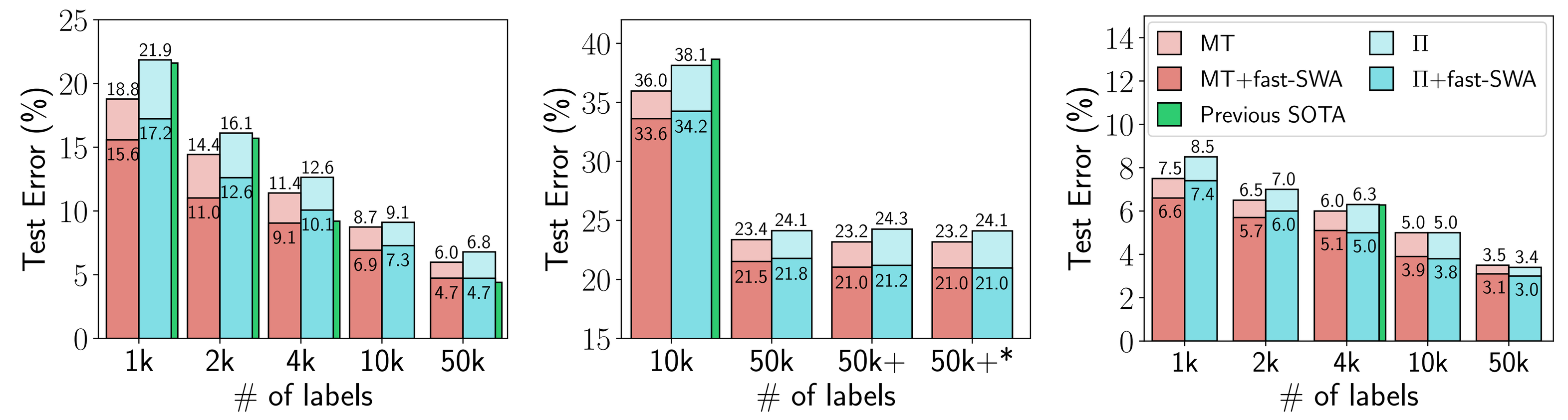
There are Many Consistent Explanations of Unlabeled Data: Why You Should Average

Ben Athiwaratkun, Marc Finzi, Pavel Izmailov, Andrew Gordon Wilson

Cornell University

Introduction

- We observe that for consistency-based methods, SGD does not converge to a single point but continues to explore many solutions with high distances apart.
- We propose to apply fast-SWA, a novel modification of Stochastic Weight Averaging (SWA), to the Π and Mean Teacher models. fast-SWA runs SGD with a cyclical learning rate schedule and averages weights of multiple SGD iterates within each cycle.
- Applying weight averaging to the Π and Mean Teacher models we improve the best reported results on multiple consequential benchmarks.



(a) C10, CNN (b) C100, CNN (c) C10, Shake-Shake

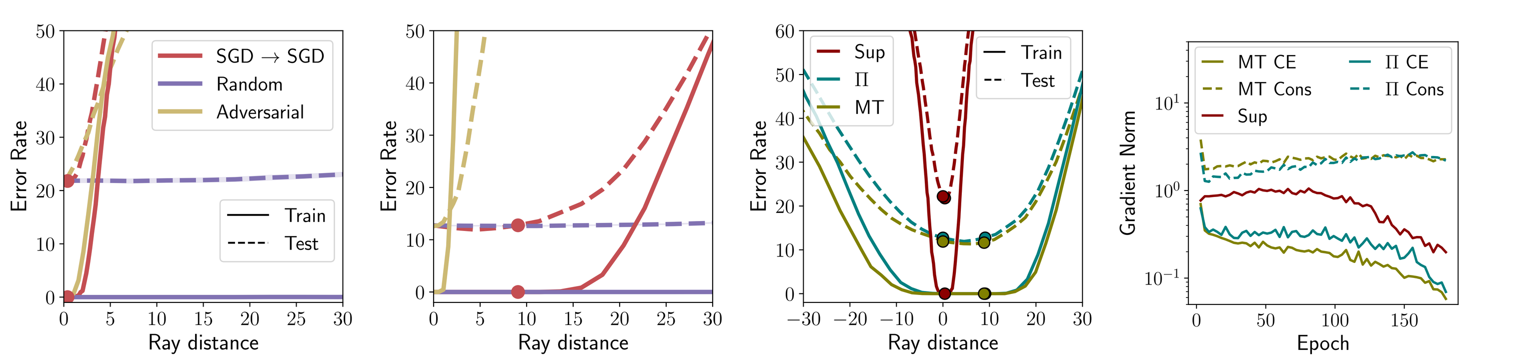
Consistency Enforcing methods

Consistency methods for SSL penalize change in network predictions with respect to input perturbations $x \to x'$ like random translations and horizontal flips with an additional loss term that can be computed on unlabeled data.

$$\sum_{(x,y)\in\mathcal{D}_L} \ell_{\mathrm{CE}}(x,y) + \lambda \sum_{x\in\mathcal{D}_L\cup\mathcal{D}_U} \|f(x) - f(x')\|^2$$

- For small additive normal perturbations, $x' = x + \epsilon z$, $z \sim \mathcal{N}(0, I)$, we show that the consistency loss $\hat{Q} = \lim_{\epsilon\to 0} \frac{1}{\epsilon^2}\|f(x) - f(x')\|^2$ is an unbiased estimator for the norm of the Jacobian of the network:

$$\mathbb{E}[\hat{Q}] = \mathbb{E}_x[\|J_x\|_F^2] \quad \text{and} \quad \mathrm{Var}[\hat{Q}] = \mathrm{Var}[\|J_x\|_F^2] + 2\mathbb{E}[\|J_x^T J_x\|_F^2].$$

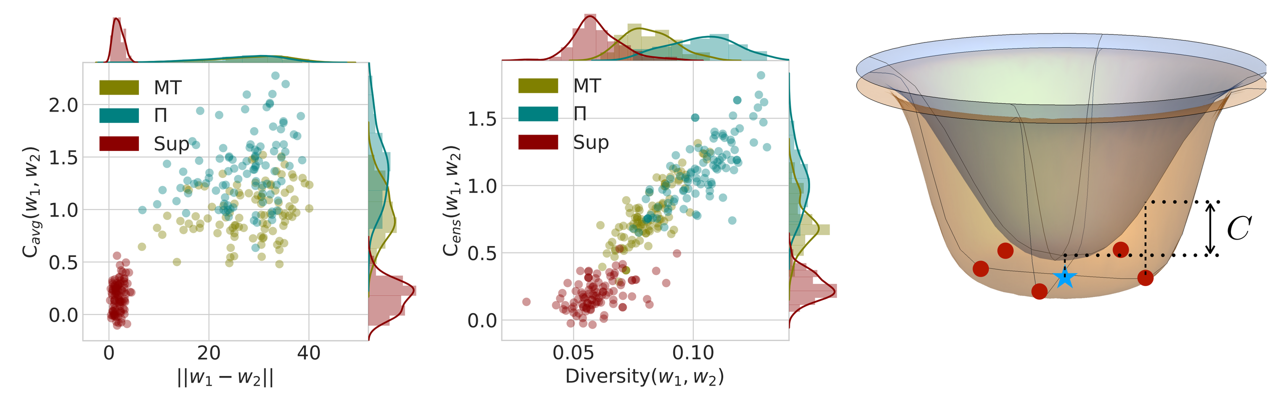


(a) (b) (c) (d)

- As measured by test error along rays from the solution parameters, we find that the consistency enforcing methods, Π and Mean Teacher, find minima which are less sharp than supervised only solutions.
- Optimizing the consistency loss, SGD continues to explore a diverse set of candidate models late into training, both as measured by distance and the fraction of different predictions on unseen data.

Ensembling and Weight Averaging

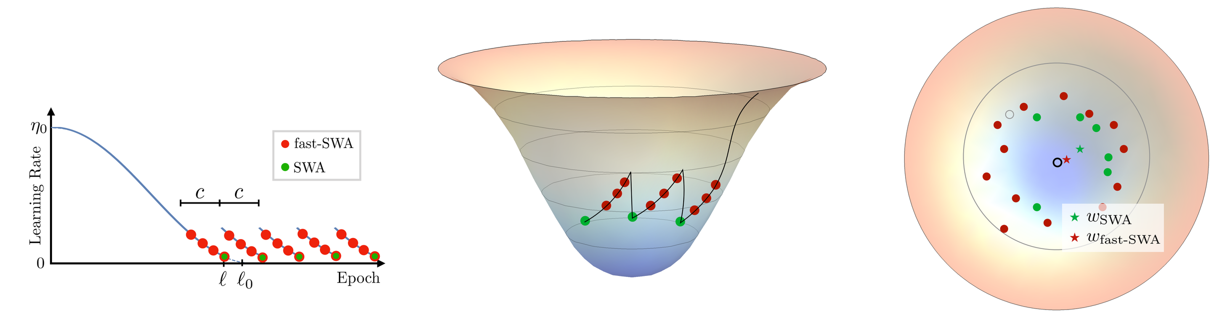
- This additional diversity as a result of the consistency loss can be converted into substantially greater performance through the ensembling predictions and averaging weights of the networks at different epochs in the training procedure.
- The improvement is much larger for the Π and Mean Teacher models compared to supervised training.
- Averaging the weights instead of predictions yields comparable performance, but major computational benefits for inference.



Left: Scatter plot of the decrease in error C_{avg} for weight averaging versus distance. **Middle**: Scatter plot of the decrease in error C_{ens} for prediction ensembling versus diversity. **Right**: Train error surface (orange) and Test error surface (blue). The SGD solutions (red dots) around a locally flat minimum are far apart due to the flatness of the train surface which leads to large error reduction of the SWA solution (blue dot).

Semi-Supervised Learning with fast-SWA

- For the first $\ell \leq \ell_0$ epochs the network is pre-trained using the cosine annealing schedule where the learning rate at epoch i is set equal to After ℓ epochs, we use a cyclical schedule, repeating the learning rates from epochs $[\ell - c, \ell]$, where c is the cycle length.
- SWA collects the networks corresponding to the minimum values of the learning rate and averages their weights. The model with the averaged weights w_{SWA} is then used to make predictions.



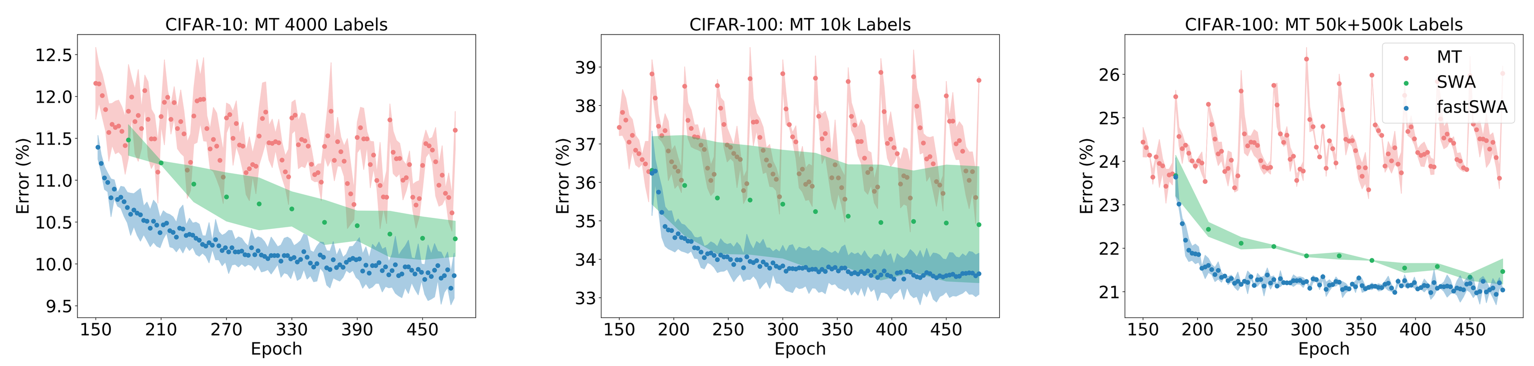
Left: Cyclical cosine learning rate schedule and SWA and fast-SWA averaging strategies. **Middle**: Illustration of the solutions explored by the cyclical cosine annealing schedule on an error surface. **Right**: Illustration of SWA and fast-SWA averaging strategies. fast-SWA averages more points but the errors of the averaged points, as indicated by the heat color, are higher.

- fast-SWA is a novel modification of SWA that uses longer learning rate cycles and averges weights more than once per cycle.
- We propose to apply SWA to the student network both for the Π and Mean Teacher models. Note that the SWA weights do not interfere with training.

Semi-Supervised Results

Table 1: Test errors against current state-of-the-art semi-supervised results.

Dataset	CIFAR-10			CIFAR-100		
No. of Images No. of Labels	50k 1k	50k 2k	50k 4k	50k 10k	50k+500k 50k	50k+237k* 50k
Previous Best CNN	18.41[4]	13.64[4]	9.22[2]	38.65[3]	23.62[3]	23.79[3]
Ours CNN	15.58	11.02	9.05	33.62	21.04	20.98
Previous Best†			6.28[1]			
Ours†	6.6	5.7	5.0‡	28.0	19.3	17.7



(a) C10, 4k labels (b) C100, 10k labels (c) C100, 50k labels

Paper and code

(a) Paper (b) Code