Learned optimizers that outperform SGD on wallclock and test loss

Google AI

Luke Metz, Niru Maheswaranathan, Jeremy Nixon, C. Daniel Freeman, Jascha Sohl-Dickstein

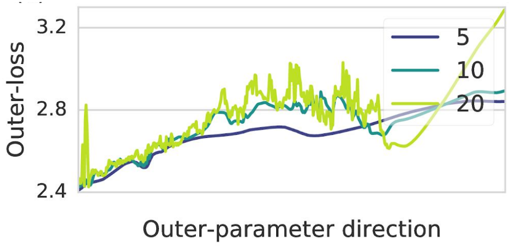
Existing optimizers are **hand designed.** Can we do better with **learning**?

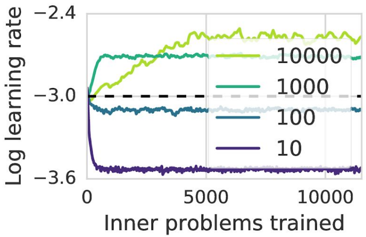
One popular strategy for training such optimizers is to leverage gradients and **truncated backpropagation through time.**

These methods, however, are notoriously **unstable**!

Careful choice of step length is required:

- Long truncations: **exploding gradients**
- Short truncations: **biased gradients**



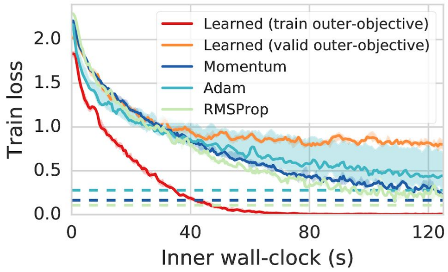


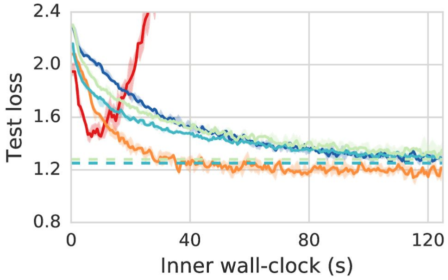
We use **variational optimization** to "smooth" the loss surface by convolving it with a Gaussian.

$$\mathcal{L}(\theta) = \mathbb{E}_{\tilde{\theta}\sim\mathcal{N}(\theta,\sigma^2 I)}\left[L\left(\tilde{\theta}\right)\right]$$

To optimize this objective, we combine **multiple gradient estimators** with difference variances.

We train **simple** MLP-based learned optimizers that are **faster in wallclock time** and **generalize better** than existing hand-designed methods.





Define two gradient estimators:

- **reparameterization trick**
- **evolutionary strategies**

Combine them!

$$g_{\text{rp}} = \frac{1}{S}\sum_{s} \nabla_\theta L\left(\theta + \sigma n_s\right), \qquad n_s \sim N\left(0, I\right)$$

$$g_{\text{es}} = \frac{1}{S}\sum_{s} L\left(\tilde{\theta}_s\right)\nabla_\theta\left[log\left(N\left(\tilde{\theta}_s;\theta,\sigma^2 I\right)\right)\right], \qquad \tilde{\theta}_s \sim N\left(\theta, \sigma^2 I\right)$$