A. Proofs

Proof for Theorem 1

Restatement of Theorem 1: There exists a loss $L$ that satisfies all the three conditions if, and only if, $f$ is affine.

Proof. The “if” part is trivial as we just need to set $L(\phi, z) = ||\phi - f(z)||^2$. To see the “only if” part, consider the sublevel set of $L$ at $0$: $S = \{(\phi, z) : L(\phi, z) \leq 0\}$. By grounding and unique recovery, $S = \{(f(z), z) : z\}$. And by the joint convexity of $L$, $S$ is convex. So for any $z_1, z_2$, $(\frac{1}{2}f(z_1) + f(z_2)), \frac{1}{2}(z_1 + z_2))$ is in $S$. But $(f(\frac{1}{2}(z_1 + z_2)), \frac{1}{2}(z_1 + z_2))$ is the only element in $S$ with the second component being $\frac{1}{2}(z_1 + z_2)$. So $\frac{1}{2}(f(z_1) + f(z_2)) = f(\frac{1}{2}(z_1 + z_2))$. So $f$ is affine.

Proof for Lemma 1

Restatement of Lemma 1: $S$ is convex, bounded, and closed. In addition,

$$\gamma_S(T) = \begin{cases} \text{tr}(T) & T \in \mathcal{T} \\ +\infty & \text{otherwise} \end{cases} \quad (18)$$

Proof. Since $\mathcal{T}$ is a convex cone, the right-hand side is a sublinear function. To show two sublinear functions $f$ and $g$ are equal, it suffices to show that their “unit balls” are equal, i.e. $\{x : f(x) \leq 1\} = \{x : g(x) \leq 1\}$. The unit ball of the left-hand side, by definition, is $S$. The unit ball of the right-hand side is: $\{T : T \in \mathcal{T}, \text{tr}(T) \leq 1\}$. But this is exactly the definition of $S$ in (7).

B. Extensions to hard tanh and non-elementwise transfers

Elementwise transfer. When using the hard tanh transfer, we have $F^e_h(\Phi) = \frac{1}{2}||\Phi||_2^2$ if the $L_\infty$ norm $||\Phi||_\infty := \max_{i,j} |\Phi_{ij}| \leq 1$, and $\infty$ otherwise. As a result, we get the same objective function as in (6), only with $T_h$ changed into $\{\Phi^' : ||\Phi^'||_\infty \leq 1\}$ and the domain of $A$ changed into $\{A : \sum_i |A_{ij}| \leq 1, \forall j\}$. Given the negative gradient $G \succeq 0$ of the objective, the polar operator boils down to solving

$$\max_{\Phi \in \mathbb{R}^{h \times t} : ||\Phi||_\infty \leq 1} \text{tr}(G^\prime \Phi^\prime) = h \max_{\phi \in [0,1]^t} \phi^\prime G \phi = h \max_{\phi \in [0,1]^t} ||A\phi||^2, \quad \text{where} \quad A^\prime A = G. \quad (19)$$

This problem is NP-hard, but an approximate solution with constant multiplicative guarantee can be found in $O(t^2)$ time (Steinberg, 2005). Note for computation we do not even need an expression of the convex hull of $T_h$.

Non-elementwise transfer. The Bregman divergence can be further leveraged to convexify transfer functions that are not applied elementwise. For example, consider the soft-max function that is commonly used in machine learning and deep learning:

$$f(x) = \left(\sum_{k=1}^h e^{x_k}\right)^{-1} (e^{x_1}, \ldots, e^{x_h})'. $$

Clearly the range of $f$ is $S^h = \{z \in \mathbb{R}^h : z > 0, 1'z = 1\}$. The potential function $F(x)$ is simply

$$F(x) = \log \sum_{k=1}^h e^{x_k}, \quad (20)$$

and its Fenchel dual is

$$F^*(\phi) = \begin{cases} \sum_{k=1}^h \phi_k \log \phi_k & \text{if } \phi \in S^h \\ \infty & \text{otherwise} \end{cases} \quad (21)$$
Therefore the objective in (4) can be instantiated into

$$\min_{\phi_j \in S^h} \max_{R_1 \in 0, \lambda_j \in S^h} \sum_{j=1}^{t} F^*(\phi_j) - \frac{1}{2} \| (\Phi - \Lambda) X' \|^2 - \frac{1}{2} \| \Phi R' \|^2 - F^*(\Lambda) - \ell^*(R). \quad (22)$$

where $\Phi = (\phi_1, \ldots, \phi_t) \in \mathbb{R}^{h \times t}$ and $\Lambda = (\lambda_1, \ldots, \lambda_t) \in \mathbb{R}^{h \times t}$. Here $S^h$ is the closure of $S^h$: $\{ z \in \mathbb{R}^h : 1'z = 1 \}$, i.e. the $h$ dimensional probability simplex.

When $h = 2$, $F^*(\phi)$ is the negative entropy function, and it can be approximated by $\frac{a}{2} [(\phi_1 - 0.5)^2 + (\phi_2 - 0.5)^2] + c$, where $a$ and $c$ are chosen such that $c = F^*(\frac{1}{h} 1) = \log \frac{1}{h}$ and $\frac{a}{2} (0.5^2 + 0.5^2) + c = F^*((0,1)'') = 0$. For general $h$, we can similarly approximate $F^*(\phi)$ by $\frac{a}{h} \| \phi - \frac{1}{h} 1 \|^2 + c$, with $c = F^*(\frac{1}{h} 1) = \log \frac{1}{h}$ and $\frac{a}{2} [(1 - \frac{1}{h})^2 + \frac{h-1}{h^2}] + c = F^*((1, 0, \ldots, 0)'') = 0$. Since $1'\phi = 1$, this approximation is in turn equal to $a \| \phi \|^2 + d$ where $d = c - a/(2h)$. As a result, (22) can be approximated by (setting $a = 1$ to ignore scaling)

$$\min_{\phi_j \in S^h} \max_{R_1 \in 0, \lambda_j \in S^h} \frac{1}{2} \| \Phi \|^2 - \frac{1}{2} \| (\Phi - \Lambda) X' \|^2 - \frac{1}{2} \| \Phi R' \|^2 - \frac{1}{2} \| \Lambda \|^2 - \ell^*(R). \quad (23)$$

Once more we can apply change of variable by $\Lambda = \Phi A$. Since $\Phi \geq 0$, $\Lambda \geq 0$, $\Phi' = 1$, and $\Lambda' = 1$, we easily derive the domain of $A$ as $A'1 = 1$ and $A \geq 0$. So using $T = \Phi^* \Phi$, we finally arrive at the convexified objective:

$$\min_{T \in T_h} \max_{R_1 \in 0, A \geq 0, A' = 1} \frac{1}{2} \text{tr}(T) - \frac{1}{2} \text{tr}((I - A)X'X(I - A')) - \frac{1}{2} \text{tr}(TR'R) - \frac{1}{2} \text{tr}(TAA') - \ell^*(R), \quad (24)$$

where $T_h$ is the convex hull of $\{ \Phi^* \Phi : \Phi \in \mathbb{R}^{h \times t}, \Phi' = 1 \}$. So given the negative gradient $G \geq 0$ of the objective, the polar operator aims to compute

$$\max_{\Phi \in \mathbb{R}^{h \times t} : \Phi' = 1} \text{tr}(G' \Phi^* \Phi) = \max_{\phi_1, \ldots, \phi_h \in \mathbb{R}^t} \sum_{k=1}^{h} \| A \phi_k \|^2 \quad \text{s.t.} \quad \sum_{k=1}^{h} \phi_k = 1, \quad \text{where} \quad A' A = G. \quad (25)$$

This problem is NP-hard (Steinberg, 2005), but an approximate solution with provable guarantee is still possible. For example, in the case that $h = 2$, we have $\phi_2 = 1 - \phi_1$, and the problem becomes

$$\max_{\phi_1 \in [0,1]^t} \| A \phi_1 \|^2 + \| A (1 - \phi_1) \|^2 = \max_{\phi_1 \in [0,1]^t} \| A (\phi_1 - \frac{1}{2} 1) \|^2 + \text{constant} \quad (26)$$

$$= \max_{\phi \in \{-\frac{1}{2}, \frac{1}{2}\}^t} \| A \phi \|^2 + \text{constant}. \quad (27)$$

This again admits an approximate solution with constant multiplicative guarantee that can be computed in $O(t^2)$ time (Steinberg, 2005).

Note the $T_h$ in this case, as well as that in the hard tanh case above, is closely related to the completely positive matrix cone, because $\Phi \in \mathbb{R}^{h \times t}_+$.  

C. Dataset description

The experiments made use of 4 “real” world datasets - G241N (241 × 1500) from (Chapelle), Letter (vowel letters A-E vs non vowel letters B-F) (16 × 20000) from (UCI, 1990), CIFAR-SM (bicycle and motorcycle vs mower and tank) (256 × 1526) from (Aslan et al., 2013) and (Krizhevsky & Hinton, 2009) and CIFAR-10 (ship vs truck) (256 × 12000) from (Krizhevsky & Hinton, 2009), where red channel features are preprocessed by averaging pixels in both the CIFAR datasets.
D. Additional results

Here we include run time results of our baselines FFNN and LOCAL.

<table>
<thead>
<tr>
<th></th>
<th>100</th>
<th>200</th>
<th>1000</th>
<th>2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Letter</td>
<td>0.05</td>
<td>0.09</td>
<td>1.84</td>
<td>2.53</td>
</tr>
<tr>
<td>G241N</td>
<td>0.035</td>
<td>0.057</td>
<td>0.45</td>
<td>N/A</td>
</tr>
<tr>
<td>XOR</td>
<td>0.03</td>
<td>0.04</td>
<td>0.16</td>
<td>1.41</td>
</tr>
<tr>
<td>CIFAR-10</td>
<td>0.051</td>
<td>0.1</td>
<td>1.9</td>
<td>2.55</td>
</tr>
</tbody>
</table>

Table 6. Training times (in minutes) for LOCAL on 100, 200, 1000, and 2000 training examples

<table>
<thead>
<tr>
<th></th>
<th>100</th>
<th>200</th>
<th>1000</th>
<th>2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Letter</td>
<td>0.0031</td>
<td>0.0025</td>
<td>0.006</td>
<td>0.0075</td>
</tr>
<tr>
<td>G241N</td>
<td>0.023</td>
<td>0.028</td>
<td>0.054</td>
<td>N/A</td>
</tr>
<tr>
<td>XOR</td>
<td>0.02</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>CIFAR-10</td>
<td>0.047</td>
<td>0.039</td>
<td>0.073</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Table 7. Training times (in minutes) for FFNN on 100, 200, 1000, and 2000 training examples
References


