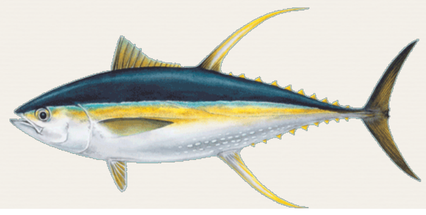
YellowFin and the Art of Momentum Tuning

Jian Zhang[1], Ioannis Mitliagkas[2]
[1]Stanford University, [2]MILA, University of Montreal



Adaptive Optimization

Hyperparameter tuning is a big cost of deep learning.

Momentum: a key hyperparameter to SGD and variants.

Adaptive methods, e.g. Adam[1], don't tune momentum.

YellowFin optimizer

[1.[Kingma et. al 15]

- Based on the robustness properties of momentum.
- Auto-tuning of momentum and learning rate in SGD.
- Closed-loop momentum control for async. training.

Robustness of Momentum

Momentum operation

- SGD step: $x_{t+1} = x_t - \alpha \nabla f(x_t) + \mu(x_t - x_{t-1})$.
- In a 1-D case, matrix form with $\nabla f(x_t) = h(x_t)(x_t - x^\star)$:

$$\begin{pmatrix} x_{t+1} - x^* \\ x_t - x^* \end{pmatrix} = \begin{bmatrix} 1-\alpha h(x_t)+\mu & -\mu \\ 1 & 0 \end{bmatrix} \begin{pmatrix} x_t - x^* \\ x_{t-1} - x^* \end{pmatrix} \triangleq A_t \begin{pmatrix} x_t - x^* \\ x_{t-1} - x^* \end{pmatrix}$$

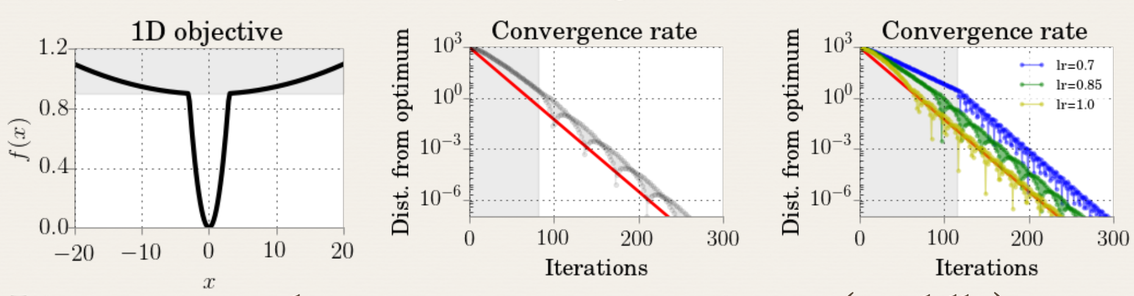
h is curvature in quadratics.

Robust region

$$\frac{(1-\sqrt{\mu})^2}{h_{\max}} \le \alpha \le \frac{(1+\sqrt{\mu})^2}{h_{\min}}, \text{ given } h_{\min} \le h(x_t) \le h_{\max}.$$

Spectral radius $\rho(A_t) = \sqrt{\mu}$ → linear convergence[2]

[2. Not guarrenteed for non-quadratics



- Linear rate robust to curvature variance (middle).
- Linear rate robust to a range of learning rates (left).

YellowFin

Noisy quadratic model

[3.[Schaul et. al 13]

- Model stochastic setting with gradient variance C.
- Local quadratic approximation[3]: 1-D case.

In robust region, distance to optimum is approx. by

$$\mathbb{E}(x_t - x^*)^2 \approx \mu^t (x_0 - x^*)^2 + (1-\mu^t)\frac{\alpha^2 C}{1-\mu}$$

Greedy tuning strategy

Principle I: Stay in the robust region

Principle II: Minimize $\mathbb{E}(x_0 - x^\star)^2$ after one step ($t=1$)

- Solve learning rate and momentum in closed-form sol.

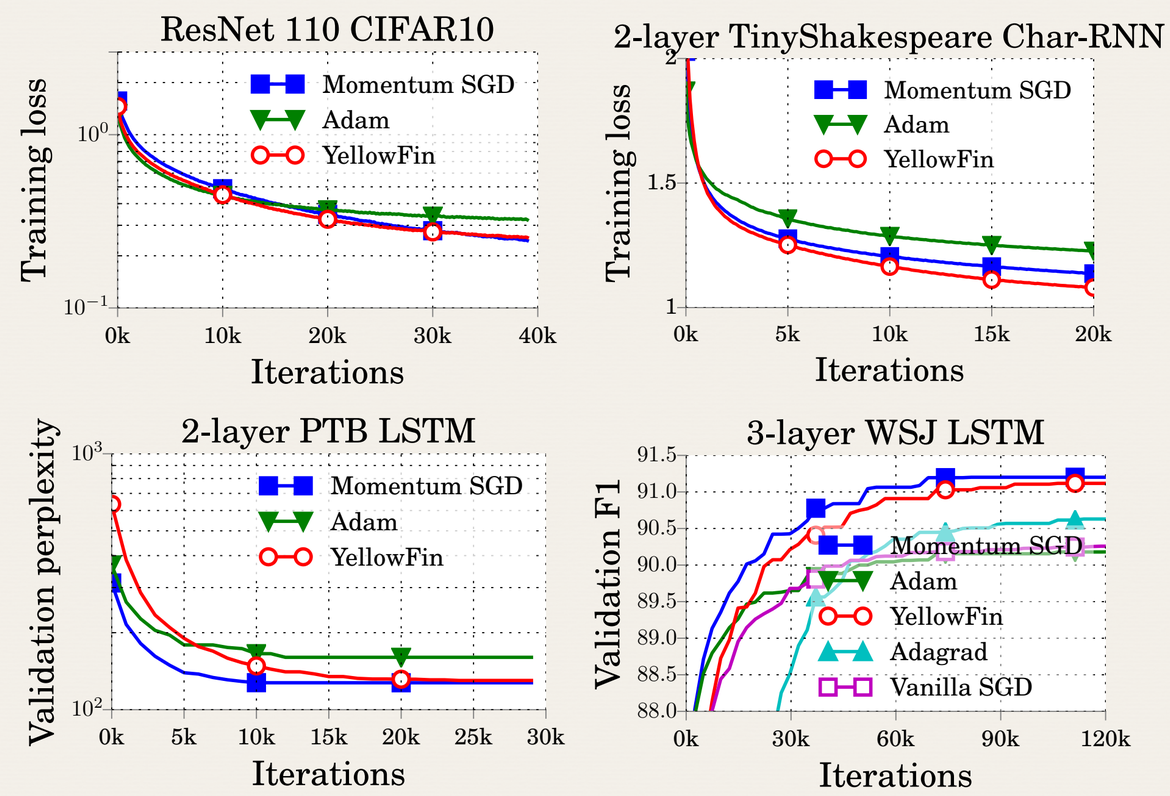
Experiments

ResNet and LSTM

YellowFin runs with no tuning.

Adam, mom. SGD etc. are tuned on learning rate grids.

YellowFin can outperform tuned SoA on train/val. metrics.



Facebook Convolutional Seq-to-seq model

IWSLT 2014 German-English translation

YellowFin outperforms the hand-tuned default optimizer

	Val. loss	Val. BLEU@4
Default Nesterov Momentum	2.86	30.75
YellowFin	2.75	31.59

Github for PyTorch: https://github.com/JianGoForIt/YellowFin_Pytorch

Github for TensorFlow: https://github.com/JianGoForIt/YellowFin

"Send us your bug reports!"

Extension: Closed-loop YellowFin

Async. distributed training: fast, no sync. barrier.

However, Asynchrony induces additional $\bar{\mu}$ to μ. [Mitliagkas et. al 16]

Can we auto-match total momentum μ_T to YF-tuned $\mu^\star$?

Closed-loop momentum control

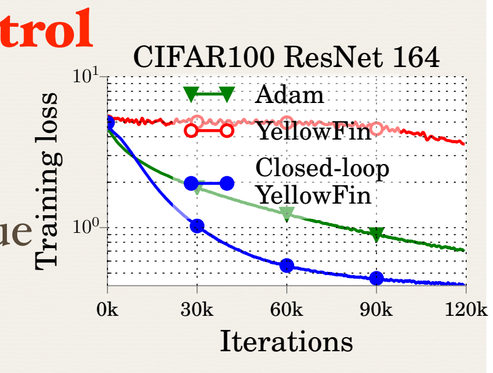
Async. induced $\bar{\mu}$

Algorithmic μ → Effective total value $\mu_T = \mu + \bar{\mu}$

Feedback control

YF target value

$$\mu \leftarrow \mu + \gamma \cdot (\mu^* - \mu_T) \longleftarrow \mu^\star$$



Closed-loop mechanism improves YellowFin in async..