Particle Networks: A Variation on Multilayer Perceptrons with Spatial Pairwise Kernels

Adrian W. Lange

January 6, 2016

Abstract

We present a formulation of an Artificial Neural Network akin to the classic Multilayer Perceptron model but with weight matrices replaced by implicit pairwise kernels between units. We call our neural network model a Particle Network. Each unit in a Particle Network is conceptualized as a particle with a variable amplitude, phase, and position in $\mathbb{R}^n$. Units interact between layers via a chosen kernel function, and input data is passed through the network in the usual feed-forward pattern of Multilayer Perceptrons. We exhibit our model with an experiment on the MNIST digit dataset and discuss possible advantages of Particle Networks, such as $O(N)$ computer memory cost, $O(N \log N)$ algorithms, and parallelism strategies.

1 Introduction

The Multilayer Perceptron (MLP) model is an Artificial Neural Network (ANN) that, in an extremely abstract sense, resembles how information might be received and transformed in an intelligent biological system like a human brain. For instance, the firing of a neuron is modeled mathematically as an activation function associated with each unit of the MLP. Additionally, interneuron connections of a brain are modeled in MLPs as cross-layer unit pair connections controlled by a weight variable.

We are interested in alternative representations of inter-unit connections. The reduction of an interneuron connection into a single independent weight variable seems to us to lack certain features of biological brains, namely spatial relationships of neurons. We have been motivated to consider if incorporating the notion of spatial relationships between units yields any benefits.[1] We have thus formulated an alternative version of the MLP in which each node is conceptualized as a particle with a variable amplitude, phase, and position in $\mathbb{R}^n$. We refer to our model as the Particle Network (PN).

Our intent with PNs is not necessarily to make a model more closely resemble a biological system but more of an exploratory study. We are curious to see if the inter-unit connections in PNs might self-organize into interesting clusters or structures, which may be useful for visualization and/or model interpretation—contrasting to the near non-interpretability of weights in MLPs. If such structures appear regularly as part of optimized PNs, it may be possible that one could pre-construct PNs with common structural motifs when fitting a new PN to potentially enhance deep learning, similar to starting with trained autoencoders followed by fine tuning with a deep MLP. Also, PNs could be used as an alternative in the "fully connected" layers often seen

[1] It is pretty unlikely that we are the first to envision using positions as part of an ANN. However, we (the author, me, Adrian) are not particularly well-versed in the vast amount of ANN literature. And so, we make no claim of novelty here but do claim to have arrived at this idea independently (and naïvely). Others may have come up with similar models, but we still felt like writing a up study about it anyway.

in deep convolution networks and recurrent netoworks. Or, since convolution/recurrent layers are ultimately abstractions of MLPs, one could imagine using PNs as the foundation for convolution/recurrent layers instead (*e.g.* a convolution filter based on a PN as opposed to a MLP).

Furthermore, there may exist certain reusable or optimized sets of pairwise kernels and parameters for PNs, much like the force fields used widely in the field of molecular mechanics simulations. Also along such lines, PNs could maybe benefit from certain algorithmic approaches frequently used in N-body simulations, such as distance cutoffs or Fast Multipole Methods, to reduce computational cost to $O(N \log N)$.

We present the formulation of the PN model in the following section, including its analytic gradient (*i.e.*, backpropagation gradient) and an overview of the algorithm we have used to implement it with reasonable efficiency. We then in Section 3 perform an experiment with a PN on the commonly studied MNIST digit dataset. Comparisons are drawn between PNs and MLPs and discussed throughout this work.

In summary of what follows, we claim that PNs potentially offer these benefits:

- $O(N)$ number of parameters and memory cost with respect to number of units in each layer.

- $O(N \log N)$ computational cost if implemented with a Fast Multipole Method approach.

- Reduced computational cost with distance cutoffs, potentially scaling roughly linearly (*e.g.* with neighbor lists). Can be used in combination with domain decomposition in parallelism strategies.

- Automatic feature clustering in spatial dimensions, which can be useful for interpretation and/or for pruning less important features (or units) in the network.

2 Model Formulation and Implementation

We briefly review the formulation of MLPs before introducing the formulation of Particle Networks in order to make clear comparisons. We follow the notation of Graves [] as much as possible.

2.1 Multilayer Perceptron Formulation

Consider a hidden unit h in a MLP with activation function θ_h that is connected to I input units, each with input data x_i. The output b_h is given by the activation function applied to the weighted sum

$$a_h = \sum_i^I w_{ih} x_i \tag{1}$$

$$b_h = \theta_h(a_h) \tag{2}$$

where w_{ij} is the weight between unit i and unit j. It is common practice to include an additional fixed input $x_0 = 1$ such that the weight w_{0h} constitutes a constant offset, or bias term. The output of each hidden unit is propagated forward recursively to the l-th layer hidden layer H_l

$$a_h = \sum_{h' \in H_{l-1}} w_{h'h} b_{h'} \tag{3}$$

A bias term may also be included in Eq. 3. The output layer of a MLP, with units a_k for K output units, follows similar to Eq. 3. The output layer generally may employ any activation function, though it is typical for classification problems to chose the so-called softmax function to yield output class probabilities, y_k,

$$y_k = \frac{e^{a_k}}{\sum_{k'}^K e^{a_{k'}}} \tag{4}$$

Finally, a loss function $\mathcal{L}(x, z)$ is chosen to score the accuracy of the MLP mapping of input x to predicted output y against the true value z

to which the MLP is fit. Several choices for loss function exist, but when fitting a MLP to predict multiple class output, the preferred choice is the categorical cross-entropy function

$$\mathcal{L}(x,z) = -\sum_{k}^{K} z_k \ln(y_k) \tag{5}$$

To fit a MLP for a supervised learning problem, the loss function is minimized, which can be accomplished in a number of ways but most often through a gradient-based approach, like stochastic gradient descent. The gradient of the loss function with respect to each unit's weights is computed via the backpropagation technique involving a backward pass of information through the network. For a MLP employing the categorical cross entropy loss function [Eq. 5] in conjunction with an output layer softmax activation function [Eq. 4], one has

$$\frac{\partial \mathcal{L}}{\partial a_k} = y_k - z_k \tag{6}$$

For the output layer, Then, one can recursively compute the following quantity at each layer (and similarly for the output layer)

$$\delta_h = \frac{\partial \mathcal{L}(x,z)}{\partial a_h} = \theta'_h(a_h) \sum_{h' \in H_{l+1}} \delta_{h'} w_{hh'} \tag{7}$$

where $\theta'(a_h)$ is the function $d\theta(a)/da$. Eq. 7 then enters the gradient for each unit through the chain rule:

$$\frac{\partial \mathcal{L}}{\partial w_{ij}} = \frac{\partial \mathcal{L}(x,z)}{\partial a_j}\frac{\partial a_j}{\partial w_{ij}} = \delta_j b_i \tag{8}$$

2.2 Particle Network Formulation

The central concept of a Particle Network (PN) is to consider letting each h-th unit of a MLP be given a single amplitude (or weight) q_h, a phase ϕ_h, and a position $\vec{r}_h$ in $\mathbb{R}^n$. We choose for the remainder of this work to simply use three dimensional Euclidean space, $\mathbb{R}^3$.[2] We then allow the particles (or units) of layer l to interact in a pairwise fashion with the particles of layer $l+1$ as follows,

$$a_h = q_h \sum_{h' \in H_{l-1}} k(\vec{r}_h, \vec{r}_{h'}, \phi_h, \phi_{h'}) b_{h'} \tag{9}$$

The pairwise kernel function, $k(\vec{r}_h, \vec{r}_{h'}, \phi_h, \phi_{h'}) = k_{hh'}$, can generally be any function to couple the coordinates and phases of particles h and h'. For now, we suggest a function that decreases as inter-particle distance increases, does not diverge at zero distance, and is modulated by the relative phase of the particles. In this work, we select a product of a Gaussian function and a cosine function as our kernel:

$$k_{hh'} = e^{-|\vec{r}_h - \vec{r}_{h'}|^2} \cos(\phi_h - \phi_{h'}) \tag{10}$$

In addition to Eq. 9, we treat the input layer specially such that it too carries a phase, and position but that it's amplitude is simply the input data:

$$a_i = b_i = x_i \quad \text{for } i \in I \tag{11}$$

Eqs. 9 and 11 along with the choice of a kernel, like Eq. 10, constitute the major variation of a PN from the usual MLP. All other features of the MLP are assumed to be the same in a PN, such as the use of activation functions θ_h.

The gradient of a PN can then be derived and adapted into backpropagation with the following:

$$\frac{\partial a_h}{\partial q_h} = \sum_{h' \in H_{l-1}} k(\vec{r}_h, \vec{r}_{h'}) b_{h'} \tag{12}$$

$$\frac{\partial a_i}{\partial q_i} = x_i \quad \text{for } i \in I \tag{13}$$

$$\frac{\partial a_h}{\partial \vec{r}_h} = \sum_{h' \in H_{l-1}} b_{h'} \frac{\partial k(\vec{r}_h, \vec{r}_{h'})}{\partial \vec{r}_h} \tag{14}$$

$$\frac{\partial k_{hh'}}{\partial \vec{r}_h} = -2e^{-|\vec{r}_h - \vec{r}_{h'}|^2} \cos(\phi_h - \phi_{h'})(\vec{r}_h - \vec{r}_{h'}) \tag{15}$$

$$\frac{\partial k_{hh'}}{\partial \phi_h} = -e^{-|\vec{r}_h - \vec{r}_{h'}|^2} \sin(\phi_h - \phi_{h'}) \tag{16}$$

[2]It is entirely conceivable to make other choices, such as $\mathbb{R}^2$ or, say, hyperbolic space, if that's your thing.

By symmetry of the pairwise kernel (i.e., Newton's 3^{rd} Law), we have

$$\frac{\partial a_h}{\partial \vec{r}_{h'}} = -\sum_{h' \in H_{l-1}} b_{h'} \frac{\partial k_{hh'}}{\partial \vec{r}_h} \tag{17}$$

and

$$\frac{\partial a_h}{\partial \phi_{h'}} = -\sum_{h' \in H_{l-1}} b_{h'} \frac{\partial k_{hh'}}{\partial \phi_h} \tag{18}$$

which can also be applied to the input layer by letting $h' \to i$ and $H_{l-1} \to I$. Thus, the modified version of Eq. 7 for a PN is

$$\delta_h = \frac{\partial \mathcal{L}(x,z)}{\partial a_h} = \theta'_h(a_h) \sum_{h' \in H_{l+1}} \delta_{h'} q_{h'} k(\vec{r}_h, \vec{r_{h'}}) \tag{19}$$

The loss function gradient with respect to amplitudes is then

$$\frac{\partial \mathcal{L}}{\partial q_h} = \frac{\partial \mathcal{L}(x,z)}{\partial a_h} \frac{\partial a_h}{\partial q_h} = \delta_h a_h / q_h \tag{20}$$

$$\frac{\partial \mathcal{L}}{\partial q_i} = \frac{\partial \mathcal{L}(x,z)}{\partial a_i} \frac{\partial a_i}{\partial q_i} = \delta_i x_i \tag{21}$$

For the loss function gradient with respect to position, one must be careful to sum Eq. 14 and Eq. 18 across each layer in which the h-th particle's position appears. For example, for unit h with $h \in H_l$ and $h' \in H_{l+1}$. The position gradients are

$$\frac{\partial \mathcal{L}}{\partial \vec{r}_h} = \frac{\partial \mathcal{L}(x,z)}{\partial a_h} \frac{\partial a_h}{\partial \vec{r}_h} + \sum_{h' \in H_{l+1}} \frac{\partial \mathcal{L}(x,z)}{\partial a_{h'}} \frac{\partial a_{h'}}{\partial \vec{r}_h} \tag{22}$$

which, for completeness, can be expanded by substitution of the above equations into each term

$$\frac{\partial \mathcal{L}(x,z)}{\partial a_h} \frac{\partial a_h}{\partial \vec{r}_h} = \delta_h \sum_{h'' \in H_{l-1}} b_{h''} \frac{\partial k(\vec{r}_h, \vec{r}_{h''})}{\partial \vec{r}_h} \tag{23}$$

$$\frac{\partial \mathcal{L}(x,z)}{\partial a_{h'}} \frac{\partial a_{h'}}{\partial \vec{r}_h} = -\delta_{h'} \sum_{h'' \in H_{l+1}} b_{h''} \frac{\partial k(\vec{r}_h, \vec{r}_{h''})}{\partial \vec{r}_h} \tag{24}$$

For the input layer, having no layer preceding it, Eq. 22 is reduced to

$$\frac{\partial \mathcal{L}}{\partial \vec{r}_i} = \sum_{h \in H_{l+1}} \frac{\partial \mathcal{L}(x,z)}{\partial a_h} \frac{\partial a_h}{\partial \vec{r}_i} \tag{25}$$

which can be computed similar to Eq. 24.

Similar equations are derived for the phase gradients:

$$\frac{\partial \mathcal{L}}{\partial \phi_h} = \frac{\partial \mathcal{L}(x,z)}{\partial a_h} \frac{\partial a_h}{\partial \phi_h} + \sum_{h' \in H_{l+1}} \frac{\partial \mathcal{L}(x,z)}{\partial a_{h'}} \frac{\partial a_{h'}}{\partial \phi_h} \tag{26}$$

$$\frac{\partial \mathcal{L}}{\partial \phi_i} = \sum_{h \in H_{l+1}} \frac{\partial \mathcal{L}(x,z)}{\partial a_h} \frac{\partial a_h}{\partial \phi_i} \tag{27}$$

2.3 Particle Network Implementation

The equations for the PN model appear somewhat daunting at first glance in comparison to the MLP model, but their implementation in code is more straightforward than might appear. We have implemented the PN model (for arbitrary number of units and layers) in our experimental neural network code, called Calrissian (available on GitHub []), which is written in Python and makes heavy use of the NumPy library for efficient vectorized computation.

2.3.1 Linear Scaling Memory and Computational Complexity

An interesting feature of PNs is that it does not explicitly require storing weight matrices in memory for each layer (or weight vectors per unit, depending on implementation) like a MLP does. In fact, one can recover MLP weights from PNs by setting

$$q_j k_{ij} \to w_{ij} \tag{28}$$

where w_{ij} is the corresponding MLP weight from 3.[3] In other words, one may interpret PNs

[3]It is not clear to use at this point that the reverse mapping from MLP weight to PN amplitude, phase, and position is always uniquely defined or computable.

as a decomposition of MLP weights into a product of amplitude and a position and phase dependent kernel function.

In PNs, one could explicitly construct a distance/relative phase matrix (or the full pairwise kernel matrix) and yield the same memory requirement of MLPs, but this is not strictly necessary. A more memory efficient approach for PNs is to implicitly build these matrices elementwise by computing k_{ij} on-the-fly as needed and subsequently discarding the value from memory when the matrix element is no longer needed.

With such an implementation, the total number of free parameters to be fit for a PN in $\mathbb{R}^3$ is given by

$$4N_I + 6N_K + \sum_{l=0}^{L} 6N_l \tag{29}$$

where N_I is the number of input layer units, N_K is the number of output layer units, and N_l is the number of units in the l-th hidden layer. N_I and N_K are fixed by the problem domain, which leaves N_l as the number which determines scaling of number of parameters. Thus, we see that the number of parameters (and the memory required to store them) scales linearly with respect to the number of hidden units for any l and independently of the number of parameters in any other layer.

For comparison, the total number of free parameters (weights and biases) to be fit in a MLP is given by

$$(N_I+1)N_0+(N_L+1)N_K+\sum_{l=1}^{L}(N_{l-1}+1)N_l \tag{30}$$

Note that the "plus one" appearing in each term is accounts for the bias parameter in each unit. Here we see that the number of parameters scales like $O(N_{l-1}N_l + N_l N_{l+1})$, which technically is linear with respect to N_l but with possibly large N_{l-1} or N_{l+1} prefactors. That is, increasing N_l results in an overall parameter increase dependent on both N_{l-1} and N_{l+1}. If we assume $N_{l-1} \sim N_l$, then we have that the number of parameters in MLPs scales as roughly quadratic with respect to the number of hidden units.

In other words, MLPs require an explicit parameter to modulate each inter-unit connection, whereas in PNs the inter-unit connection is modulated implicitly via spatial relationships of the units. The result is that PNs can model the same number of connections as MLPs—requiring the same overall computational complexity scaling—but with far fewer (linear scaling number of) parameters/memory.

Linear scaling memory could be a potentially useful feature of PNs in parallelization strategies requiring communication of parameters across distributed memory and/or between CPU and GPU memory. The latter is sometimes cited as a concerning bottleneck when using GPUs for parallelization, and having to transfer a linear scaling amount of memory as opposed to a quadratic scaling amount of memory could yield a performance boost. Or, one might possibly be able to squeeze larger/deeper PN networks on a GPU. Experiments will be needed to evaluate such a claim, of course, as well as to understand the balance of memory benefit and model predictive power.

PNs, however, require more floating point operations than MLPs per inter-unit connection. On the other hand, PNs might be able to exploit spatial sparsity to reduce computational complexity by invoking distance cutoffs, wherein beyond a certain threshold distance where the kernel is approximately zero, one ignores the inter-unit connection computation. Cutoffs can yield $O(N)$ computational scaling in N-body simulations provided the particles maintain a roughly homogeneous spatial distribution, which may not be likely in PNs. Alternatively, Fast Multipole Methods and/or domain decomposition could possibly be used to achieve $O(N_{l-1} \log N_l)$ complexity per input datum and/or scalable model parallelization, respectively. This is to be compared to the dense matrix-vector multiplication of MLPs, which is $O(N_{l-1}N_l)$ per input datum (or per forward/backward pass). Again, the realization of any of these claimed potential benefits will need experimental evidence as support.

2.3.2 Forward Pass

The forward pass of a PN proceeds quite similar to that of an MLP forward pass. To take advantage of linear scaling memory, though, we need to structure the algorithm in such a way that each element of the kernel matrix computed on-the-fly is available for each input instance. See Algorithm 1, which saves w_{ij} in memory only for the innermost loop over the N_D input instances of data. Of course, if the feed forward procedure were being called many times for a constant set of PN parameters, it might be useful to explicitly cache w_{ij} as a matrix rather than recompute for every forward pass. But, for PN learning, the parameters change with every update in the optimization algorithm, defeating the purpose of caching. Naturally, it is up to the developer to decide if caching is appropriate or not.

Algorithm 1 PN Forward Pass for input data set with N_D instances

```
procedure ForwardPass
    for $i \in$ input, $n \in N_D$ instances do
        $b_i^n = x_i^n$
    end for
    for $l \in H_l$ do
        $a = 0$
        for $i \in l-1,\ j \in l$ do
            $w_{ij} = q_j k_{ij}$
            for $n \in N_D$ do
                $a_j^n = a_j^n + w_{ij} b_i^n$
            end for
        end for
        $b = \theta_l(a)$
    end for
    return $b$
end procedure
```

2.3.3 Backpropagation

To compute the analytic gradient of a PN, one can apply the backpropagation technique as is done in MLPs. Like the forward pass, PN backpropagation can be written in such a way as to avoid explicit construction of matrices (Algorithm 2). Much like MLP backpropagation, a forward pass is first performed to compute the activations of each layer. Then, a backward pass is performed in which the gradient information of one layer is passed back to previous layers recursively. As with the forward pass, the loops can be structured so as to avoid explicit matrix construction.

Algorithm 2 PN Backpropagation for input data set with N_D instances

```
procedure Backpropagation
    ForwardPass to compute $b, \theta'(a)\ \forall\ l \in H_l$
    for $k \in K,\ n \in N_D$ do
        $\delta_k^n = y_k^n - z_k^n$
    end for
    $l = L$
    while $l >= 0$ do
        $\delta' = 0$
        for $i \in l-1,\ j \in l$ do
            $w_{ij} = q_j k_{ij}$
            for $n \in N$ do
                ${\delta'}_i^n = {\delta'}_i^n + \theta'(a_i) w_{ij} \delta_j^n$
                $\partial_q \mathcal{L}_j^l = \partial_q \mathcal{L}_j^l + k(\vec{r}_i, \vec{r}_j) \sum_n b_i^n \delta_j^n$
                $\Delta_{\vec{r}_j} = \sum_n \partial_{\vec{r}_j} k_{ij} b_i^n \delta_j^n$
                $\partial_{\vec{r}} \mathcal{L}_j^l = \partial_{\vec{r}} \mathcal{L}_j^l + \Delta_{\vec{r}_j}$
                $\partial_{\vec{r}} \mathcal{L}_i^l = \partial_{\vec{r}} \mathcal{L}_i^l - \Delta_{\vec{r}_j}$
                $\Delta_{\phi_j} = \sum_n \partial_{\phi_j} k_{ij} b_i^n \delta_j^n$
                $\partial_\phi \mathcal{L}_j^l = \partial_\phi \mathcal{L}_j^l + \Delta_{\phi_j}$
                $\partial_\phi \mathcal{L}_i^l = \partial_\phi \mathcal{L}_i^l - \Delta_{\phi_j}$
            end for
        end for
        $\delta = \delta'$
        $l = l - 1$
    end while
    for $j \in l_{input}$ do
        $\partial_q \mathcal{L}_j^{l_{input}} = \partial_q \mathcal{L}_j^{l_{input}} + \sum_n b_j^n \delta_j^n$
    end for
    return $\partial_q \mathcal{L}, \partial_{\vec{r}} \mathcal{L}$
end procedure
```

3 Experiment

Experiment

MNIST digits with no augmentation.

4 Discussion

This is a discussion.

Fewer parameters, less likely to overfit. Overfitting a concern with so many parameters. Cite AlexNet paper. Mention Circulant matrix paper. Limited to be square matrix; use cutoffs and/or zero filling to fake out rectangular matrix. (This is not an issue in PNs)

Connections in PNs a bit weaker for fitting than connections in MLPs. May require more connections to be comparable to MLPs. Tradeoff.

5 Conclusions

This is a conclusion.

6 Appendix: Modified Rprop algorithm

Describe modified Rprop algorithm used for fitting.