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

Dissecting Adam
Adam [2] updates \( \theta_{t+1} = \theta_t - \alpha \frac{m_t}{\sqrt{v_t} + \epsilon} \), where
\[
m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t, \quad v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2.
\]
We can rewrite the update direction as
\[
m_t = \frac{\text{sign}(m_t)}{\sqrt{v_t}} = \frac{1}{\sqrt{1 + \frac{m_t}{v_t}}} \odot \text{sign}(m_t),
\]
The magnitude of \( m_t \) only appears in the ratio
\[
\frac{m_t}{v_t} = \frac{m_t}{1 + \frac{m_t}{v_t}}
\]
if we assume \( m_t \approx E[g_t] \) and \( v_t \approx E[g_t^2] \).

Adam combines two aspects

Taking the sign: Equal update magnitude for each coordinate.

Variance Adaptation: The update magnitudes are scaled with
\[
\gamma_t := \frac{1}{\sqrt{1 + \frac{m_t}{v_t}}}
\]
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.
- SGD 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 \( \nabla L (\text{or sign(\nabla L))} \) but only have \( g \).
- If we update \( \gamma \odot g \) (or \( \gamma \odot \text{sign(g)} \)), what is the best \( \gamma \)?

Lemma
\[
\mathbb{E}[\| \gamma \odot g - \nabla L \|_2^2] \quad \text{and} \quad \mathbb{E}[\| \gamma \odot \text{sign(g)} - \text{sign(\nabla L)} \|_2^2]
\]
is minimized by
\[
\gamma_t = \frac{1}{1 + \frac{m_t}{v_t}} \quad \text{and} \quad \gamma_t = 2 \rho_t - 1,
\]
respectively, where \( \rho_t := \mathbb{P}[\text{sign}(g_t) = \text{sign}(\nabla L_t)] \).

SSD + variance adaptation \( \approx \) Adam
If \( g \sim \mathcal{N}(\nabla L, \Sigma) \), then \( 2 \rho_t - 1 = \text{erf} \left( \frac{\sqrt{2} \Sigma_{11}^{1/2}}{1 + \eta_t} \right) \approx (1 + \eta_t)^{-1/2} \).

SGD + variance adaptation \( = \) SVAG

Stochastic Variance-Adapted Gradient:
\[
\theta_{t+1} = \theta_t - \alpha \frac{1}{1 + \eta_t} \odot g_t
\]
- The idealized version converges at \( O(1/t) \) with constant \( \alpha \).
- Practical realization is similar to Adam.

Connection to Generalization
Wilson et al. [3] show that Adam has adverse effects on generalization.
- Binary least-squares classification problem:
  \[
  \min_{\theta \in \mathbb{R}^d} \frac{1}{2n} \| X \theta - y \|_2^2, \quad X \in \mathbb{R}^{n \times d}, y \in \{\pm 1\}^n.
  \]
- SGD converges to the max-margin solution.
- They construct instances where Adam finds arbitrarily bad solutions.
- Lemma 3.1 in [3]: Suppose \( X \text{sign}(X^T y) \neq 0 \) for all \( i \), and \( 3c \in \mathbb{R} \) such that \( X \text{sign}(X^T y) = c \).
  Then, for \( \theta_t = 0 \), the iterates of full-batch Adam satisfy \( \theta_t \propto \text{sign}(X^T y) \).

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 \( \theta \leftarrow \theta_0, \; m = 0, \; v \leftarrow 0 \)
for \( t = 0, \ldots, T - 1 \) do
  \( m \leftarrow \beta m + (1 - \beta) g(\theta) \), \( \hat{v} \leftarrow \beta \hat{v} + (1 - \beta) g(\theta)^2 \)
  \( m \leftarrow (1 - \beta)^{t+1} m, \; \hat{v} \leftarrow (1 - \beta)^{t+1} \hat{v} \)
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

This work will be presented at ICML 2018.

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