(z) = \sum_{k=1}^{d}\xi_k\,e^{-\lambda_k z} \ \text{for}\ z\geq 0 \ ,$$

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

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

[4]Notice that if $x_k = 0$ the random variable $\omega_k x_k$ equals zero too, therefore we assume without loss of generality that $x_k \neq 0$.

difference and get,

$$
\begin{aligned}
& f_{\sum_{k=1}^d A_k - B_k}(z) = \left(f_{\sum_{k=1}^d A_k} * f_{-\sum_{k=1}^d B_k}\right)(z) \\
&= \int_{-\infty}^{\min(z,0)} \left(\sum_{m=1}^d \xi_m e^{\lambda_m t}\right)\left(\sum_{k=1}^d \xi_k e^{\lambda_k (z-t)}\right) dt \\
&= \sum_{m,n=1}^d \xi_m \xi_k e^{-\lambda_k z} \frac{e^{(\lambda_m+\lambda_k)t}}{\lambda_m+\lambda_k}\Bigg|_{-\infty}^{\min(z,0)} \\
&= \sum_{m,n=1}^d \frac{\xi_m\xi_k}{\lambda_m+\lambda_k} e^{-\lambda_k|z|} = \sum_{k=1}^d \psi_k e^{-\lambda_k|z|} \\
& \quad \text{for } \psi_k = \psi_k(\boldsymbol{x}) = \sum_{m=1}^d \frac{\xi_m\xi_k}{\lambda_m+\lambda_k} \ .
\end{aligned}
$$

We integrate to get the CDF,

$$
\begin{aligned}
& \ell_{cdf}\left(y(\boldsymbol{\omega}\cdot\boldsymbol{x})\right) = \int_{z=-\infty}^{-y(\boldsymbol{\mu}\cdot\boldsymbol{x})} \sum_{k=1}^d \psi_k e^{-\lambda_k|z|} \\
&= \begin{cases} \sum_{k=1}^d \frac{\psi_k}{\lambda_k} e^{-\lambda_k y(\boldsymbol{\mu}\cdot\boldsymbol{x})} & y(\boldsymbol{\mu}\cdot\boldsymbol{x}) \geq 0 \\ 1 - \sum_{k=1}^d \frac{\psi_k}{\lambda_k} e^{\lambda_k y(\boldsymbol{\mu}\cdot\boldsymbol{x})} & y(\boldsymbol{\mu}\cdot\boldsymbol{x}) < 0 \end{cases}
\end{aligned}
$$

Finally, we define $\alpha_k(\boldsymbol{x}) = \frac{\psi_k(\boldsymbol{x})}{\lambda_k(\boldsymbol{x})}$ and obtain for $\xi = \text{sort}(|x|)$ (3),

$$
\begin{aligned}
\alpha_k(\boldsymbol{x}) &= \xi_k \left(\prod_{j=1}^d \xi_j\right)^{-2} \prod_{j=1, j\neq k}^d \left|\xi_j^{-1} - \xi_k^{-1}\right|^{-1} \\
&\times \sum_{m=1}^d (-1)^{m+k} \left(\xi_k^{-1} + \xi_m^{-1}\right)^{-1} \prod_{j=1, j\neq m}^d \left|\xi_j^{-1} - \xi_m^{-1}\right|^{-1} .
\end{aligned}
$$

In particular, from the symmetry of $f_{\sum_{k=1}^d A_k - B_k}(z)$, we have for $\boldsymbol{\mu} = 0$, that

$$
\frac{1}{2} = \Pr_{\omega\sim\mathcal{Q}}\left(y(\boldsymbol{\omega}\cdot\boldsymbol{x}) < 0\right) = \sum_{k=1}^d \alpha_i \ ,
$$

which concludes the proof. ■

C Proof of Theorem 4

Proof: From the assumption that the data is linearly separable we conclude that the set $\left\{\boldsymbol{\mu}_Q \,|\, y_i \boldsymbol{x}_i \cdot \boldsymbol{\mu}_Q \geq 0, i = 1, \ldots, m\right\}$ is not empty. Additionally, the set is defined via linear constraints and thus convex. The objective (7) is convex in σ as its second derivative with respect to σ is $d\sigma^{-2} > 0$.

The regularization term of (7) is convex in $\boldsymbol{\mu}$ 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 $z \approx 0$).

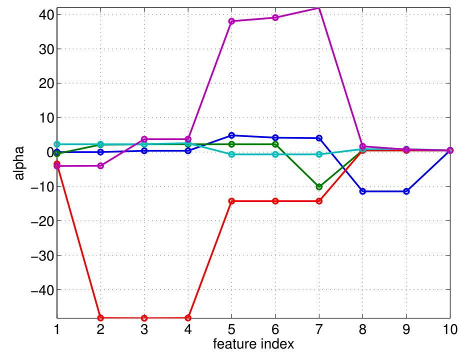

Figure 6: Illustration of the cumulative sums, $\sum_{i=1}^k \alpha_i(\boldsymbol{x})$, for five 10-dimensional vectors.

As for the loss term $\ell\left(y_i \boldsymbol{x}_i \cdot \boldsymbol{\mu}\right)$, we use the following auxiliary lemma.

Lemma 10 *The following set of probability density functions over the reals*

$$
\begin{aligned}
\mathbb{S} = \Big\{ f_{pdf} \ \Big| \ & f \in \mathcal{C}_1 \ , \ f(z) = f(-z), \\
& \textit{and } \forall z_1, z_2, |z_2| > |z_1| \Rightarrow f(z_2) < f(z_1) \Big\}
\end{aligned}
$$

is closed under convolution, i.e $f, g \in \mathbb{S} \rightarrow f * g \in \mathbb{S}$.

Since the random variables $\omega_1, .., \omega_d$ are independent, the density $f_{Z_i}(z)$ of the margin $Z_i = y_i\left(\boldsymbol{\omega} - \boldsymbol{\mu}_Q\right) \cdot \boldsymbol{x}_i$, is obtained by convoluting d independent zero-mean Laplace distributed random variables $y_i(\omega_k - \mu_{i,k})x_{i,k}$. Since the 1-dimensional Laplace pdf is in $\mathbb{S}$, it follows from Lemma 10 by induction that so is f_{Z_i}. As a member of $\mathbb{S}$, 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 ℓ_{cdf}, the cumulative density function, $\mathcal{E}(\boldsymbol{x}_i, y_i, \boldsymbol{\mu}_Q, \sigma_Q) = \int_{-\infty}^{-y_i \boldsymbol{\mu}_Q \cdot \boldsymbol{x}_i} f_{Z_i}(z) dz$. Thus, the second derivative of $\mathcal{E}(\boldsymbol{x}_i, y_i, \boldsymbol{\mu}_Q, \sigma_Q)$ 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. ■

D Proof of Lemma 10

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

$f * \left(\frac{dg}{dz}\right)$. We compute for the convolution derivative,

$$\begin{aligned} h'(z) &= \int_{-\infty}^{\infty} f(z-t) \cdot \left(\frac{dg(t)}{dt}\right) dt \\ &= \int_{-\infty}^{0} f(z-t) \cdot \left(\frac{dg(t)}{dt}\right) dt + \int_{0}^{\infty} f(z-t) \cdot \left(\frac{dg(t)}{dt}\right) dt \\ &= \int_{-\infty}^{0} f(z-t) \cdot \left(\frac{dg(t)}{dt}\right) dt + \int_{-\infty}^{0} f(z+t) \cdot \left(\frac{dg(-t)}{dt}\right) dt \\ &= \int_{-\infty}^{0} \Big[f(z-t) - f(z+t)\Big] \left(\frac{dg(t)}{dt}\right) dt \ , \end{aligned}$$

where the last equality follows the fact $\frac{dg(t)}{dt}$ is an odd function as a derivative of an even function. Since $f, g \in \mathbb{S}, h(z) \in \mathcal{C}_1$ (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,

$$\begin{aligned} &\int_{-\infty}^{0} \Big[f(z-t) - f(z+t)\Big] \cdot \left(\frac{dg(t)}{dt}\right) dt \\ &= -\operatorname{sign}(z) \cdot \int_{-\infty}^{0} \Big|f(z-t) - f(z+t)\Big| \left|\frac{dg(t)}{dt}\right| dt \ . \end{aligned}$$

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 $\boldsymbol{\mu} = \mathbf{0}$ and $\sigma = 1$ the objective becomes $0 + cm\eta$. Since the loss is non-negative we get that the minimizers satisfy,

$$\begin{aligned} &cm\eta \geq \\ &-d \log \sigma^* e + \sigma^* \sum_{k=1}^{d} \left[|\mu_k^*| + e^{-|\mu_k^*|}\right] \\ &+ c \sum_i \ell(y_i \boldsymbol{x}_i \cdot \boldsymbol{\mu}^*) \geq \\ &-d \log \sigma^* e + \sigma^* \sum_{k=1}^{d} \left[|\mu_k^*| + e^{-|\mu_k^*|}\right] \ . \end{aligned}$$

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

$$\begin{aligned} cm\eta &\geq -d \log \frac{ed}{\sum_{k=1}^{d} |\mu_k^*| + e^{-|\mu_k^*|}} + d \\ &= d \log \frac{\sum_{k=1}^{d} |\mu_k^*| + e^{-|\mu_k^*|}}{d} . \end{aligned}$$

Rearranging, we get,

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

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 $\boldsymbol{\mu}$, and was proved to be strictly convex for examples that satisfies $y_i \boldsymbol{x}_i \cdot \boldsymbol{\mu} \geq 0$ in theorem 4, the regularization term is optimized over both $\boldsymbol{\mu}, \sigma$. Incorporating the optimal value for sigma from (8) into the objective yields the following:

$$\begin{aligned} \mathcal{F}(\boldsymbol{\mu}, \sigma^*(\boldsymbol{\mu})) = &d \log \left(\sum_{k=1}^{d} |\mu_k| + e^{-|\mu_k|}\right) \\ &+ c \sum_{i=1}^{m} \ell(y \boldsymbol{x}_i \cdot \boldsymbol{\mu}). \end{aligned}$$

Differentiating the regularization term twice with respect to $\boldsymbol{\mu}$ results in the following Hessian matrix,

$$\begin{aligned} H(\boldsymbol{\mu}) = &\frac{d}{\sum_{k=1}^{d} |\mu_k| + e^{-|\mu_k|}} \times \\ &\left[\operatorname{diag}(\exp[-\boldsymbol{\mu}]) - \frac{\boldsymbol{v} \cdot \boldsymbol{v}^\top}{\sum_{k=1}^{d} |\mu_k| + e^{-|\mu_k|}}\right] \ , \end{aligned}$$

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

$$\begin{aligned} &\max_j \lambda_j \left(\frac{d}{\left(\sum_{k=1}^{d} |\mu_k| + e^{-|\mu_k|}\right)^2}\right) \\ &\leq \frac{d\boldsymbol{v}^\top \boldsymbol{v}}{\left(\sum_{k=1}^{d} |\mu_k| + e^{-|\mu_k|}\right)^2} \\ &= \frac{d \sum_{k=1}^{d} \left(1 - e^{-|\mu_k|}\right)^2}{\left(\sum_{k=1}^{d} |\mu_k| + e^{-|\mu_k|}\right)^2} \\ &< \frac{d \times d}{d^2} = 1 \ . \end{aligned}$$

Thus, the minimal eigenvalue of $H(\boldsymbol{\mu})$ is bounded from below by (-1), and the Hessian of the sum of the objective and $\frac{1}{2}\|\boldsymbol{\mu}\|^2$ 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 $\|\boldsymbol{\mu}\|_\infty \leq 1$ 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,

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

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

Fixing μ_k the left-hand-side is decomposed to a sum of one variable convex functions μ_j. We minimize it for each μ_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 $\|\boldsymbol{\mu}\|_\infty \leq a$ for a scalar $a \geq 0$ 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. ■

G Proof of Lemma 7

Proof: We first need to compute ℓ_{lin} directly, as $\alpha_k(\boldsymbol{x})$ is not defined on the standard basis, which contains few elements of the same value,

$$\Pr\left[\boldsymbol{e}_k \cdot \boldsymbol{\omega} \leq 0\right] = \Pr\left[\omega_k \leq 0\right] = \Pr\left[(\omega_k - \mu_k) < -\mu_k\right] = \int_{-\infty}^{-\mu_k}(2\sigma)^{-1}e^{-\frac{|\omega_k|}{\sigma}}d\omega_k\ .$$

Thus, if $\mu_k \geq 0$ we get (the convex part) $\Pr\left[\boldsymbol{e}_k \cdot \boldsymbol{\omega} \leq 0\right] = \frac{1}{2}\exp\left(-|\mu_k|\right)$. Otherwise, we bound $\Pr\left[\boldsymbol{e}_k \cdot \boldsymbol{\omega} \leq 0\right]$ with the linear extension and get $\frac{1}{2}(1 + |\mu_k|)$. To conclude, for each element k we get that, $\sum_{y=\pm 1}\ell_{lin}(y\boldsymbol{e}_k \cdot \boldsymbol{\mu}) = \frac{1}{2}\left(\exp\{-|\mu_k|\} + (1 + |\mu_k|)\right)$. Taking the sum over k and multiplying by 2 yields the above regularization term. ■

H RobuCop Pseudo-code

Input: Training set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$,
$c > 0$, Artifical set $\mathcal{A} = \{(\boldsymbol{e}_k, y) : k = 1...d, y \in \mathcal{Y}\}$
Initialization: $\boldsymbol{\mu}^{(1)} = 0$
Loop do until convergence criterion met:

- Set: $\sigma^{(n+1)} = d\left(\sum_{k=1}^{d}|\mu_k| + \exp\{-|\mu_k|\}\right)^{-1}$
- Solve $\mu^{(n+1)} = \arg\min_{\boldsymbol{\mu}}\left\{\sum_{S\cup\mathcal{A}}\tilde{c}_i \cdot \ell_{lin}(y\boldsymbol{x}_i \cdot \boldsymbol{\mu})\right\}$

  for: $\tilde{c}_i = \begin{cases} c & (\boldsymbol{x}_i, y_i) \in S \\ 2\sigma^{(n+1)} & (\boldsymbol{x}_i, y_i) \in \mathcal{A} \end{cases}$

Output: $\boldsymbol{\mu}, \sigma$

I Proof of Lemma 8

Proof: Denote the change of the loss term of (12) by,

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

We start by bounding Δ_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,

$$\begin{aligned}\frac{\Delta_{t,i}}{c} =& -\log\left(\frac{1+D_i e^{-y_i \boldsymbol{x}_i\cdot\boldsymbol{\mu}_Q^{(t+1)}}}{1+D_i e^{-y_i \boldsymbol{x}_i\cdot\boldsymbol{\mu}_Q^{(t)}}}\right)\\ =& -\log\left(\frac{1}{1+D_i e^{-y_i \boldsymbol{x}_i\cdot\boldsymbol{\mu}_Q^{(t)}}} + \frac{D_i e^{-y_i \boldsymbol{x}_i\cdot\boldsymbol{\mu}_Q^{(t+1)}}}{1+D_i e^{-y_i \boldsymbol{x}_i\cdot\boldsymbol{\mu}_Q^{(t)}}}\right)\\ =& -\log\left(1-\frac{D_i}{D_i+e^{y_i \boldsymbol{x}_i\cdot\boldsymbol{\mu}_Q^{(t)}}} + \frac{D_i e^{-y_i \boldsymbol{x}_i\left[\boldsymbol{\mu}_Q^{(t+1)}-\boldsymbol{\mu}_Q^{(t)}\right]}}{D_i+e^{y_i \boldsymbol{x}_i\cdot\boldsymbol{\mu}_Q^{(t)}}}\right)\\ =& -\log\left(1-q_t(i)\left[1-e^{-y_i x_{i,k}\delta_k^{(t)}}\right]\right)\ .\end{aligned}$$

By using $-\log(1-z)\geq z$ for $z<1$ we get,

$$\begin{aligned}&-\log\left(1-q_t(i)\left[1-e^{-y_i x_{i,k}\delta_k^{(t)}}\right]\right)\\ \geq& q_t(i)\left[1-e^{-y_i x_{i,k}\delta_k^{(t)}}\right]\ .\end{aligned}$$

Convexity of the exponent, for every $\sigma_{Q,k}\in(0,1)$, yields,

$$\begin{aligned}e^{-y_i x_{i,k}\delta_k^{(t)}} \leq& \sigma_{Q,k}\left|x_{i,k}\right| e^{-\operatorname{sign}(y_i x_{i,k})\frac{\delta_k^{(t)}}{\sigma_{Q,k}}}\\ &+\left(1-\sigma_{Q,k}\left|x_{i,k}\right|\right)e^0\ .\end{aligned}$$

Summing over the examples,

$$\begin{aligned}\Delta_t \geq& c\sum_{i=1}^m q_t(i)\sigma_{Q,k}\left|x_{i,k}\right|\left(1-e^{-\operatorname{sign}(y_i x_{i,k})\frac{\delta_k^{(t)}}{\sigma_{Q,k}}}\right)\\ =& c\sum_{i=1,y_i x_{i,k}\geq 0}^m q_t(i)\sigma_{Q,k}\left|x_{i,k}\right|\left(1-e^{-\frac{\delta_k^{(t)}}{\sigma_{Q,k}}}\right)\\ &+c\sum_{i=1,y_i x_{i,k}<0}^m q_t(i)\sigma_{Q,k}\left|x_{i,k}\right|\left(1-e^{\frac{\delta_k^{(t)}}{\sigma_{Q,k}}}\right)\\ =& c\,\sigma_{Q,k}\left(\gamma_k^+\left[1-e^{-\frac{\delta_k^{(t)}}{\sigma_{Q,k}}}\right]+\gamma_k^-\left[1-e^{\frac{\delta_k^{(t)}}{\sigma_{Q,k}}}\right]\right)\ ,\end{aligned}$$

adding the regularization terms completes the proof. ■

J Proof of Lemma 9

Proof: Without loss of generality we assume that $\gamma_k^+ e^{\frac{\mu_{Q,k}^{(t)}}{\sigma_{Q,k}}} - \gamma_k^- e^{-\frac{\mu_{Q,k}^{(t)}}{\sigma_{Q,k}}} > 0$, and in addition we assume for the sake of contradiction that, $\mu_{Q,k}^{(t)}+\delta_k^{(t)}<0$. Differentiating the objective with respect for $\delta_t^{(t)}$ and equating to zero yields:

$$\begin{aligned}&=\frac{\partial}{\partial\delta_t}\Bigg\{-\mu_{Q,k}^{(t)}-\delta_k^{(t)}+\sigma_{Q,k}e^{\frac{\mu_{Q,k}^{(t)}+\delta_k^{(t)}}{\sigma_{Q,k}}}\\ &\quad+c\sigma_{Q,k}\left(\gamma_k^+ e^{-\frac{\delta_k^{(t)}}{\sigma_{Q,k}}}+\gamma_k^- e^{\frac{\delta_k^{(t)}}{\sigma_{Q,k}}}\right)\Bigg\}\\ &=-1+e^{\frac{\mu_{Q,k}^{(t)}+\delta_k^{(t)}}{\sigma_{Q,k}}}-c\gamma_k^+ e^{-\frac{\delta_k^{(t)}}{\sigma_{Q,k}}}+c\gamma_k^- e^{\frac{\delta_k^{(t)}}{\sigma_{Q,k}}}=0\ .\end{aligned}$$

Arranging the terms:

$$-c\gamma_k^+ e^{-\frac{\delta_k^{(t)}}{\sigma_{Q,k}}}+c\gamma_k^- e^{\frac{\delta_k^{(t)}}{\sigma_{Q,k}}}=1-e^{\frac{\mu_{Q,k}^{(t)}+\delta_k^{(t)}}{\sigma_{Q,k}}}\ ,$$

the right hand side is assumed to be strictly positive, and as for the left hand side:

$$\begin{aligned}&-\gamma_k^+ e^{-\frac{\delta_k^{(t)}}{\sigma_{Q,k}}}+\gamma_k^- e^{\frac{\delta_k^{(t)}}{\sigma_{Q,k}}}\\ &<-\gamma_k^+ e^{-\frac{\delta_k^{(t)}}{\sigma_{Q,k}}}\cdot e^{\frac{\mu_{Q,k}^{(t)}+\delta_k^{(t)}}{\sigma_{Q,k}}}+\gamma_k^- e^{\frac{\delta_k^{(t)}}{\sigma_{Q,k}}}\cdot e^{-\frac{\mu_{Q,k}^{(t)}+\delta_k^{(t)}}{\sigma_{Q,k}}}\\ &=-\left(\gamma_k^+ e^{\frac{\mu_{Q,k}^{(t)}}{\sigma_{Q,k}}}-\gamma_k^- e^{-\frac{\mu_{Q,k}^{(t)}}{\sigma_{Q,k}}}\right)<0\ .\end{aligned}$$

This is a contradiction, so we must have that $\delta_k^{(t)}+\mu_{Q,k}^{(t)}\geq 0$. The proof for the symmetric case follows similarly. ■

K Experiments- Data Details:

Synthetic data: We generated $4,000$ vectors $\boldsymbol{x}_i\in\mathbb{R}^8$ sampled from a zero mean isotropic normal distribution $\boldsymbol{x}_i\sim\mathcal{N}(\mathbf{0},\mathbf{I})$. Labels were assigned by generating once per run $\boldsymbol{\omega}\in\mathbb{R}^8$ at random and using: $y_i=\operatorname{sign}(\boldsymbol{\omega}\cdot\boldsymbol{x}_i)$. Each input $\boldsymbol{x}_i$ *training* data was then corrupted with probability p by adding to it a random vector sampled from a zero mean isotropic Gaussian, $\epsilon_i\sim\mathcal{N}(\mathbf{0},\sigma\mathbf{I})$, with some positive standard-deviation σ. 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 σ.