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

Presented by Fangcheng Fu

2019/03/07

Generic Adaptive Method

Algorithm 1 Generic Adaptive Method Setup

Input: $x_1 \in \mathcal{F}$, step size $\{\alpha_t > 0\}_{t=1}^{T}$, sequence of functions $\{\phi_t, \psi_t\}_{t=1}^{T}$

for $t = 1$ **to** T **do**

$g_t = \nabla f_t(x_t)$

$m_t = \phi_t(g_1, \ldots, g_t)$ and $V_t = \psi_t(g_1, \ldots, g_t)$

$\hat{x}_{t+1} = x_t - \alpha_t m_t / \sqrt{V_t}$

$x_{t+1} = \Pi_{\mathcal{F}, \sqrt{V_t}}(\hat{x}_{t+1})$

end for

	SGD	SGDM	ADAGRAD	RMSPROP	ADAM
ϕ_t	g_t	$\sum_{i=1}^{t} \gamma^{t-i} g_i$	g_t	g_t	$(1-\beta_1)\sum_{i=1}^{t} \beta_1^{t-i} g_i$
ψ_t	$\mathbb{I}$	$\mathbb{I}$	$\mathrm{diag}(\sum_{i=1}^{t} g_i^2)/t$	$(1-\beta_2)\mathrm{diag}(\sum_{i=1}^{t} \beta_2^{t-i} g_i^2)$	$(1-\beta_2)\mathrm{diag}(\sum_{i=1}^{t} \beta_2^{t-i} g_i^2)$

Adam has attracted a surge of research interests due to its fast speed

Dissecting Adam

$$\tilde{m}_t = \beta_1 \tilde{m}_{t-1} + (1-\beta_1) g_t, \quad m_t = \frac{\tilde{m}_t}{1-\beta_1^{t+1}},$$

$$\tilde{v}_t = \beta_2 \tilde{v}_{t-1} + (1-\beta_2) g_t^2, \quad v_t = \frac{\tilde{v}_t}{1-\beta_2^{t+1}},$$

with $\beta_1, \beta_2 \in (0, 1)$ and updates

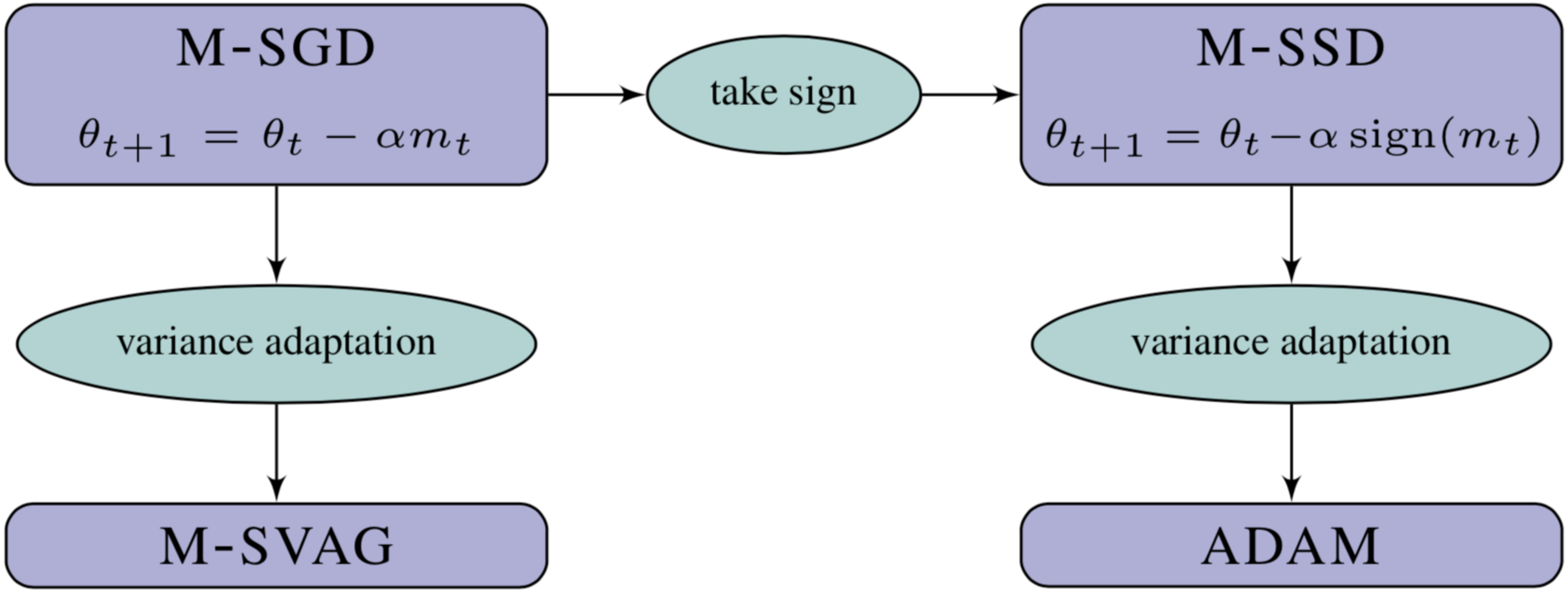
$$\theta_{t+1} = \theta_t - \alpha \frac{m_t}{\sqrt{v_t} + \varepsilon}$$

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

➔ Update direction is given by the sign $m_{t,i}$

➔ Update magnitude is determined by relative variance $\hat{\eta}_{t,i}^2 := \frac{v_{t,i} - m_{t,i}^2}{m_{t,i}^2} \approx \frac{\sigma_{t,i}^2}{\nabla \mathcal{L}_{t,i}^2} =: \eta_{t,i}^2$

Dissecting Adam



The Sign

What is the success probability of the direction?

$$\rho_i := \mathbf{P}\left[s_i = \operatorname{sign}(\nabla \mathcal{L}_i)\right] \text{ where } s = \operatorname{sign}(g)$$

By Central Limit Theorem, we can assume $g_i \sim N(0,1)$, then we have

$$\rho_i = \frac{1}{2} + \frac{1}{2}\operatorname{erf}\left(\frac{|\nabla \mathcal{L}_i|}{\sqrt{2}\sigma_i}\right)$$

The Sign

Get some intuition from a simple but insightful problem:

Model Problem (Stochastic Quadratic Problem, sQP). *Consider the loss function $\ell(\theta; x) = 0.5\,(\theta - x)^T Q(\theta - x)$ with a symmetric positive definite matrix $Q \in \mathbb{R}^{d\times d}$ and 'data" coming from the distribution $x \sim \mathcal{N}(x^*, \nu^2 I)$ with $\nu \in \mathbb{R}_+$. The objective $\mathcal{L}(\theta) = \mathbf{E}_x[\ell(\theta; x)]$ evaluates to*

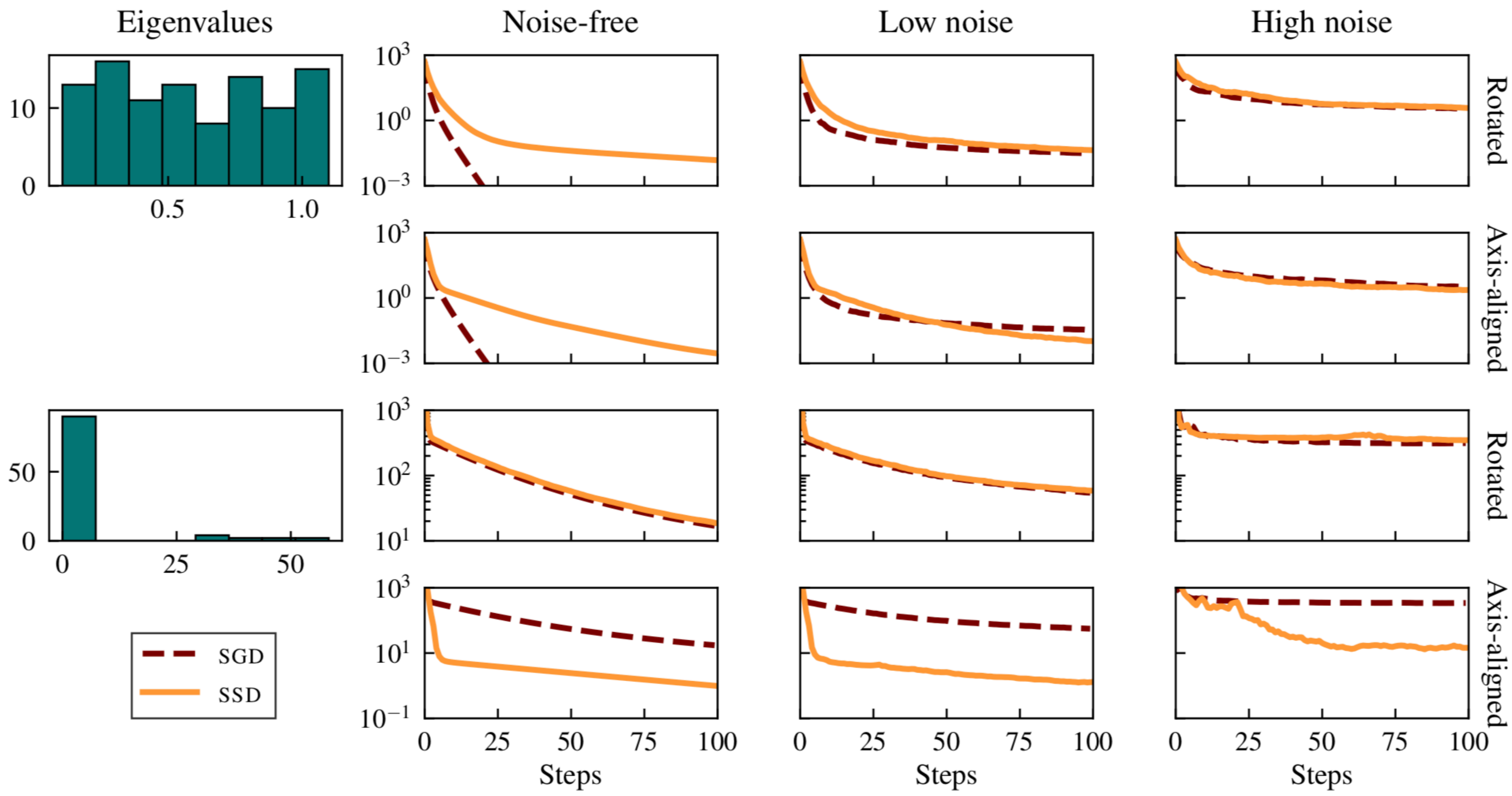
$$\mathcal{L}(\theta) = \frac{1}{2}(\theta - x^*)^T Q(\theta - x^*) + \frac{\nu^2}{2}\operatorname{tr}(Q), \tag{11}$$

with $\nabla\mathcal{L}(\theta) = Q(\theta - x^*)$. Stochastic gradients are given by $g(\theta) = Q(\theta - x) \sim \mathcal{N}(\nabla\mathcal{L}(\theta), \nu^2 QQ)$.

Conclusion drawn from theoretical analysis:

1) sign-based: noisy, ill-conditioned problems with diagonally dominant Hessians
2) non-sign-based: low-noise, arbitrarily-rotated eigenbases

The Sign



Eigenvalues
Noise-free
Low noise
High noise
Rotated
Axis-aligned
Rotated
Axis-aligned
Steps
Steps
Steps
SGD
SSD

Variance Adaptation

Assume we want to update a direction p, but only know $\hat{p}$ $s.t.$ $\mathbf{E}[\hat{p}] = p$

How to update?

Lemma 1. *Let $\hat{p} \in \mathbb{R}^d$ be a random variable with $\mathbf{E}[\hat{p}] = p$ and $\mathbf{var}[p_i] = \sigma_i^2$. Then $\mathbf{E}[\|\gamma \odot \hat{p} - p\|_2^2]$ is minimized by*

$$\gamma_i = \frac{\mathbf{E}[\hat{p}_i]^2}{\mathbf{E}[\hat{p}_i^2]} = \frac{p_i^2}{p_i^2 + \sigma_i^2} = \frac{1}{1 + \sigma_i^2/p_i^2} \tag{15}$$

and $\mathbf{E}[\|\gamma \odot \operatorname{sign}(\hat{p}) - \operatorname{sign}(p)\|_2^2]$ is minimized by

$$\gamma_i = (2\rho_i - 1), \tag{16}$$

where $\rho_i := \mathbf{P}[\operatorname{sign}(\hat{p}_i) = \operatorname{sign}(p_i)]$. *(Proof in §B.3)*

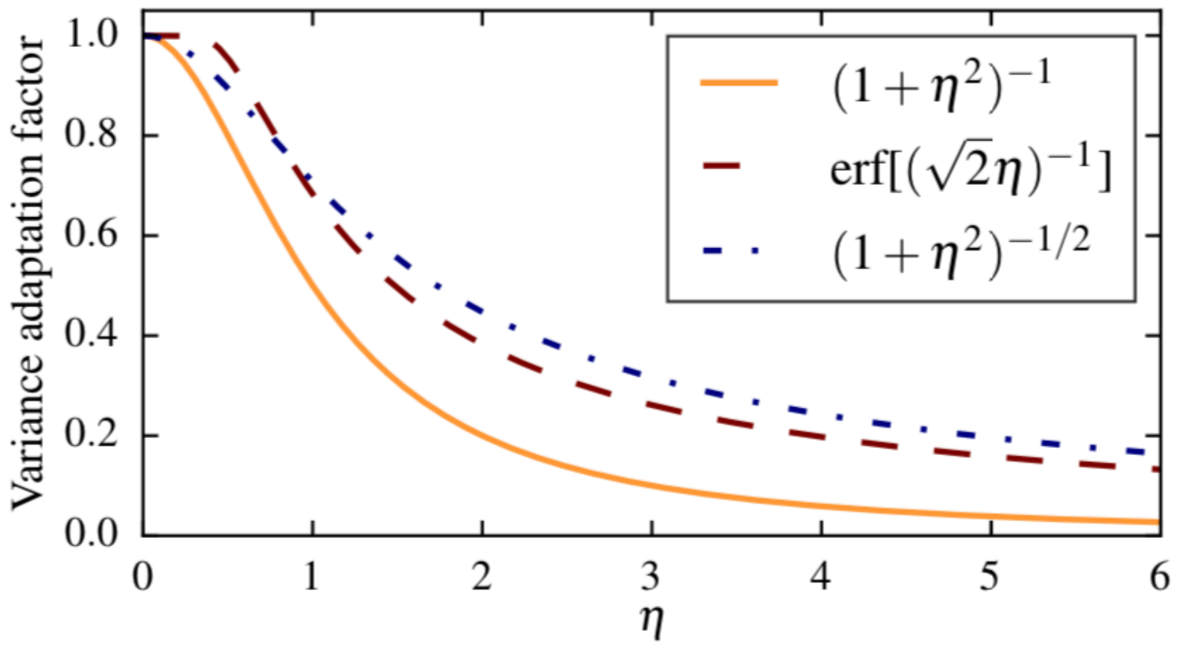
Variance Adaptation

Adam as a Variance-Adapted Sign Descent

Optimal: $\gamma_i = 2\rho_i - 1 = \mathrm{erf}[(\sqrt{2}\eta_i)^{-1}]$

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

$$\hat{\eta}^2_{t,i} := \frac{v_{t,i} - m^2_{t,i}}{m^2_{t,i}} \approx \frac{\sigma^2_{t,i}}{\nabla\mathcal{L}^2_{t,i}} =: \eta^2_{t,i}$$

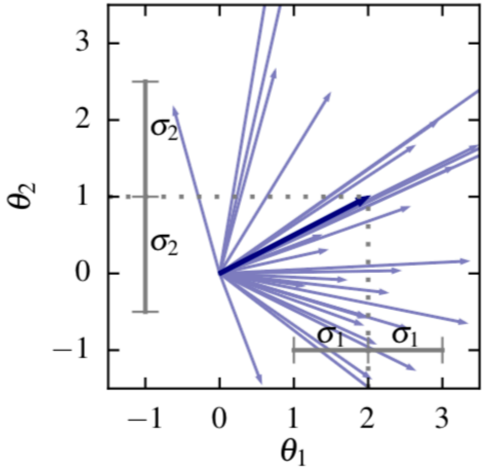


Adam is a realization of optimal variance adaptation regarding m instead of g

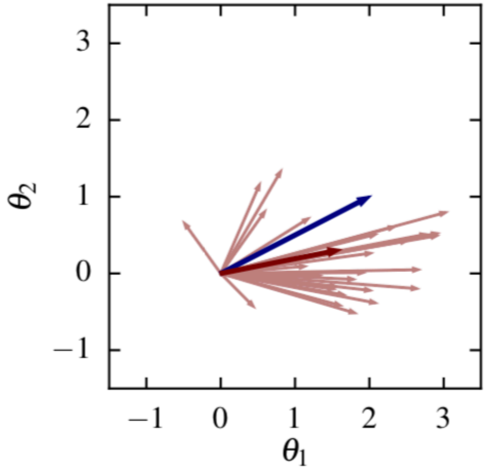
Variance Adaptation

Stochastic Variance-Adapted Gradient (SVAG)

Optimal: Let $\hat{p} = g$, we have $\gamma_i^g = \dfrac{\nabla\mathcal{L}_i^2}{\nabla\mathcal{L}_i^2 + \sigma_i^2} = \dfrac{1}{1 + \sigma_i^2/\nabla\mathcal{L}_i^2} = \dfrac{1}{1 + \eta_i^2}$



$(\sigma_1, \sigma_2) = (1, 1.5)$



$(\eta_1, \eta_2) = (2.25, 0.25)$

Shorten the axis with higher variance

Provably $O(1/t)$ convergence rate

$$\mathbf{E}[f(\theta_t) - f_*] \in \mathcal{O}\left(\frac{1}{t}\right)$$

We skip the extension from SVAG to M-SVAG.
Please refer to the paper for more details.

Experiments

Algorithms: M-SGD, Adam, M-SSD, M-SVAG

Algorithm 1 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**
$\quad \tilde{m} \leftarrow \beta\tilde{m} + (1-\beta)g(\theta), \quad \tilde{v} \leftarrow \beta\tilde{v} + (1-\beta)g(\theta)^2$
$\quad m \leftarrow (1-\beta^{t+1})^{-1}\tilde{m}, \quad v \leftarrow (1-\beta^{t+1})^{-1}\tilde{v}$
$\quad s \leftarrow (1-\rho(\beta,t))^{-1}(v - m^2)$
$\quad \gamma \leftarrow m^2/(m^2 + \rho(\beta,t)s)$
$\quad \theta \leftarrow \theta - \alpha(\gamma \odot m)$
end for

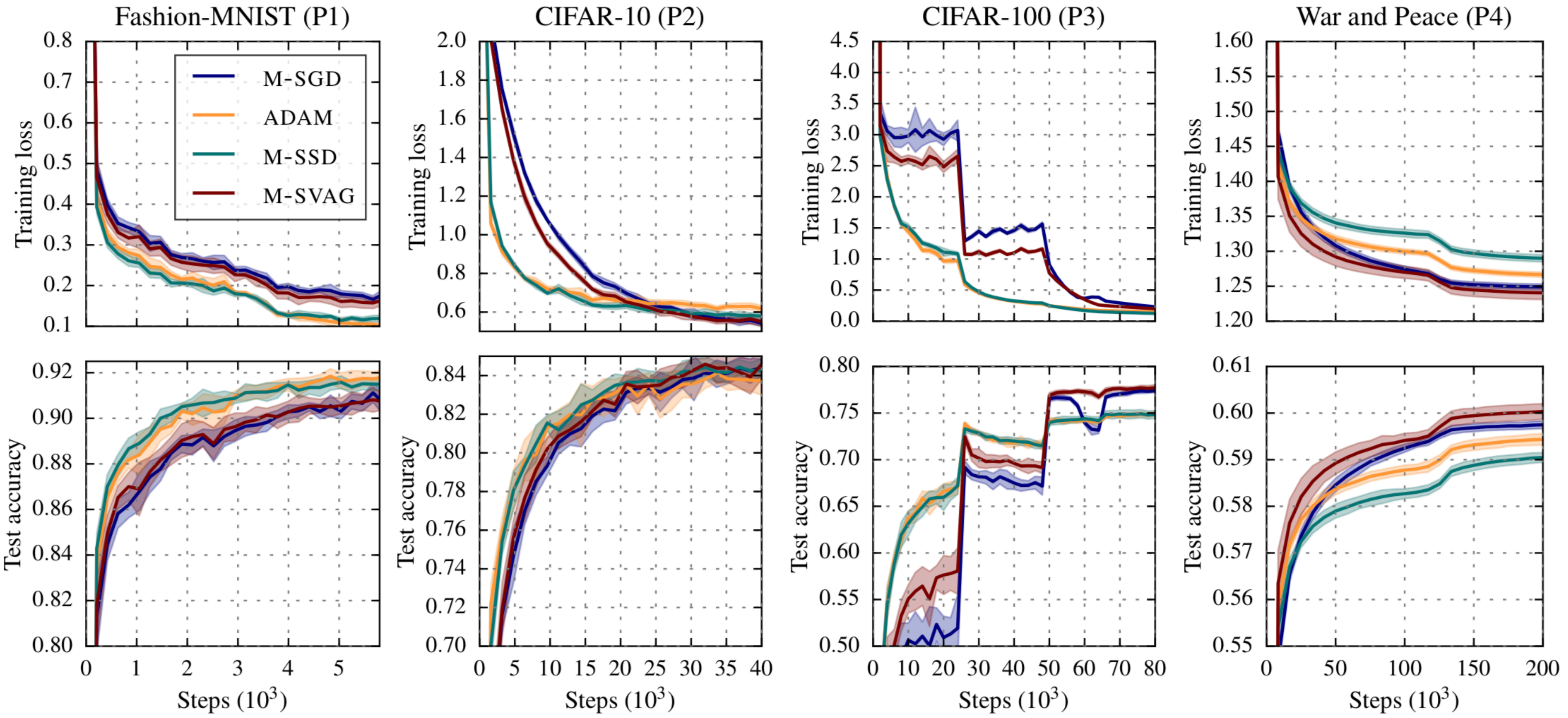
Note: M-SVAG exposes two hyperparameters, α and β.

Algorithm 2 M-SGD and M-SSD

Input: $\theta_0 \in \mathbb{R}^d$, $\alpha > 0$, $\beta \in [0,1]$, $T \in \mathbb{N}$
Initialize $\theta \leftarrow \theta_0$, $\tilde{m} \leftarrow 0$
for $t = 0, \dots, T-1$ **do**
$\quad \tilde{m} \leftarrow \beta\tilde{m} + (1-\beta)g(\theta)$
$\quad m \leftarrow (1-\beta^{t+1})^{-1}\tilde{m}$
$\quad \theta \leftarrow \theta - \alpha m \qquad \theta \leftarrow \theta - \alpha\,\mathrm{sign}(\tilde{m})$
end for

Experiments

1) Sign aspect dominates
2) Usefulness of sign is problem-dependent
3) Variance adaption helps
4) Generalization effect are caused by the sign



Experiments

Additional question: signed-based methods require small learning rate?

Problem 1: Fashion-MNIST

M-SGD:
$3, 1, 6 \cdot 10^{-1}, 3 \cdot 10^{-1}, \mathbf{1 \cdot 10^{-1}}, 6 \cdot 10^{-2}, 3 \cdot 10^{-2}, 1 \cdot 10^{-2}, 6 \cdot 10^{-3}, 3 \cdot 10^{-3}$

ADAM:
$3 \cdot 10^{-2}, 10^{-2}, 6 \cdot 10^{-3}, 3 \cdot 10^{-3}, \mathbf{1 \cdot 10^{-3}}, 6 \cdot 10^{-4}, 3 \cdot 10^{-4}, 1 \cdot 10^{-4}$

M-SSD:
$10^{-2}, 6 \cdot 10^{-3}, 3 \cdot 10^{-3}, 1 \cdot 10^{-3}, 6 \cdot 10^{-4}, \mathbf{3 \cdot 10^{-4}}, 1 \cdot 10^{-4}$

M-SVAG:
$3, 1, 6 \cdot 10^{-1}, \mathbf{3 \cdot 10^{-1}}, 1 \cdot 10^{-1}, 6 \cdot 10^{-2}, 3 \cdot 10^{-2}, 1 \cdot 10^{-2}, 6 \cdot 10^{-3}, 3 \cdot 10^{-3}$

Problem 2: CIFAR-10

M-SGD:
$6 \cdot 10^{-1}, 3 \cdot 10^{-1}, 1 \cdot 10^{-1}, 6 \cdot 10^{-2}, \mathbf{3 \cdot 10^{-2}}, 1 \cdot 10^{-2}, 6 \cdot 10^{-3}, 3 \cdot 10^{-3}$

ADAM:
$6 \cdot 10^{-3}, 3 \cdot 10^{-3}, 1 \cdot 10^{-3}, \mathbf{6 \cdot 10^{-4}}, 3 \cdot 10^{-4}, 1 \cdot 10^{-4}, 6 \cdot 10^{-5}$

M-SSD:
$6 \cdot 10^{-3}, 3 \cdot 10^{-3}, 1 \cdot 10^{-3}, 6 \cdot 10^{-4}, 3 \cdot 10^{-4}, \mathbf{1 \cdot 10^{-4}}, 6 \cdot 10^{-5}, 3 \cdot 10^{-5}$

M-SVAG:
$1, 6 \cdot 10^{-1}, 3 \cdot 10^{-1}, 1 \cdot 10^{-1}, \mathbf{6 \cdot 10^{-2}}, 3 \cdot 10^{-2}, 1 \cdot 10^{-2}, 6 \cdot 10^{-3}$

Experiments

Additional question: signed-based methods require small learning rate?

Problem 3: CIFAR-100

M-SGD:

$6, \mathbf{3}, 1, 6\cdot 10^{-1}, 3\cdot 10^{-1}, 1\cdot 10^{-1}, 6\cdot 10^{-2}, \mathbf{3\cdot 10^{-2}}, 1\cdot 10^{-2}$

ADAM:

$1\cdot 10^{-2}, 6\cdot 10^{-3}, 3\cdot 10^{-3}, 1\cdot 10^{-3}, 6\cdot 10^{-4}, \mathbf{3\cdot 10^{-4}}, 1\cdot 10^{-4}, 6\cdot 10^{-5}, 3\cdot 10^{-5}$

M-SSD:

$1\cdot 10^{-2}, 6\cdot 10^{-3}, 3\cdot 10^{-3}, 1\cdot 10^{-3}, 6\cdot 10^{-4}, 3\cdot 10^{-4}, \mathbf{1\cdot 10^{-4}}, 6\cdot 10^{-5}, 3\cdot 10^{-5}$

M-SVAG:

$6, \mathbf{3}, 1, 6\cdot 10^{-1}, 3\cdot 10^{-1}, 1\cdot 10^{-1}, 6\cdot 10^{-2}, \mathbf{3\cdot 10^{-2}}, 1\cdot 10^{-2}$

Problem 4: War and Peace

M-SGD:

$10, 6, \mathbf{3}, 1, 6\cdot 10^{-1}, 3\cdot 10^{-1}, 1\cdot 10^{-1}, 6\cdot 10^{-2}$

ADAM:

$1\cdot 10^{-2}, 6\cdot 10^{-3}, \mathbf{3\cdot 10^{-3}}, 1\cdot 10^{-3}, 6\cdot 10^{-4}, 3\cdot 10^{-4}, 1\cdot 10^{-4}, 6\cdot 10^{-5}$

M-SSD:

$1\cdot 10^{-2}, 6\cdot 10^{-3}, 3\cdot 10^{-3}, \mathbf{1\cdot 10^{-3}}, 6\cdot 10^{-4}, 3\cdot 10^{-4}, 1\cdot 10^{-4}, 6\cdot 10^{-5}$

M-SVAG:

$30, \mathbf{10}, 6, 3, 1, 6\cdot 10^{-1}, 3\cdot 10^{-1}, 1\cdot 10^{-1}$

Thank you