Supplementary material

Appendix A Variable Scaling

Note that the new model (9) has introduced $L+1$ more hyperparameters. We can use variable scaling and the dual formulation to show how to effectively reduce this to only *one* hyperparameter. Consider the model with ReLU activations, that is, the biconvex function as in (5) and regularization functions $\pi_l(W_l) = \|W_l\|_F^2$ for $l = 0, \dots, L$. Note that B_ϕ is homogeneous of degree 2, that is for any U, V and γ we have

$$\gamma B_\phi(V, U) = B_\phi(\sqrt{\gamma}V, \sqrt{\gamma}U)$$

Define $\lambda_{-1} = 1$ and the scalings

$$\bar{X}_l := \sqrt{\lambda_{l-1}} X_l, \;\; \bar{W}_l := \sqrt{\frac{\lambda_l}{\lambda_{l-1}}} W_l,$$

Then (9) becomes

$$\begin{aligned} G(\lambda) := \min_{(\bar{W}_l)_{l=0}^{L}, (\bar{X}_l)_1^{L+1}} \; & \mathcal{L}(Y, \sqrt{\lambda_L}(\bar{W}_L \bar{X}_L)) \\ & + \sum_{l=0}^{L} \rho_l \pi_l(\sqrt{\frac{\lambda_{l-1}}{\lambda_l}} W_l) + \sum_{l=0}^{L-1} B_l(\bar{X}_{l+1}, \bar{W}_l \bar{X}_l) \\ \text{s.t. } & \bar{X}_0 = X, \; \bar{X}_l \geq 0, \; l = 0, \dots, L \end{aligned} \tag{21}$$

Using the fact $\pi_l(W_l) = \|W_l\|_F^2$ and defining $\bar{\rho}_l = \rho_l \dfrac{\lambda_{l-1}}{\lambda_l}$ we have

$$\begin{aligned} G(\lambda) := \min_{(\bar{W}_l)_{l=0}^{L}, (\bar{X}_l)_1^{L+1}} \; & \mathcal{L}(Y, \sqrt{\lambda_L}(\bar{W}_L \bar{X}_L)) \\ & + \sum_{l=0}^{L} \bar{\rho}_l \|\bar{W}_l\|_F^2 + \sum_{l=0}^{L-1} B_l(\bar{X}_{l+1}, \bar{W}_l \bar{X}_l) \\ \text{s.t. } & \bar{X}_0 = X, \; \bar{X}_l \geq 0, \; l = 0, \dots, L \end{aligned} \tag{22}$$

where $G(\lambda)$ is now only a function of one variable λ_L as opposed to L variables. Note that this argument for variable scaling still works when we use average pooling or convolution operations in conjunction with a ReLU activation since they are linear operations. Note furthermore that the same scaling argument works in place of any norm due to the homogeneity of norms – the only thing that would change is how $\bar{\rho}$ is scaled by λ_{l-1} and λ_l.

Another way to show that we only require one hyperparameter λ is to note the equivalence

$$B_l(v, u) \leq 0 \;\; \forall l \iff \sum_l B_l(v, u) \leq 0$$

Then we may replace the L biconvex constraints in (6) by the equivalent constraint $\sum_l B_l(v, u) \leq 0$. Since this is only one constraint, when we dualize we only introduce *one* Lagrange multiplier λ.

Appendix B One-layer Regression Setting

In this section, we show that for a one layer network we are able to convert a non-convex optimization problem into a convex one by using the BC condition described in the main text.

Consider a regression setting where $Y = \phi(W^*X)$ for some fixed $W^* \in \mathbb{R}^{p\times n}$ and a given data matrix $X \in \mathbb{R}^{n\times m}$. Given a training set (X, Y) we can solve for W by solving the following non-convex problem

$$\min_W \ \|Y - \phi(WX)\|_F^2. \tag{23}$$

We could also solve the following relaxation of (23) based on the BC condition

$$\min_W \ B_\phi(Y, WX) \tag{24}$$

Note (24) is trivially convex in W by definition of $B_\phi(\cdot,\cdot)$. Furthermore, by construction $B_\phi(Y, WX) \geq 0$ and $B_\phi(Y, WX) = 0$ if and only if $Y = \phi(WX)$. Since $Y = \phi(W^*X)$, it follows W^* (which is the minimizer of (23)) is a global minimizer of the convex program (24). Therefore, we can solve the original non-convex problem (23) to global optimality by instead solving the convex problem presented in (24).

Appendix C Hyperparameters for Experiments

For all experiments that used batching, the batch size was fixed at 500 and $K = 1$. We observed empirically that larger batch sizes improved the performance of the lifted models. To speed up computations, we set $K = 1$ and empirically find this does not affect final test set performance. For batched models, we do not use $\pi_l(\cdot)$ since we explicitly regularize through batching (see (13)) while for the non-batched models we set $\pi_l(W_l) = \|W_l\|_F^2$ for all l. For models trained using Adam, the learning rate was set to $\eta = 10^{-3}$ and for models trained using SGD, the learning rate was set to $\eta = 10^{-2}$. The learning rates were a hyperparamter that we picked from $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$ to give the best final test performance for both Adam and SGD.

For the network architechtures described in the experimental results, we used the following hyperparamters:

- Fenchel Lifted Network for LeNet-5 architecture
  1. $\rho_1 = 1e-4, \lambda_1 = 5$
  2. $\rho_2 = 1e-2, \lambda_2 = 5$
  3. $\rho_3 = 1, \quad \lambda_3 = 1$
  4. $\rho_4 = 1, \quad \lambda_4 = 1$
  5. $\rho_5 = 1$
- Fenchel Lifted Network for 784-300-10 architecture (batched)
  1. $\rho_1 = 1, \quad \lambda_1 = 0.1$
  2. $\rho_2 = 100$
- Fenchel Lifted Network for 784-300-10 architecture (non-batched)
  1. $\rho_1 = 1e-2, \ \lambda_1 = 0.1$
  2. $\rho_2 = 10$

For all weights the initialization is done through Xavier initialization implemented in TensorFlow. The ρ variables are chosen to balance the change of variables across layers in iterations. Although the theory in Appendix A states we can collapse all λ hyperparameters into a single hyperparameter, due to time constraints, we were unable to implement this change upon submission. We also stress that the hyperparamter search over the ρ's were very coarse and a variety of ρ values worked well in practice; for simplicity we only present the ones we used to produce the plots in the experimental results.

Appendix D Fenchel Conjugates and Proximal Operators

Here we discuss the similarities between Li et al. (2019) and the approach of this paper (for simplicity, we only concern ourselves with the ReLU activation since it is convex). In what follows, when we refer to equation numbers, they are the equation numbers in Li et al. (2019). First we derive an elementary result relating conjugate functions and proximal operators.

Lemma 1. *Let $\lambda > 0$ and let $f(x)$ be a closed, convex and proper function. Define $\tilde{f}(x) = \lambda f(x) + \frac{1}{2}\|x\|_2^2$ and let $f^*(y)$ be the fenchel conjugate of $f(x)$. Furthermore, define the proximal operator as $\boldsymbol{prox}_{\lambda f}(x) = \arg\min_y f(y) + \frac{1}{2\lambda}\|x-y\|_2^2$ and for a given x, let $y^*(x) = \arg\max x^\top y - \tilde{f}(y)$. Then $\boldsymbol{prox}_{\lambda f}(x) = y^*(x)$.*

Proof.

$$\begin{aligned}\arg\min_y f(y) + \frac{1}{2\lambda}\|y-x\|_2^2 &= \arg\min_y f(y) + \frac{1}{2\lambda}\|x\|_2^2 - \frac{1}{\lambda}x^\top y + \frac{1}{2\lambda}\|y\|_2^2 \\ &= \arg\min_y \left(\lambda f(y) + \frac{1}{2}\|y\|_2^2\right) - x^\top y \\ &= \arg\max_y x^\top y - \tilde{f}(y)\end{aligned}$$

The left hand side is exactly $\texttt{prox}_{\lambda f}(x)$ and the right hand side is exactly $y^*(x)$. Note furthermore that the problem defining $\texttt{prox}_{\lambda f}(x)$ is strongly convex and hence there is only one unique global optima and similarly for the problem defining $y^*(x)$. □

The above lemma shows the natural connection between proximal operators and fenchel conjugates. We now highlight this in the case of the ReLU function $\phi(x) = \max(0, x)$ and make the connection explicit. Below we consider the scalar case, and the multivariate case is a simple generalization of the argument below.

As in Li et al. (2019), if we set $f(x) = \int_0^x \phi^{-1}(z) - z\, dz$ as defined below (11) and set $g(x) = \int_0^x \phi(z) - z\, dz$ as defined below (18) in the aforementioned reference and, we then have

$$\begin{aligned}f(x) &= \int_0^x \phi^{-1}(z) - z \ = 0 \\ g(x) &= \int_0^x \phi(z) - z\, dz = \frac{1}{2}\max(x,0)^2 - \frac{1}{2}x^2\end{aligned}$$

where we use the fact that $\phi^{-1}(z) = z$ for $z \in [0, \infty)$ and set $\phi^{-1}(z) = +\infty$ otherwise. Modulo hyperparameters in their objective function, the term inside the summand in (18) (in the scalar case), reduces to

$$\begin{aligned}&f(x^i) + \frac{1}{2}(x^i - w^{i-1}x^{i-1})^2 + g(w^{i-1}x^{i-1}) \\ =&0 + \frac{1}{2}(x^i - w^{i-1}x^{i-1})^2 + \frac{1}{2}(w^{i-1}x^{i-1})_+^2 - \frac{1}{2}(w^{i-1}x^{i-1})^2 \\ =&\frac{1}{2}(x^i)^2 - \langle w^{i-1}x^{i-1}, x^i\rangle + \frac{1}{2}(w^{i-1}x^{i-1})_+^2 \\ =&B_\phi(x^i, w^{i-1}x^{i-1})\end{aligned}$$

Hence

$$B_\phi(v,u) = f(v) + \frac{1}{2}\|v-u\|_2^2 + g(u)$$

As a result, the term in the summand the authors use in (18) is equivalent to the fenchel lifted formulation.