Ellipsoidal Trust Region Methods for Neural Nets

L. Adolphs*, J. Kohler*, A. Lucchi, T. Hofmann

Institute for Machine Learning, ETH Zürich

Introduction

We consider finite-sum optimization problems of the form

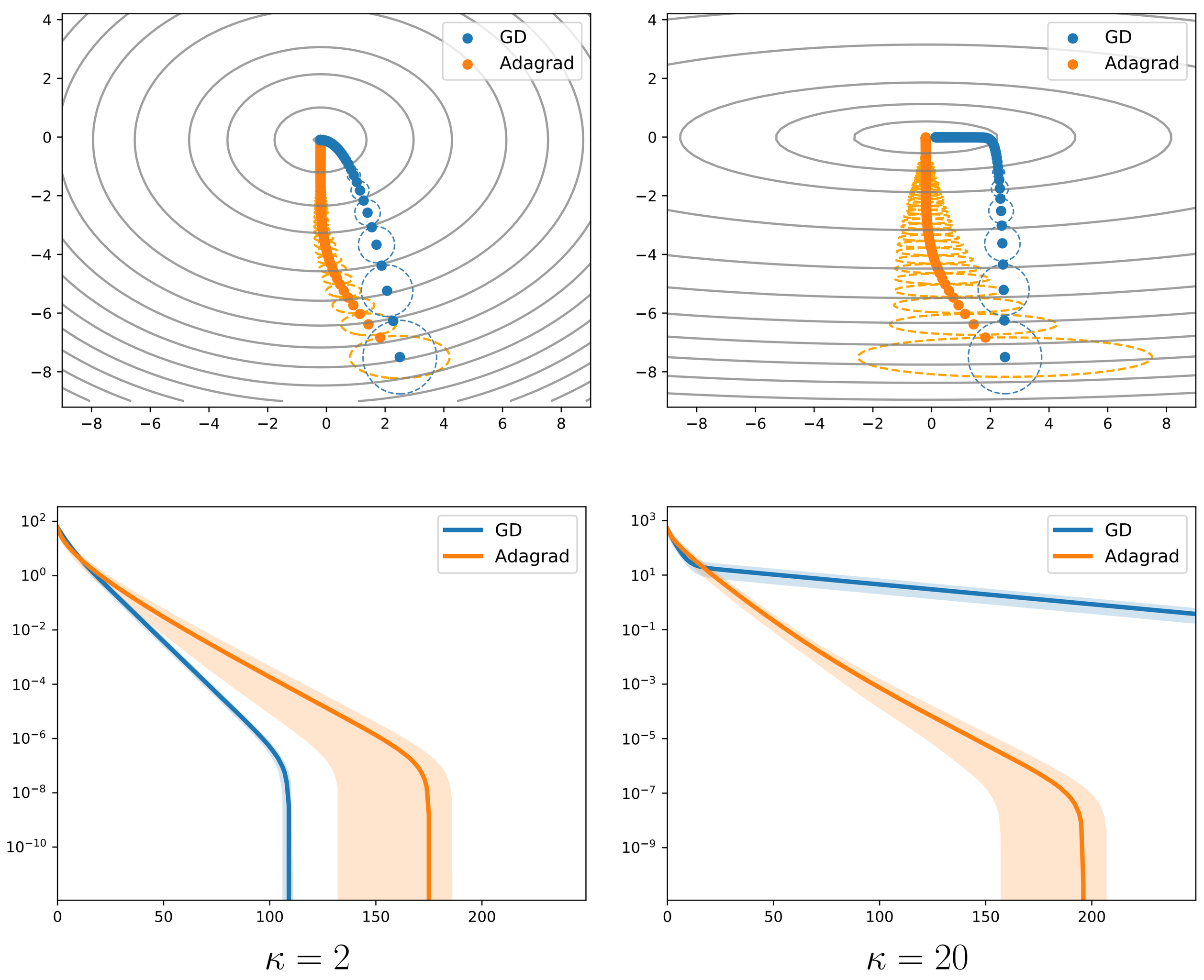
$$\min_{\mathbf{w}\in\mathbb{R}^d}\left[\mathcal{L}(\mathbf{w}) := \sum_{i=1}^{n}\ell(f(\mathbf{w},\mathbf{x}_i,\mathbf{y}_i))\right].$$

- Most widely used training algorithm in Neural Networks: SGD.
- SGD is known to be inadequate to optimize not well-conditioned functions → adaptive first-order methods (e.g. RMSProp, Adagrad, Adam).
- Newton methods have stronger theoretical guarantees (superlinear local convergence & provable escape from saddle points) by transforming ill-conditioned regions using Hessian information [CGT00].
- Recent stochastic extensions to the Trust-Region (TR) [CGT00] framework [XRKM17, YXRKM18, KL17, GRVZ17] make them applicable for Deep Learning.

We here propose to use ellipsoidal constraints in TR methods to make them even more suitable for Neural Network training.

Alternative View on Adaptive Gradient Methods

While gradient descent can be interpreted as a spherically constrained first-order TR method, preconditioned gradient methods—such as Adagrad—can be seen as first-order TR methods with ellipsoidal trust region constraint.



Theorem 1. *A preconditioned gradient step*

$$\mathbf{w}_{t+1} - \mathbf{w}_t = \mathbf{s}_t := -\eta_t \mathbf{A}_t^{-1}\mathbf{g}_t$$

with stepsize $\eta_t > 0$, symmetric positive definite preconditioner $\mathbf{A}_t \in \mathbb{R}^{d\times d}$ and $\mathbf{g}_t \neq 0$ minimizes a first-order model around $\mathbf{w}_t \in \mathbb{R}^d$ in an ellipsoid given by $\mathbf{A}_t$ in the sense that

$$\mathbf{s}_t := \arg\min_{\mathbf{s}\in\mathbb{R}^d}\left[m_t^1(\mathbf{s}) = \mathcal{L}(\mathbf{w}_t) + \mathbf{s}^\intercal\mathbf{g}_t\right], \qquad s.t. \quad \|\mathbf{s}\|_{\mathbf{A}_t} \leq \eta_t\|\mathbf{g}_t\|_{\mathbf{A}_t^{-1}}.$$

References

[CGT00] Andrew R Conn, Nicholas IM Gould, and Philippe L Toint. *Trust region methods*. SIAM, 2000.

[GRVZ17] Serge Gratton, Clément W Royer, Luís N Vicente, and Zaikun Zhang. Complexity and global rates of trust-region methods based on probabilistic models. *IMA Journal of Numerical Analysis*, 2017.

[KL17] Jonas Moritz Kohler and Aurelien Lucchi. Sub-sampled cubic regularization for non-convex optimization. In *International Conference on Machine Learning*, 2017.

[LBOM12] Yann A LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer, 2012.

[VDSH98] Patrick Van Der Smagt and Gerd Hirzinger. Solving the ill-conditioning in neural network learning. In *Neural networks: tricks of the trade*, pages 193–206. Springer, 1998.

[XRKM17] Peng Xu, Farbod Roosta-Khorasani, and Michael W Mahoney. Newton-type methods for non-convex optimization under inexact hessian information. *arXiv preprint arXiv:1708.07164*, 2017.

[YXRKM18] Zhewei Yao, Peng Xu, Farbod Roosta-Khorasani, and Michael W Mahoney. Inexact non-convex newton-type methods. *arXiv preprint arXiv:1802.06925*, 2018.

Second-order Trust Region Methods

$$\min_{\mathbf{s}\in\mathbb{R}^d}\left[m_t(\mathbf{s}) := \mathcal{L}(\mathbf{w}_t) + \mathbf{g}_t^\intercal\mathbf{s} + \frac{1}{2}\mathbf{s}^\intercal\mathbf{B}_t\mathbf{s}\right], \quad \text{s.t. } \|\mathbf{s}\|_{\mathbf{A}_t} \leq \Delta_t$$

- $\mathbf{A}_t$ induces the shape of the constraint set. Common choice for NN training: $\mathbf{A}_t = \mathbf{I}$.
- We prove that any TR method with an ellipsoidal constraint of the preconditioning matrix of RMSProp,

  $$\mathbf{A}_{rms,t} := \left((1-\beta)\mathbf{G}_t \operatorname{diag}(\beta^t,\dots,\beta^0)\mathbf{G}_t^\intercal\right) + \epsilon\mathbf{I},$$

  inherits all convergence guarantees ([CGT00], Theorem 6.6.8).

Why Ellipsoids?

- There are many sources for ill-conditioning in Neural Networks, e.g. uncentered and correlated inputs [LBOM12], saturated hidden units, and different weight scales in different layers [VDSH98].
- The spherical constraint is blind towards the loss surface. The RMS ellipsoid adaptively adjust its shape to fit the current region of the non-convex loss landscape.

Algorithm

Algorithm 1 Stochastic Ellipsoidal Trust Region Method

1: **Input:** $\mathbf{w}_0 \in \mathbb{R}^d, \gamma > 1, 1 > \eta > 0, \Delta_0 > 0$
2: **for** $t = 0, 1, \dots,$ until convergence **do**
3: Compute approximations $\mathbf{g}_t$ and $\mathbf{B}_t$. **If** $\|\mathbf{g}_t\| \leq \epsilon_g$, set $\mathbf{g}_t := 0$.
4: Set $\mathbf{A}_t := \mathbf{A}_{rms,t}$ or $\mathbf{A}_t := \operatorname{diag}(\mathbf{A}_{rms,t})$.
5: Obtain $\mathbf{s}_t$ by solving $m_t(\mathbf{s}_t)$ approximately.
6: Compute ratio of function over model decrease

$$\rho_t = \frac{\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_t + \mathbf{s}_t)}{m_t(\mathbf{0}) - m_t(\mathbf{s}_t)}$$

7: Set

$$\Delta_{t+1} = \begin{cases}\gamma\Delta_t & \text{if } \rho_{\mathcal{S},t} > \eta\\ \Delta_t/\gamma & \text{if } \rho_{\mathcal{S},t} < \eta\end{cases}, \quad \mathbf{w}_{t+1} = \begin{cases}\mathbf{w}_t + \mathbf{s}_t & \text{if } \rho_t \geq \eta \quad \text{(successful)}\\ \mathbf{w}_t & \text{otherwise} \quad \text{(unsuccessful)}\end{cases}$$

8: **end for**

Experiments

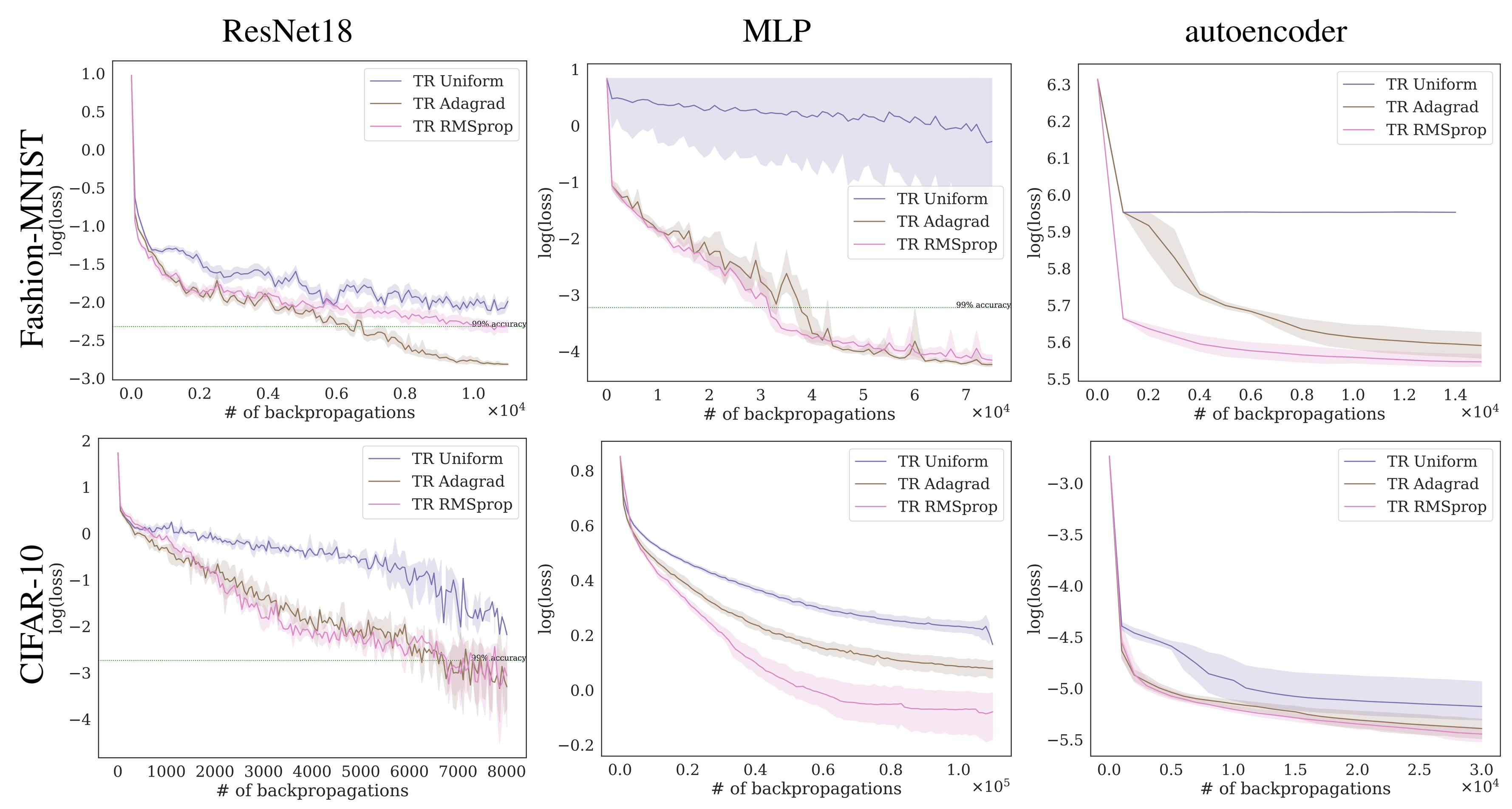


Figure 1: Log loss over number of backpropss. Average and 95% CI of 10 independent runs. Green dotted line indicates 99% accuracy.

- Ellipsoids are based on *diagonal* pre-conditioners and we employ Steihaug-Toint Conjugate Gradient method as subproblem solver.
- Ellipsoidal TR methods consistently outperform the spherical counterparts.
- An empirical comparison to common first-order methods suggests that further improvements in hardware are needed before Newton-type methods will replace them in Deep Learning.