An Improved Empirical Fisher Approximation for Natural Gradient Descent

Xiaodong Wu[1*], Wenyi Yu[2*], Chao Zhang[2], Philip Woodland[1]

[1]Dept. of Engineering, University of Cambridge
[2]Dept. of Electronic Engineering, Tsinghua University
[*]Equal Contribution
Email: **xw338@cam.ac.uk**

Introduction

- Natural Gradient Descent (NGD) enjoys improved convergence
  - Exact Fisher matrix is too large to store for large models $\mathbf{F} \in \mathbb{R}^{P \times P}$
  - Preconditioned update $\mathbf{F}^{-1} \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})$ is impossible to compute for large models

- Empirical Fisher (EF) is a commonly used approximation for NGD

$$\tilde{\mathbf{F}} := \sum_n \left[\nabla_{\boldsymbol{\theta}} \log p_n(y_n) \nabla_{\boldsymbol{\theta}} \log p_n(y_n)^\top \right] = \nabla_{\boldsymbol{\theta}} \boldsymbol{l}^\top \nabla_{\boldsymbol{\theta}} \boldsymbol{l}$$

  - $\nabla_{\boldsymbol{\theta}} \boldsymbol{l} \in \mathbb{R}^{N \times P}$ is the empirical per-sample gradient, can be collected during back-propagation.
    - Easy to implement.
  - EF is not theoretically well-supported.
    - Approximation quality is limited.

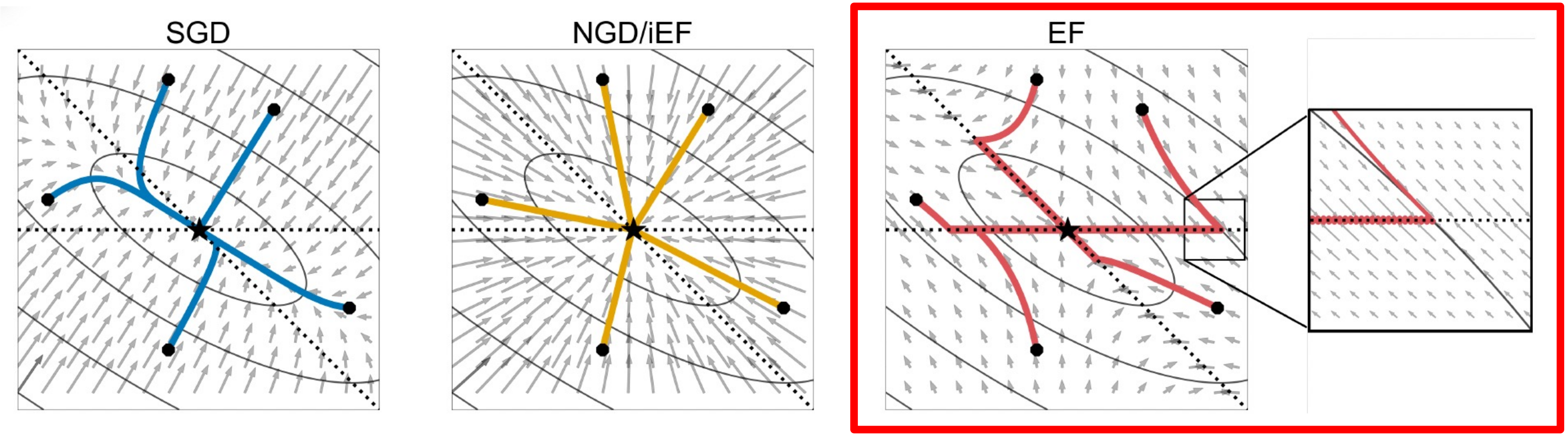
Inversely-Scaled Projection of Empirical Fisher (EF)

- EF update enforces Equal Per-sample Loss Reduction

$$\Delta \boldsymbol{l}_{\mathrm{EF}} = -\nabla_{\boldsymbol{\theta}} \boldsymbol{l} \Delta \boldsymbol{\theta}_{\mathrm{EF}} = -\eta \nabla_{\boldsymbol{\theta}} \boldsymbol{l} \nabla_{\boldsymbol{\theta}} \boldsymbol{l}^{\top} (\nabla_{\boldsymbol{\theta}} \boldsymbol{l} \nabla_{\boldsymbol{\theta}} \boldsymbol{l}^{\top} + \lambda \mathbf{I})^{-1} \mathbf{1} \approx -\eta \mathbf{1}$$

- Better trained samples get significantly more updated

Visualisation on Least-Squares Toy Problem



- Distorted Training Trajectory
- When one sample is nearly converged, the update norm becomes larger (inversely-scaled)

Improved Empirical Fisher (iEF)

- Induced Per-sample Loss Reduction: convergence-level aware for every sample

$$\Delta(l_n)_{\text{iEF}} = \nabla_{\boldsymbol{\theta}} l_n^\top \Delta\boldsymbol{\theta}_{\text{iEF}} \approx -\eta \, \|\nabla_{\boldsymbol{z}_n} l_n\|_2^2$$

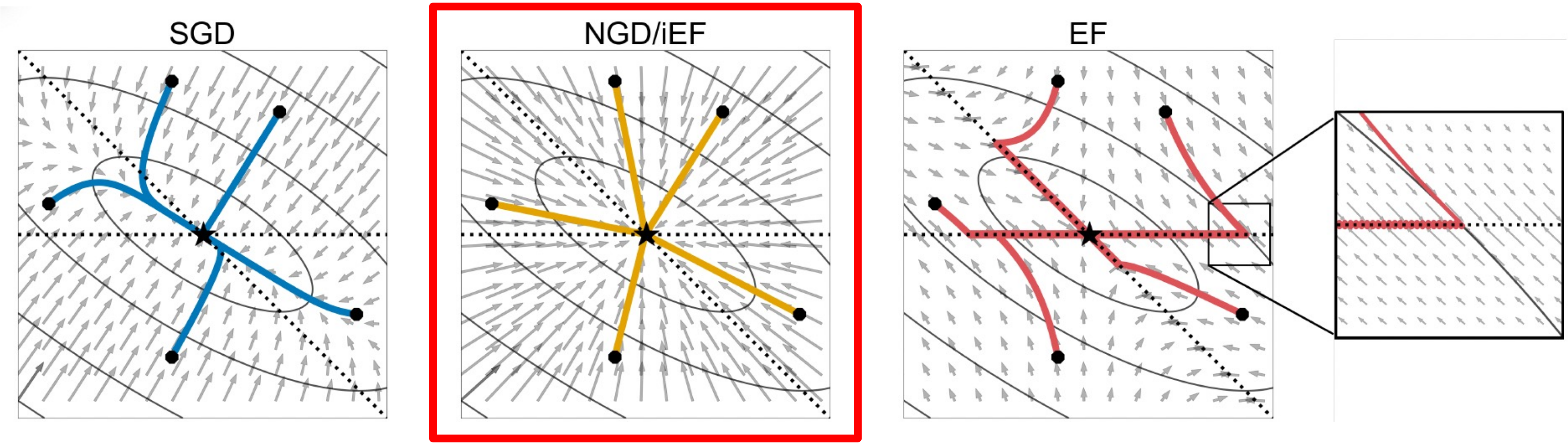
  - $\|\nabla_{\boldsymbol{z}_n} l_n\|_2$ is the logits-level gradient norm

- Inspiration : iEF Approximates per-sample loss reduction of Gauss-Newton algorithm

$$\Delta(l_n)_{\text{GN}} \approx -\eta \, \|\nabla_{\boldsymbol{z}_n} l_n\|_2^2$$

  - Gauss-Newton algorithm is a Generalised NGD method.

Visualisation on Least-Squares Toy Problem



- iEF Adapted to the Curvature of the loss landscape
- No more Distorted Training Trajectory

Experiments

- We compare ***EXACT*** EF (empirical Fisher), iEF (improved EF) and SF (sampled Fisher) for practical and up-to-date optimization setups.
- We consider Parameter-Efficient Finetuning setup for pretrained Transformer models for GLUE (textual classification) and CIFAR (image classification) tasks.
    - Optimisation Performance
    - Approximation Quality to NG Updates

Optimisation Performance

- Overall Test Performance:

	AdamW	Adafactor	SGD	EF	SF	iEF
GLUE + T5 + Prompt Tuning	-	77.1	67.4	48.1	69.7	**79.3**
GLUE + T5 + LoRA	**80.1**	-	77.3	63.1	76.5	79.3
CIFAR100 + ViT + LoRA	93.9	-	91.3	31.0	92.8	**94.3**

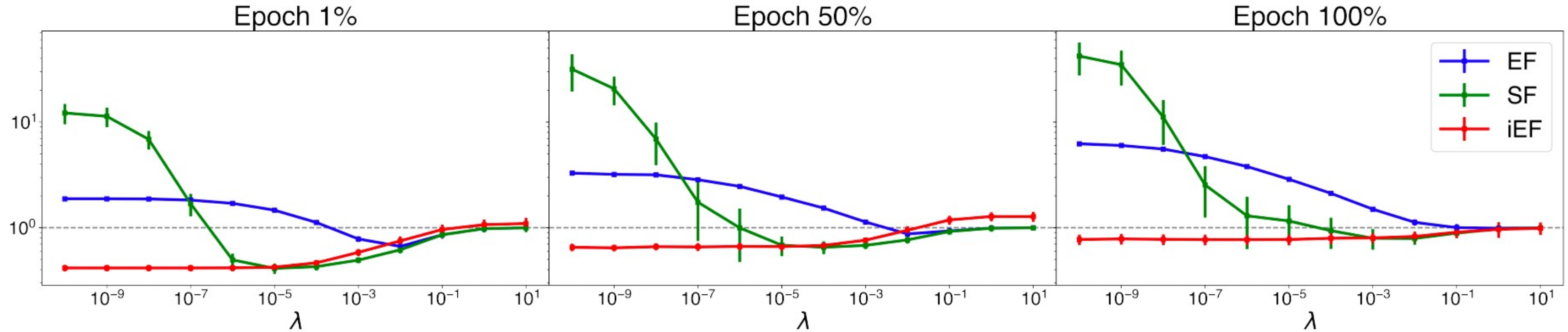
  - iEF achieves comparable performance with well-tuned baseline optimisers
  - iEF consistently outperforms SGD, EF and SF optimisers
  - EF consistently suffers from unstable training and is unable to train a decent model

Evaluation Framework for Approximation Quality

- Traditional Methods usually requires Computation of the Exact Fisher matrix $\mathbf{F} \in \mathbb{R}^{P \times P}$
  - Too Expensive!
- Our Efficient Evaluation Framework for Large Scale Neural Networks
  - This Framework requires only a matrix-vector-product with Fisher matrix
    - Efficient to compute
    - Theoretically well-supported

Approximation Quality *w.r.t.* Time and Damping

- Approximation quality of EF/SF/iEF updates for different damping values (x-axis is the damping value) (y-axis is the relative approximation indicator↓) at different training stages



  - EF and SF updates are sensitive to damping values
  - Optimal damping values for EF and SF vary greatly across training stages (and tasks)
  - iEF has comparable performance to optimally damped SF updates
  - iEF is robust to damping values (small damping would suffice)

Conclusions

- Identify a crucial flaw of EF: ***the inversely-scaled projection*** issue.
- We proposed the improved EF (iEF), which is shown to be robust and achieve better quality.
- We proposed an efficient evaluation framework for the approximation quality to NG update.

Thank you

Contact Email: xw338@cam.ac.uk

Preliminaries

- Natural Gradient Descent (NGD) enjoys improved convergence

$$\mathbf{F} := \sum_n \sum_c p_n(c) \left[\nabla_{\boldsymbol{\theta}} \log p_n(c) \nabla_{\boldsymbol{\theta}} \log p_n(c)^\top \right]$$

  - Preconditioned update $\mathbf{F}^{-1} \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})$ is impossible to compute for large scaled neural networks
- Monte-Carlo Sampled Fisher (SF) is a well-supported approximation method

$$\hat{\mathbf{F}}(K) = \frac{1}{K} \sum_{n=1}^{N} \sum_{k=1}^{K} [\nabla_{\boldsymbol{\theta}} \log p_n(\hat{y}_n^{(k)}) \nabla_{\boldsymbol{\theta}} \log p_n(\hat{y}_n^{(k)})^\top]$$

  - $\hat{y}_n^{(k)} \sim p_{\boldsymbol{\theta}}(y|\boldsymbol{x}_n)$ Too expensive, Hard to implement, Even for $K = 1$.
- Empirical Fisher (EF) is a commonly used approximation for NGD

$$\tilde{\mathbf{F}} := \sum_n \left[\nabla_{\boldsymbol{\theta}} \log p_n(y_n) \nabla_{\boldsymbol{\theta}} \log p_n(y_n)^\top \right] = \nabla_{\boldsymbol{\theta}} \boldsymbol{l}^\top \nabla_{\boldsymbol{\theta}} \boldsymbol{l}$$

  - $\nabla_{\boldsymbol{\theta}} \boldsymbol{l}$ is the empirical per-sample gradient. Fast, Easy to implement. Poor Quality.