Dissecting Adam: The Sign, Magnitude and Variance of Stochastic Gradients

Lukas Balles and Philipp Hennig

Max Planck Institute for Intelligent Systems and University of Tübingen

MAX-PLANCK-GESELLSCHAFT

Dissecting Adam

Adam [2] updates $\theta_{t+1} = \theta_t - \alpha \frac{m_t}{\sqrt{v_t + \varepsilon}}$, where

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t, \qquad v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2.$$

We can rewrite the update direction as

$$\frac{m_t}{\sqrt{v_t}} = \frac{\mathrm{sign}(m_t)}{\sqrt{\frac{v_t}{m_t^2}}} = \sqrt{\frac{1}{1 + \frac{v_t - m_t^2}{m_t^2}}} \odot \mathrm{sign}(m_t),$$

The magnitude of $m_{t,i}$ only appears in the ratio

$$\hat{\eta}_{t,i}^2 := \frac{v_{t,i} - m_{t,i}^2}{m_{t,i}^2} \quad \text{which approximates} \quad \frac{\sigma_{t,i}^2}{\nabla \mathcal{L}_{t,i}^2} =: \eta_{t,i}^2$$

if we assume $m_t \approx \mathbf{E}[g_t]$ and $v_t \approx \mathbf{E}[g_t^2]$.

Adam combines two aspects

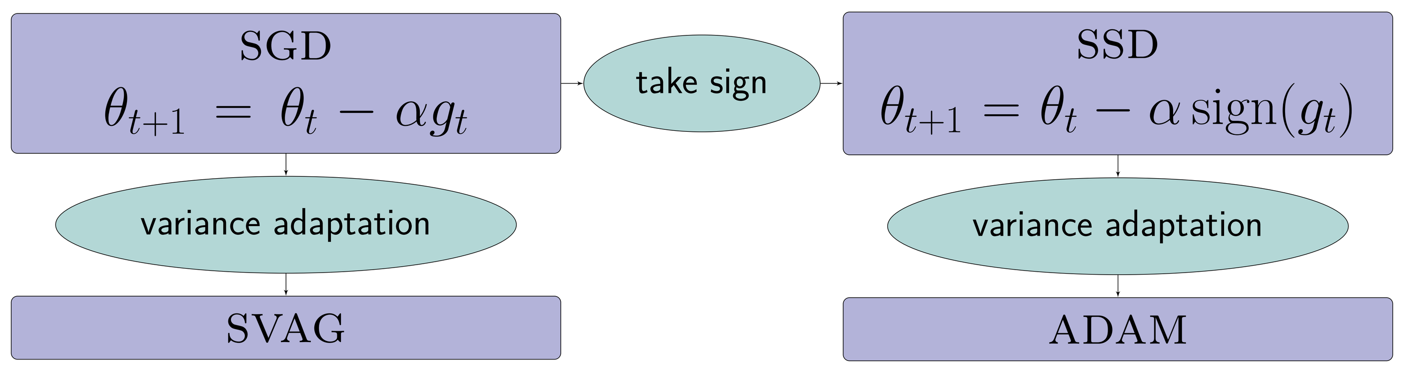
Taking the sign: Equal update magnitude for each coordinate.
Variance adaptation: The update magnitudes are scaled with

$$\gamma_{t,i} := \sqrt{\frac{1}{1 + \hat{\eta}_{t,i}^2}},$$

i.e., based on the *noise-to-signal ratio* of the gradient coordinates.

Overview

We disentangle these two aspects to discuss and analyze them in isolation.



Why the Sign?

We compare SGD and SSD on the "noisy quadratic" toy problem.

- SSD is less sensitive to the eigenspectrum than SGD.
- SSD avoids interaction between noise and eigenspectrum.
- SSD depends on the axis-alignment of the QP.

More detailed analysis in concurrent work [1].

Variance Adaptation

- We want the direction $\nabla\mathcal{L}$ (or $\mathrm{sign}(\nabla\mathcal{L})$) but only have g.
- If we update $\gamma \odot g$ (or $\gamma \odot \mathrm{sign}(g)$), what is the best γ?

Lemma

$\mathbf{E}[\|\gamma \odot g - \nabla\mathcal{L}\|_2^2]$ and $\mathbf{E}[\|\gamma \odot \mathrm{sign}(g) - \mathrm{sign}(\nabla\mathcal{L})\|_2^2]$ are minimized by

$$\gamma_i = \frac{1}{1 + \eta_i^2} \quad \text{and} \quad \gamma_i = 2\rho_i - 1,$$

respectively, where $\eta_i^2 = \sigma_i^2 / \nabla\mathcal{L}_i^2$ and $\rho_i := \mathbf{P}[\mathrm{sign}(g_i) = \mathrm{sign}(\nabla\mathcal{L}_i)]$.

SSD + variance adaptation ≈ Adam

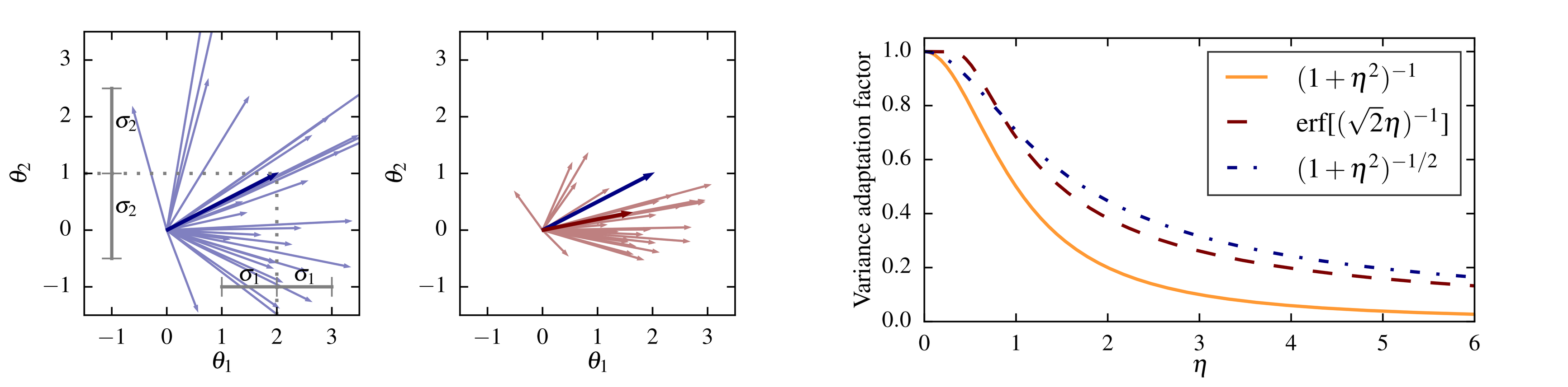
If $g \sim \mathcal{N}(\nabla\mathcal{L}, \Sigma)$, then $2\rho_i - 1 = \mathrm{erf}\left((\sqrt{2}\eta_i)^{-1}\right) \approx (1 + \eta_i)^{-1/2}$.

SGD + variance adaptation = SVAG

Stochastic Variance-Adapted Gradient:

$$\theta_{t+1} = \theta_t - \alpha_t \frac{1}{1 + \eta_t} \odot g_t$$

- The idealized version converges at $\mathcal{O}(1/t)$ with constant $\alpha_t \equiv \alpha$.
- Practical realization is similar to Adam.



TL;DR

Adam combines two aspects, the *sign aspect* and the *variance adaptation aspect*, which we disentangle in this paper.

Connection to Generalization

Wilson et al. [3] show that Adam has adverse effects on generalization.

- Binary least-squares classification problem:
$$\min_{\theta \in \mathbb{R}^d} R(\theta) = \frac{1}{2n}\|X\theta - y\|^2, \quad X \in \mathbb{R}^{n\times d}, y \in \{\pm 1\}^n.$$
- SGD converges to the max-margin solution.
- They construct instances where Adam finds arbitrarily bad solutions.
- *Lemma 3.1 in [3]*: Suppose $[X^T y]_i \neq 0$ for all i, and $\exists c \in \mathbb{R}$ such that $X\,\mathrm{sign}(X^T y) = cy$. Then, for $\theta_0 = 0$, the iterates of full-batch Adam satisfy $\theta_t \propto \mathrm{sign}(X^T y)$.

The Lemma easily extends to sign descent, but not to (M-)SVAG.

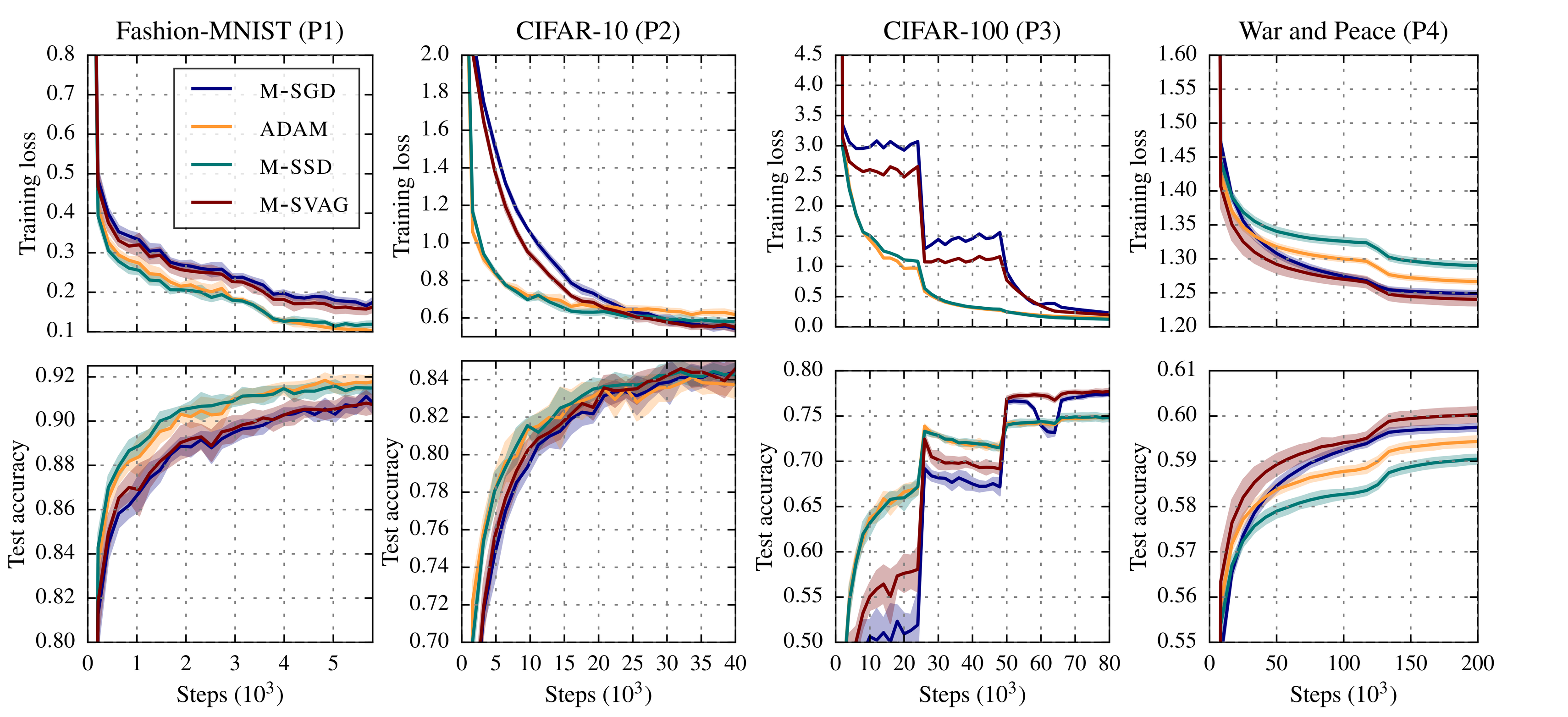
Is the sign aspect the perpetrator?

Practical Implementation of M-SVAG

Input: $\theta_0 \in \mathbb{R}^d$, $\alpha > 0$, $\beta \in [0,1]$, $T \in \mathbb{N}$
Initialize $\theta \leftarrow \theta_0$, $\tilde{m} \leftarrow 0$, $\tilde{v} \leftarrow 0$
for $t = 0, \dots, T-1$ **do**
  $\tilde{m} \leftarrow \beta\tilde{m} + (1-\beta) g(\theta)$, $\tilde{v} \leftarrow \beta\tilde{v} + (1-\beta) g(\theta)^2$
  $m \leftarrow (1-\beta^{t+1})^{-1}\tilde{m}$, $v \leftarrow (1-\beta^{t+1})^{-1}\tilde{v}$
  $\rho \leftarrow \frac{(1-\beta)(1+\beta^{t+1})}{(1+\beta)(1-\beta^{t+1})}$
  $s \leftarrow (1-\rho)^{-1}(v - m^2)$
  $\gamma \leftarrow m^2/(m^2 + \rho s)$
  $\theta \leftarrow \theta - \alpha(\gamma \odot m)$
end for

Find a TensorFlow implementation at
`github.com/lballes/svag`

Experimental Results



Observations:

- The sign aspect is dominant.
- Usefulness of the sign depends on the problem.
- Variance adaptation helps.
- Adverse generalization effects can be attributed to the sign aspect.

References

[1] Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Anima Anandkumar. signSGD: compressed optimisation for non-convex problems. In *ICML*, 2018.

[2] Diederik Kingma and Jimmy Ba. ADAM: A method for stochastic optimization. *ICLR*, 2015.

[3] Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. In *NIPS*, 2017.

This work will be presented at ICML 2018.