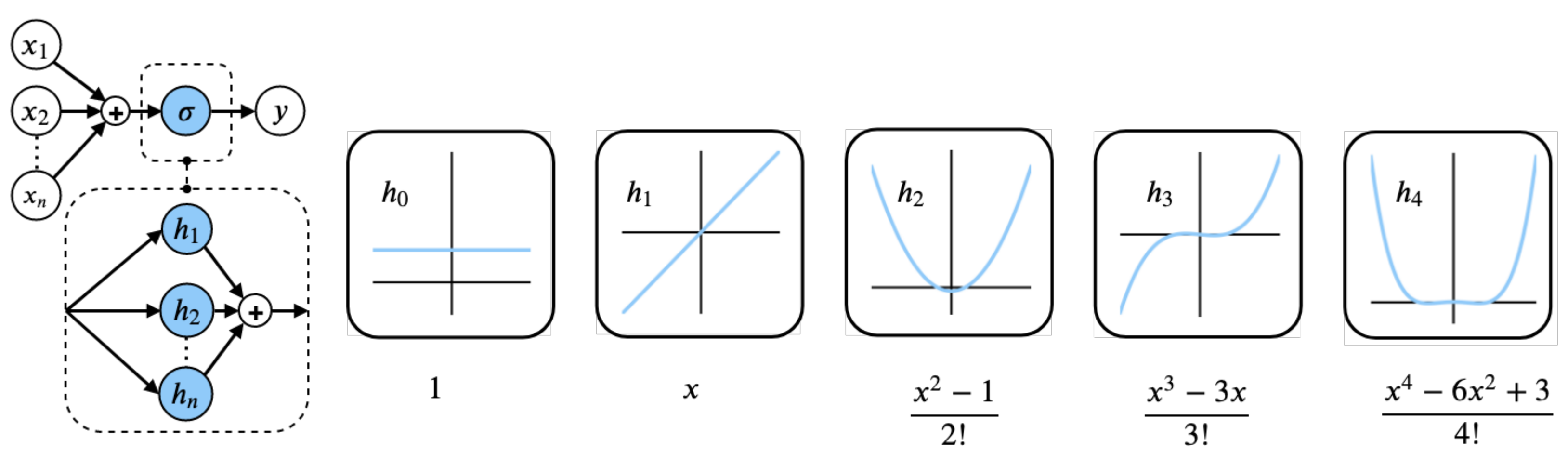
Generating Accurate Pseudo-labels in Semi-Supervised Learning and Avoiding Overconfident Predictions via Hermite Polynomial Activations

Vishnu Lokhande, Sathya N. Ravi, Songwong Tasneeyapant, Abhay Venkatesh, Vikas Singh
University of Wisconsin-Madison

Hermite Polynomials

Hermite Polynomials as activations

- The lower order terms in the Hermite polynomial series expansion of ReLU is used as an **activation function** with **the coefficients as trainable parameters**.
- Optionally, a SoftSign function is added to handle large numerical values attained by the polynomials.



Gap in the Literature

- Ge et al. [3] showed that for one hidden layer network, one could **avoid spurious local minima** by utilizing an orthogonal basis expansion for ReLUs.

Conclusion: Theoretical results not empirically investigated on regular computer vision architectures

- Nar et al. [4] showed that **smoother landscapes** enable the use of a larger range of step sizes in order for the gradient descent algorithm to converge.

Conclusion: Smoothness of the objective helps in convergence

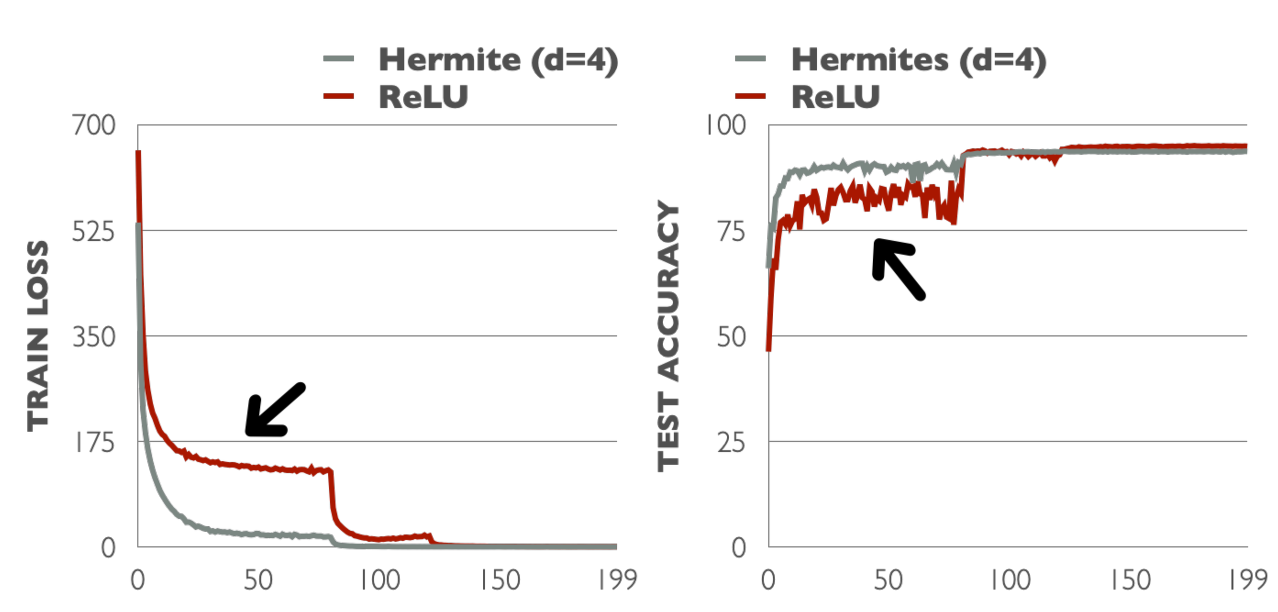
Hermite Polynomials in ResNet 152

- Test accuracy for hermite model converges in **less than half the number of epochs**.

Dataset CIFAR10	Number of Trainable Parameters	Best Test Accuracy	Epochs to reach 90% Test Accuracy
Hermite	58,145,574	95.48%	30
ReLU	58,144,842	94.5%	80

Hermite Polynomials in ResNet 18

- Hermites have **faster convergence in test accuracies over the initial epochs** but ReLU has the higher test accuracy at the end of training.



Hermite Polynomials make conscious classifications

- Unlike ReLU [5], when the test data is different from the training data, Hermite networks consciously make **(approximately) random predictions** unlike ReLU networks

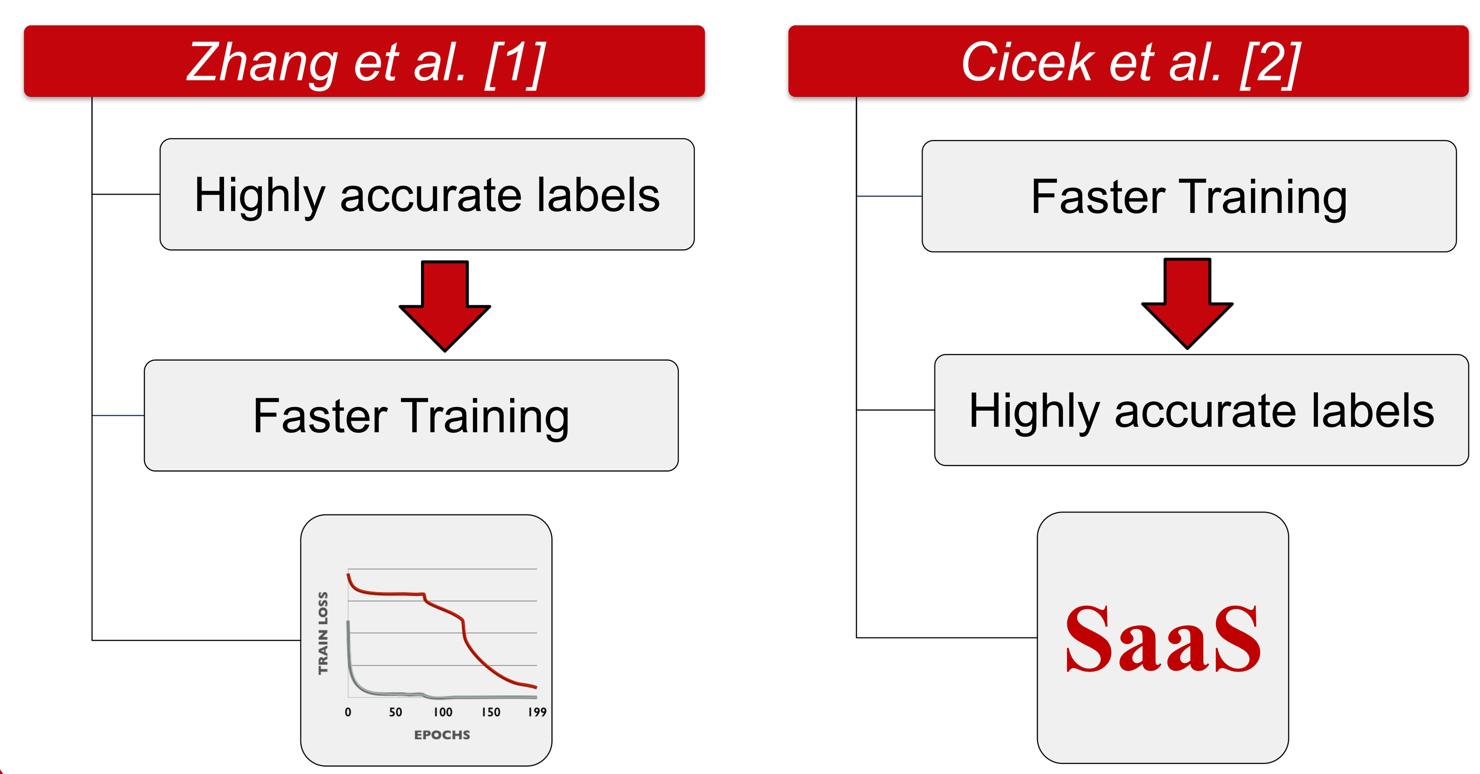
	Train Similar to Test	Train **Different From** Test
	Train Test	Train Test
ReLU	High Confidence Predictions	High Confidence Predictions
Hermite	High Confidence Predictions	**Low** Confidence Predictions

Theorem: When the training data is zero mean (mean normalized) then for a K class classification problem we have,

$$||x_{test}|| > \log\frac{1}{\epsilon} \implies \frac{1}{K} - \epsilon \le \frac{e^{f_k(x)}}{\sum_{l=1}^{K} f_l(x)} \le \frac{1}{K} + \epsilon$$

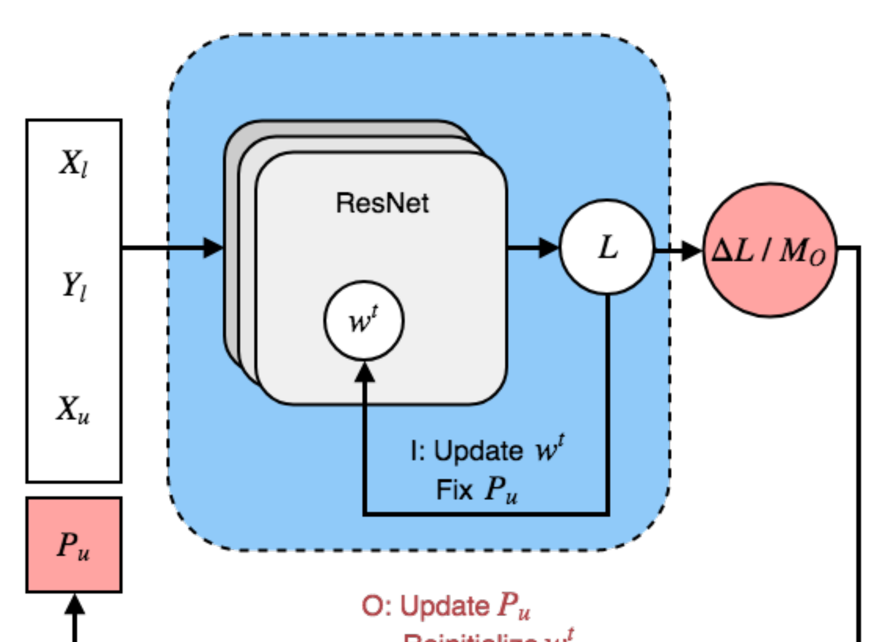
The Semi-supervised Learning Setup

- SaaS: Find labels with least training time



Speed as a Supervisor (SaaS)

- SaaS seeks to find a set of pseudolabels that maximizes the decrease in the loss function over a small number of epochs.

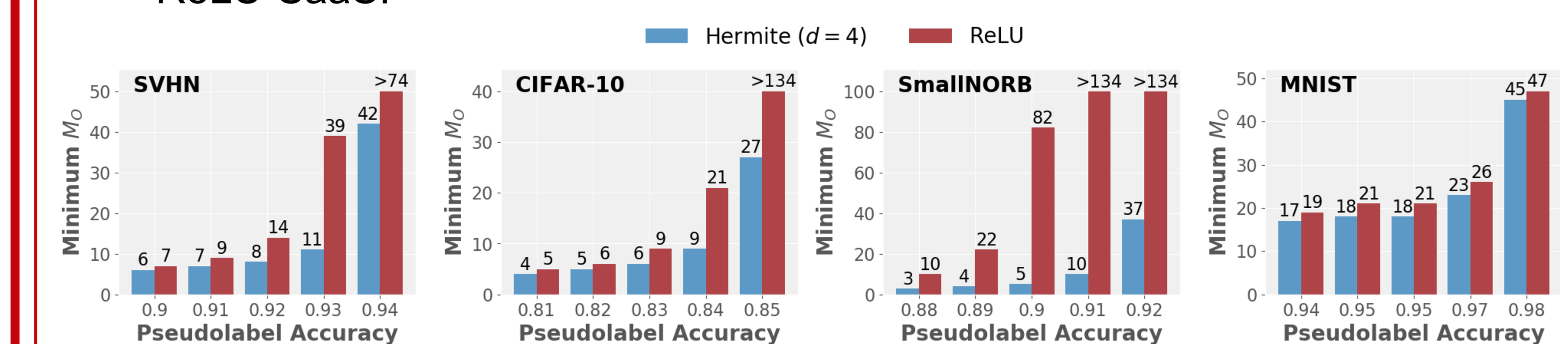


Conclusion: Smoothness of the objective helps in training SaaS faster

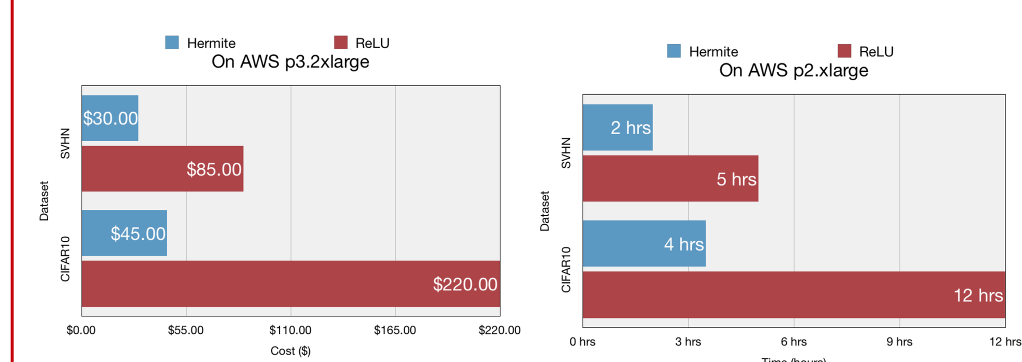
Computational Benefits

Hermite-SaaS trains faster

- Hermite-SaaS trains faster. The minimum number of epochs M_O to reach a given value of pseudolabel accuracy is **lower for Hermite-SaaS** than ReLU-SaaS.

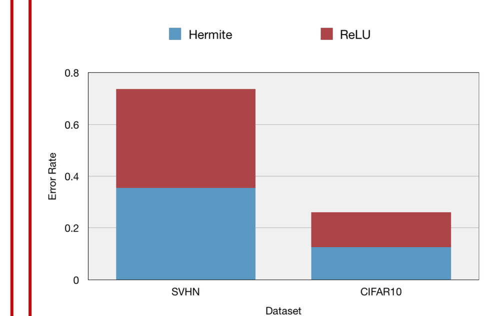


Hermite-SaaS saves time and money



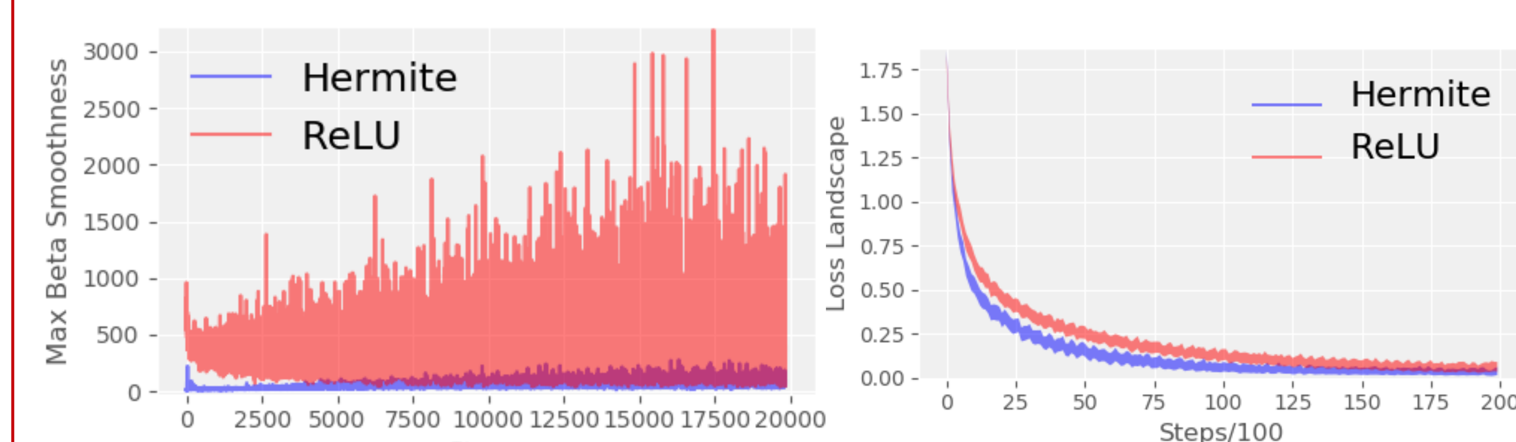
- **(Left)** Hermite-SaaS is saving $$ on AWS p3.2xlarge. **(Right)** Hermite-SaaS is saving compute time on AWS p2.xlarge.

Hermite-SaaS generalizes better



- Lower generalization error in the semi-supervised learning setup.

Hermite generates smoother landscape than ReLU



- Gradients are more stable on Hermite loss landscape: lower maximum beta smoothness.
- Magnitude of loss and its variation is lower for Hermites.

References

[1] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. arXiv preprint arXiv:1611.03530, 2016
[2] S. Cicek, A. Fawzi, and S. Soatto. Saas: Speed as a supervisor for semi-supervised learning. In The European Conference on Computer Vision (ECCV), September 2018
[3] R. Ge, J. D. Lee, and T. Ma. Learning one-hidden-layer neural networks with landscape design. arXiv:1711.00501, 2017
[4] Nar, K. and Sastry, S. Step size matters in deep learning. In Advances in Neural Information Processing Systems, 2018
[5] Hein, Matthias, Maksym Andriushchenko, and Julian Bitterwolf. "Why ReLU networks yield highconfidence predictions far away from the training data and how to mitigate the problem" *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.