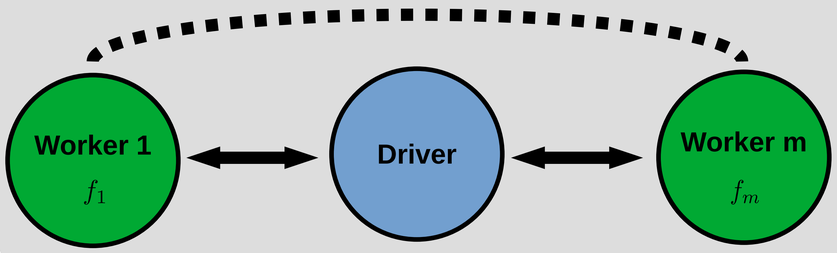
DINO: Distributed Newton-Type Optimization Method

Rixon Crane, Fred Roosta
School of Mathematics and Physics, University of Queensland, Australia.

Centralized Environment



Minimize

$$f(\mathbf{w}) = \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{w})$$

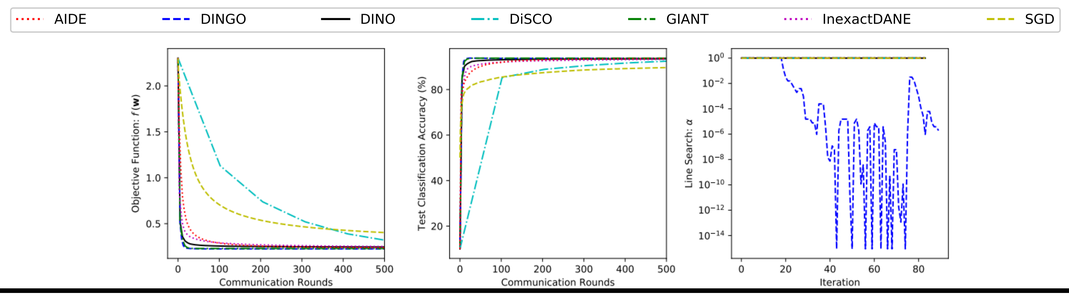
Examples

GPT-2 **~1.5B Parameters** → $\mathbf{w}_{t,i}$

~4M Searches Per Minute →

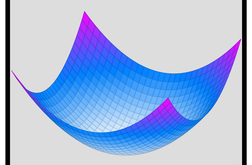
Simple Assumptions

- Lipschitz continuous gradients, gives global sub-linear convergence;
- Plus Polyak-Lojasiewicz (PL) inequality, gives linear convergence.

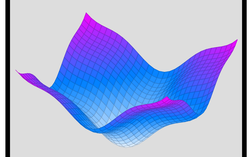


Method	Problem Class	Any Function Form	Any Data Distribution	Simple Sub-Problems.	Optimizes Function Value	Performance From 1 Node to AWS	
DINO	Non-Convex	✔	✔	✔	✔	+300%	
DINGO	Invex	✔	✔	✔	✘	+300%	
GIANT	Strongly Convex	✘	✘	✔	✔	+350%	
DiSCO	Strongly Convex	✔	✔	✔	✔	+171%	
InexactDANE	Non-Convex	✔	✔	✘	✔	+128%	
AIDE	Non-Convex	✔	✔	✘	✔	+127%	**SGD:** -32%

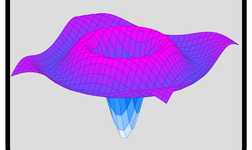
Convex



Invex



Non-Convex



Update Direction

Driver → **Workers**: Broadcast weights: $\mathbf{w}_t$; Reduce gradients: $\nabla f_i(\mathbf{w}_t)$

Driver → **Workers**: Broadcast full gradient: $\mathbf{g}_t = \nabla f(\mathbf{w}_t) = \frac{1}{m}\sum_{i=1}^{m} \nabla f_i(\mathbf{w}_t)$

Driver:

Compute:

$$\mathbf{v}_{t,i}^{(1)} = \tilde{\mathbf{H}}_{t,i}^{\dagger}\tilde{\mathbf{g}}_t \qquad \tilde{\mathbf{g}}_t = \begin{pmatrix}\mathbf{g}_t \\ \mathbf{0}\end{pmatrix} \qquad \tilde{\mathbf{H}}_{t,i} = \begin{pmatrix}\nabla^2 f_i(\mathbf{w}_t) \\ \phi\mathbf{I}\end{pmatrix}$$

If $\langle \mathbf{v}_{t,i}^{(1)}, \mathbf{g}_t\rangle \geq \theta\|\mathbf{g}_t\|^2$ then reduce: $\mathbf{p}_{t,i} = -\mathbf{v}_{t,i}^{(1)}$.

Else, compute $\mathbf{v}_{t,i}^{(2)} = (\tilde{\mathbf{H}}_{t,i}^T\tilde{\mathbf{H}}_{t,i})^{-1}\mathbf{g}_t$ then reduce:

$$\mathbf{p}_{t,i} = -\mathbf{v}_{t,i}^{(1)} - \lambda_{t,i}\mathbf{v}_{t,i}^{(2)}, \qquad \lambda_{t,i} = \frac{-\langle \mathbf{v}_{t,i}^{(1)}, \mathbf{g}_t\rangle + \theta\|\mathbf{g}_t\|^2}{\langle \mathbf{v}_{t,i}^{(2)}, \mathbf{g}_t\rangle}.$$

Driver: Update direction: $\mathbf{p}_t = \frac{1}{m}\sum_{i=1}^{m} \mathbf{p}_{t,i}$.

Step Size (Example)

Driver → **Workers**: Broadcast update direction: $\mathbf{p}_t$

Driver: Reduce: $\{f(\mathbf{w}_t + \omega^j\mathbf{p}_t)\}_{j=0}^{k}$.

Driver: Choose $\alpha_t = \omega^j$ for largest j such that:

$$f(\mathbf{w}_t + \alpha_t\mathbf{p}_t) \leq f(\mathbf{w}_t) + \alpha_t\rho\langle \mathbf{p}_t, \mathbf{g}_t\rangle.$$

Let $\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha_t\mathbf{p}_t$.