Response to the Reviewers of "The Geometry of Deep Networks: Power Diagram Subdivision"

We thank the reviewers for their careful reading, concrete suggestions, and interest in our manuscript. Responses to your Detailed Comments are given below. If our paper is accepted, the Python/TensorFlow code for all the experiments and the figures will be provided in a Github repository; this will clarify the computational details.

Reviewer 1. **[0]** Our power diagram (PD) framework exposes and advances the knowledge of the geometry of deep networks (DNs) in several ways. First, we demonstrate that MASO DN's convex regions are formed by an elegant subdivision process that we describe analytically in closed form. Second, our formulation exposes the roles played by the various parameters of the layers (the weights and biases) and the PDs' centroids and radii and opens the door to new computational geometric approaches to understanding and interpreting DNs. **[1]** We will clarify below (1) that $[A^{(\ell)}]_{k,r,.}$ represents the vector containing all the values of the last dimension. **[2]** We will clarify in the revised Introduction that x and $z^{(\ell)}$ represent either vectors or tensors depending on the context/layer and that boldface $\boldsymbol{z}^{(\ell)}$ and $\boldsymbol{x}$ represent the flattened versions of $z^{(\ell)}$ and x, respectively. **[3]** We will clarify in the text that "orthogonal" means $\langle [A]_{k,r,.}, [A]_{k',r',.}\rangle = 0, \forall r, r', k \neq k'$. **[4]** We will add below (18) the following "Remark: The centroid computation corresponds directly to the backpropagation algorithm (18) and thus can be computed precisely (up to roundoff error) and efficiently with same computational cost as a forward pass through the DN." **[5]** We will add just after the reference on line 216: "in which $\|\mu_{\boldsymbol{x}}^{(1\leftarrow L)} - \boldsymbol{x}\|$ is used as the unsupervised loss." **[6]** We will add the proof of Theorem 4 to the Supplementary Materials and augment it with additional insights on the polynomial and its use to characterize the input space partitioning. **[7]** Previous work [BB18a, BB18b] has not characterized a DN's partitioning nor studied its construction through depth. Instead, [BB18a, BB18b] focused on the affine mappings that are applied on each of the partition regions. We will clarify this in paragraph 3 of the Introduction. **[8]** We will add the definition for $\mathcal{V}([t]_1)$ in the proof of Theorem 1; it is simply shorthand to denote the region generated by unit $[t]_1$; similarly, b is shorthand that contains the elements of A and B as $b\{i,j\} = \|[A]_{1,i,.}\|^2 + 2[B]_{1,i} + \|[A]_{2,j,.}\|^2 + 2[B]_{2,j} + 2\langle [A]_{1,i,.}, [A]_{2,j,.}\rangle$. **[9]** We will add in the proof to Lemma 6 an explanation of each step. The second equality is derived by rewriting the inner product plus bias as a norm minus the remaining elements such that the equality holds. **[10]** We will add information regarding the theorem from [Joh60].

Reviewer 2. **[0]** We will add the following discussion around Fig. 3 and in the Supplementary Material regarding interpretability: In a Voronoi diagram, the centroids alone fully describe the partition regions and are guaranteed to lie inside. For a power diagram (PD), the interplay between the radii and centroids can make the centroids move out of their respective regions, complicating the visual interpretation. Interestingly, Fig. 3 suggests how, through depth, the partitioning moves from a near Voronoi partitioning, with centroids close to their associated regions (as measured by the small distances from the centroids to the input data points lying in each region) to a PD that relies heavily on the radii (as measured by the large distances from the centroids to the input data points lying in each region). This new understanding could open the door to novel DN constraints that provide more interpretable centroids (as described at the end of Sec. 3.2). **[1]** In the revised experiments section and Supplementary Materials, we will detail the training procedures (code will also be made publicly available). We will also make more explicit how training evolves the centroids and distances to the region boundaries (Figs. 3 and 4). **[2]** While the primary goal of the paper is to fully characterize the DN input space partitioning, we have provided some direction regarding how to use the results for (i) constraining the weights of the DN to impose specific PD region geometries (Appendix A.2); (ii) semi-supervised learning (end of Section 4.2); (iii) constraining the PD to be a VD for enhanced interpretability (end of Section 3.2); (iv) analysis of how the decision boundary curvature is constrained based on the DN topology and how it can be computed via a differentiable measure, enabling novel regularization techniques (Section 5.2). Unfortunately, due to the page limitation, there is no space to pursue any of these directions in greater detail. We plan to publish results in these directions in future papers.

Reviewer 3. **[0]** We will strive to make our notation and the connections more clear in the revised paper. One reason for the complications in the current notation is that we have striven to connect the standard computational geometry notation with the standard DN notation. For example, we believe that the region shape characterization (Appendix A.2) and decision boundary curvature (Section 5.2 and Appendix A.5) are more interpretable with the current notation, for the deep learning community. However, as you pointed out, some notation was not introduced nor sufficiently explained; we will add those in the revised paper (such as for (18)). **[1]** We will add a sentence setting up the notation and setting before (21) and correct ℓ to $\ell - 1$. **[2]** Indeed, the left-hand side should not have been squared; we will correct this. We will also add a paragraph in the Supplementary Material (due to the lack of space in the main text) to fully explain the distance and add a reference along with a short descriptive sentence in the main text. **[3]** The typo on line 333 will be corrected.