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, $\mathbf{f}$ is affine.

Proof. The "if" part is trivial as we just need to set $L(\boldsymbol{\phi}, \mathbf{z}) = ||\boldsymbol{\phi} - \mathbf{f}(\mathbf{z})||^2$. To see the "only if" part, consider the sublevel set of L at 0: $S = \{(\boldsymbol{\phi}, \mathbf{z}) : L(\boldsymbol{\phi}, \mathbf{z}) \leq 0\}$. By grounding and unique recovery, $S = \{(\mathbf{f}(\mathbf{z}), \mathbf{z}) : \mathbf{z}\}$. And by the joint convexity of L, S is convex. So for any $\mathbf{z}_1, \mathbf{z}_2$, $(\frac{1}{2}(\mathbf{f}(\mathbf{z}_1) + \mathbf{f}(\mathbf{z}_2)), \frac{1}{2}(\mathbf{z}_1 + \mathbf{z}_2))$ is in S. But $(\mathbf{f}(\frac{1}{2}(\mathbf{z}_1 + \mathbf{z}_2)), \frac{1}{2}(\mathbf{z}_1 + \mathbf{z}_2))$ is the only element in S with the second component being $\frac{1}{2}(\mathbf{z}_1 + \mathbf{z}_2)$. So $\frac{1}{2}(\mathbf{f}(\mathbf{z}_1) + \mathbf{f}(\mathbf{z}_2)) = \mathbf{f}(\frac{1}{2}(\mathbf{z}_1 + \mathbf{z}_2))$. So $\mathbf{f}$ is affine. $\square$

Proof for Lemma 1

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

$$\gamma_{\mathcal{S}}(T) = \begin{cases} \mathrm{tr}(T) & T \in \mathcal{T} \\ +\infty & \text{otherwise} \end{cases}. \tag{18}$$

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.* $\{\mathbf{x} : f(\mathbf{x}) \leq 1\} = \{\mathbf{x} : g(\mathbf{x}) \leq 1\}$. The unit ball of the left-hand side, by definition, is $\mathcal{S}$. The unit ball of the right-hand side is: $\{T : T \in \mathcal{T}, \mathrm{tr}(T) \leq 1\}$. But this is exactly the definition of $\mathcal{S}$ in (7). $\square$

B. Extensions to hard tanh and non-elementwise transfers

Elementwise transfer. When using the **hard tanh** transfer, we have $F_h^*(\Phi) = \frac{1}{2}\|\Phi\|^2$ if the L_∞ norm $\|\Phi\|_\infty := \max_{ij}|\Phi_{ij}| \leq 1$, and ∞ otherwise. As a result, we get the same objective function as in (6), only with $\mathcal{T}_h$ changed into $\{\Phi'\Phi : \|\Phi\|_\infty \leq 1\}$ and the domain of A changed into $\{A : \sum_i |A_{ij}| \leq 1, \ \forall\, j\}$. Given the negative gradient $G \succeq \mathbf{0}$ of the objective, the polar operator boils down to solving

$$\max_{\Phi \in \mathbb{R}^{h\times t}: \|\Phi\|_\infty \leq 1} \mathrm{tr}(G'\Phi'\Phi) \;=\; h \max_{\boldsymbol{\phi}\in[0,1]^t} \boldsymbol{\phi}' G \boldsymbol{\phi} \;=\; h \max_{\boldsymbol{\phi}\in[0,1]^t} \|A\boldsymbol{\phi}\|^2, \quad \text{where} \quad A'A = G. \tag{19}$$

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 $\mathcal{T}_h$.

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:

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

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

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

and its Fenchel dual is

$$F^*(\boldsymbol{\phi}) = \begin{cases} \sum_{k=1}^{h} \phi_k \log \phi_k & \text{if } \boldsymbol{\phi} \in S^h \\ \infty & \text{otherwise} \end{cases}. \tag{21}$$

Therefore the objective in (4) can be instantiated into

$$\min_{\boldsymbol{\phi}_j \in \mathcal{S}^h} \max_{R\mathbf{1}=\mathbf{0}, \boldsymbol{\lambda}_j \in \mathcal{S}^h} \sum_{j=1}^{t} F^*(\boldsymbol{\phi}_j) - \frac{1}{2}\left\|(\Phi - \Lambda)X'\right\|^2 - \frac{1}{2}\left\|\Phi R'\right\|^2 - F^*(\Lambda) - \ell^*(R). \tag{22}$$

where $\Phi = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_t) \in \mathbb{R}^{h\times t}$ and $\Lambda = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_t) \in \mathbb{R}^{h\times t}$. Here $\mathcal{S}^h$ is the closure of S^h: $\{\mathbf{z} \in \mathbb{R}^h_+ : \mathbf{1}'\mathbf{z} = 1\}$, *i.e.* the h dimensional probability simplex.

When $h = 2$, $F^*(\boldsymbol{\phi})$ 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 $c = F^*(\frac{1}{2}\mathbf{1}) = \log\frac{1}{2}$ and $\frac{a}{2}(0.5^2 + 0.5^2) + c = F^*((0,1)') = 0$. For general h, we can similarly approximate $F^*(\boldsymbol{\phi})$ by $\frac{a}{2}\left\|\boldsymbol{\phi} - \frac{1}{h}\mathbf{1}\right\|^2 + c$, with $c = F^*(\frac{1}{h}\mathbf{1}) = \log\frac{1}{h}$ and $\frac{a}{2}[(1-\frac{1}{h})^2 + \frac{h-1}{h^2}] + c = F^*((1,0,\ldots,0)') = 0$. Since $\mathbf{1}'\boldsymbol{\phi} = 1$, this approximation is in turn equal to $a\left\|\boldsymbol{\phi}\right\|^2 + d$ where $d = c - a/(2h)$. As a result, (22) can be approximated by (setting $a = 1$ to ignore scaling)

$$\min_{\boldsymbol{\phi}_j \in \mathcal{S}^h} \max_{R\mathbf{1}=\mathbf{0}, \boldsymbol{\lambda}_j \in \mathcal{S}^h} \frac{1}{2}\left\|\Phi\right\|^2 - \frac{1}{2}\left\|(\Phi - \Lambda)X'\right\|^2 - \frac{1}{2}\left\|\Phi R'\right\|^2 - \frac{1}{2}\left\|\Lambda\right\|^2 - \ell^*(R). \tag{23}$$

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

$$\min_{T \in \mathcal{T}_h} \max_{R\mathbf{1}=\mathbf{0}, A\geq\mathbf{0}, A'\mathbf{1}=\mathbf{1}} \frac{1}{2}\operatorname{tr}(T) - \frac{1}{2}\operatorname{tr}(T(I-A)X'X(I-A')) - \frac{1}{2}\operatorname{tr}(TR'R) - \frac{1}{2}\operatorname{tr}(TAA') - \ell^*(R), \tag{24}$$

where $\mathcal{T}_h$ is the convex hull of $\{\Phi'\Phi : \Phi \in \mathbb{R}^{h\times t}_+, \Phi'\mathbf{1} = \mathbf{1}\}$. So given the negative gradient $G \succeq \mathbf{0}$ of the objective, the polar operator aims to compute

$$\max_{\Phi \in \mathbb{R}^{h\times t}_+ : \Phi'\mathbf{1}=\mathbf{1}} \operatorname{tr}(G'\Phi'\Phi) \quad = \quad \max_{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_h \in \mathbb{R}^t_+} \sum_{k=1}^{h}\left\|A\boldsymbol{\phi}_k\right\|^2 \quad s.t. \quad \sum_{k=1}^{h}\boldsymbol{\phi}_k = \mathbf{1}, \quad \text{where} \quad A'A = G. \tag{25}$$

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 $\boldsymbol{\phi}_2 = \mathbf{1} - \boldsymbol{\phi}_1$, and the problem becomes

$$\max_{\boldsymbol{\phi}_1 \in [0,1]^t} \left\|A\boldsymbol{\phi}_1\right\|^2 + \left\|A(\mathbf{1} - \boldsymbol{\phi}_1)\right\|^2 = \max_{\boldsymbol{\phi}_1 \in [0,1]^t} \left\|A\left(\boldsymbol{\phi}_1 - \tfrac{1}{2}\mathbf{1}\right)\right\|^2 + \text{constant} \tag{26}$$

$$= \max_{\boldsymbol{\phi} \in [-\frac{1}{2}, \frac{1}{2}]^t} \left\|A\boldsymbol{\phi}\right\|^2 + \text{constant}. \tag{27}$$

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

Note the $\mathcal{T}_h$ 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 lawn- 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.

	100	200	1000	2000
Letter	0.05	0.09	1.84	2.53
G241N	0.035	0.057	0.45	N/A
XOR	0.03	0.04	0.16	1.41
CIFAR-10	0.051	0.1	1.9	2.55

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

	100	200	1000	2000
Letter	0.0031	0.0025	0.006	0.0075
G241N	0.023	0.028	0.054	N/A
XOR	0.02	0.03	0.03	0.03
CIFAR-10	0.047	0.039	0.073	0.1

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

References

Anandkumar, Animashree, Ge, Rong, Hsu, Daniel, Kakade, Sham M., and Telgarsky, Matus. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832, 2014.

Aslan, Ozlem, Cheng, Hao, Zhang, Xinhua, and Schuurmans, Dale. Convex two-layer modeling. In *Neural Information Processing Systems*, 2013.

Aslan, Ozlem, Zhang, Xinhua, and Schuurmans, Dale. Convex deep learning via normalized kernels. In *Neural Information Processing Systems*, 2014.

Auer, P., Herbster, M., and Warmuth, M. K. Exponentially many local minima for single neurons. Technical Report UCSC-CRL-96-1, Univ. of Calif.Computer Research Lab, Santa Cruz, CA, 1996. In preparation.

Banerjee, A., Merugu, S., Dhillon, I. S., and Ghosh, J. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

Bengio, Yoshua, Roux, Nicolas Le, Vincent, Pascal, Delalleau, Olivier, and Marcotte, Patrice. Convex neural networks. In *Neural Information Processing Systems*, 2005.

Berman, Abraham and Shaked-Monderer, Naomi. *Completely Positive Matrices*. World Scientific, 2003.

Brutzkus, Alon and Globerson, Amir. Globally optimal gradient descent for a ConvNet with gaussian inputs. In *Proc. Intl. Conf. Machine Learning*, 2017.

Carreira-Perpinnan, M.A. and Wang, Weiran. Distributed optimization of deeply nested systems. In *Proc. Intl. Conference on Artificial Intelligence and Statistics*, 2014.

Chapelle, Olivier. http://olivier.chapelle.cc/ssl-book/benchmarks.html.

Cheng, Hao, Yu, Yaoliang, Zhang, Xinhua, Xing, Eric, and Schuurmans, Dale. Scalable and sound low-rank tensor learning. In *Proc. Intl. Conference on Artificial Intelligence and Statistics*, 2016.

Choromanska, Anna, Henaff, Mikael, Mathieu, Michael, Arous, Gerard Ben, and LeCun, Yann. The loss surfaces of multilayer networks. In *Proc. Intl. Conference on Artificial Intelligence and Statistics*, 2014.

Dauphin, Y., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., and Bengio, Y. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Neural Information Processing Systems*, 2014.

Dickinson, Peter J. C. and Gijben, Luuk. On the computational complexity of membership problems for the completely positive cone and its dual. *Computational Optimization and Applications*, 57(2):403–415, Mar 2014. ISSN 1573-2894. doi: 10.1007/s10589-013-9594-z. URL `https://doi.org/10.1007/s10589-013-9594-z`.

Ekeland, Ivar and Témam, Roger. *Convex Analysis and Variational Problems*. SIAM, 1999.

Elad, M. and Aharon, M. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 15(12):3736–3745, 2006.

Fogel, F., Jenatton, R., Bach, F., and d'Aspremont, A. Convex relaxations for permutation problems. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1465–1488, 2015.

Freund, Robert and Grigas, Paul. New analysis and results for the Frank-Wolfe method. *Mathematical Programming*, 155(1):199–230, 2016.

Gens, Robert and Domingos, Pedro. Discriminative learning of sum-product networks. In *Neural Information Processing Systems*, 2012.

Harchaoui, Zaid, Juditsky, Anatoli, and Nemirovski, Arkadi. Conditional gradient algorithms for norm-regularized smooth convex optimization. *Mathematical Programming*, 152:75–112, 2015.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Jaggi, Martin. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proc. Intl. Conf. Machine Learning*, 2013.

Kawaguchi, Kenji. Deep learning without poor local minima. In *Neural Information Processing Systems*, 2016.

Krizhevsky, Alex and Hinton, Geoffrey. Learning multiple layers of features from tiny images. 2009.

Lacoste-Julien, S. and Jaggi, M. On the global linear convergence of Frank-Wolfe optimization variants. In *Neural Information Processing Systems*, 2015.

Lawrence, Neil. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *J. Mach. Learn. Res.*, 6:1783–1816, 2005.

Lichman, M. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

Livni, R., Shalev-Shwartz, S., and Shamir, O. An algorithm for training polynomial networks. arXiv:1304.7045v2, 2014.

Murty, Katta G. and Kabadi, Santosh N. Some NP-complete problems in quadratic and nonlinear programming. *Mathematical Programming*, 39(2):117–129, 1987.

Nemirovski, A., Roos, C., and Terlaky, T. On maximization of quadratic form over intersection of ellipsoids with common center. *Math. Program. Ser. A*, 86:463–473, 1999.

Nesterov, Yurii. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.

Nguyen, Quynh and Hein, Matthias. The loss surface of deep and wide neural networks. In *Proc. Intl. Conf. Machine Learning*, 2017a.

Nguyen, Quynh and Hein, Matthias. The loss surface and expressivity of deep convolutional neural networks. arXiv:1710.10928, 2017b.

Ranzato, Marc'aurelio, Monga, Rajat, Devin, Matthieu, Chen, Kai, Corrado, Greg, Dean, Jeff, Le, Quoc V., and Ng, Andrew Y. Building high-level features using large scale unsupervised learning. In *Proc. Intl. Conf. Machine Learning*, 2012.

Rifai, Salah, Vincent, Pascal, Muller, Xavier, Glorot, Xavier, and Bengio, Yoshua. Contractive auto-encoders: Explicit invariance during feature extraction. In *Proc. Intl. Conf. Machine Learning*, 2011.

Silver, David, Huang, Aja, Maddison, Chris J., Guez, Arthur, Sifre, Laurent, van den Driessche, George, Schrittwieser, Julian, Antonoglou, Ioannis, Panneershelvam, Veda, Lanctot, Marc, Dieleman, Sander, Grewe, Dominik, Nham, John, Kalchbrenner, Nal, Sutskever, Ilya, Lillicrap, Timothy, Leach, Madeleine, Kavukcuoglu, Koray, Graepel, Thore, and Hassabis, Demis. Mastering the game of go with deep neural networks and tree search. *Science*, 529:484–489, 2016.

Soltanolkotabi, Mahdi, Javanmard, Adel, and Lee, Jason D. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. In *Proc. Intl. Conf. Machine Learning*, 2017.

Steinberg, Daureen. *Computation of matrix norms with applications to Robust Optimization*. PhD thesis, Faculty of Industrial Engineering and Management, Technion, 2005.

Sutskever, Ilya, Vinyals, Oriol, and Le, Quoc V. Sequence to sequence learning with neural networks. In *Neural Information Processing Systems*, 2014.

Tian, Yuandong. An analytical formula of population gradient for two-layered ReLU network and its applications in convergence and critical point analysis. In *Proc. Intl. Conf. Machine Learning*, 2017.

Tishby, N., Pereira, F., and Bialek, W. The information bottleneck method. In *37-th Annual Allerton Conference on Communication, Control and Computing*, 1999.

UCI. University of California Irvine: Machine Learning Repository, 1990.

Woodbury, Max A. Inverting modified matrices. Technical Report MR38136, Memorandum Rept. 42, Statistical Research Group, Princeton University, Princeton, NJ, 1950.

Zhang, X., Yu, Y., and Schuurmans, D. Accelerated training for matrix-norm regularization: A boosting approach. In *Neural Information Processing Systems*, 2012.

Zhang, Y., Lee, JD., and Jordan, MI. L1-regularized neural networks are improperly learnable in polynomial time. In *Proc. Intl. Conf. Machine Learning*, 2016.

Zhang, Yuchen, Liang, Percy, and Wainwright, Martin. Convexified convolutional neural networks. In *Proc. Intl. Conf. Machine Learning*, 2017.

Zhong, Kai, Song, Zhao, Jain, Prateek, Bartlett, Peter, and Dhillon, Inderjit. Recovery guarantees for one-hidden-layer neural networks. In *Proc. Intl. Conf. Machine Learning*, 2017.