—Supplementary Material—

A. Experiments

A.1. Network Architectures

Fashion-MNIST We trained a simple convolutional neural network with two convolutional layers (size 5×5, 32 and 64 filters, respectively), each followed by max-pooling over 3×3 areas with stride 2, and a fully-connected layer with 1024 units. ReLU activation was used for all layers. The output layer has 10 units with softmax activation. We used cross-entropy loss, without any additional regularization, and a mini-batch size of 64. We trained for a total of 6000 steps with a constant global step size α.

CIFAR-10 We trained a CNN with three convolutional layers (64 filters of size 5×5, 96 filters of size 3×3, and 128 filters of size 3×3) interspersed with max-pooling over 3×3 areas with stride 2 and followed by two fully-connected layers with 512 and 256 units. ReLU activation was used for all layers. The output layer has 10 units with softmax activation. We used cross-entropy loss function and applied L_2-regularization on all weights, but not the biases. During training we performed some standard data augmentation operations (random cropping of sub-images, left-right mirroring, color distortion) on the input images. We used a batch size of 128 and trained for a total of 40k steps with a constant global step size α.

CIFAR-100 We use the WRN-40-4 architecture of Zagoruyko & Komodakis (2016); details can be found in the original paper. We used cross-entropy loss and applied L_2-regularization on all weights, but not the biases. We used the same data augmentation operations as for CIFAR-10, a batch size of 128, and trained for 80k steps. For the global step size α, we used the decrease schedule suggested by Zagoruyko & Komodakis (2016), which amounts to multiplying with a factor of 0.2 after 24k, 48k, and 64k steps. TensorFlow code was adapted from https://github.com/dalgu90/wrn-tensorflow.

War and Peace We preprocessed *War and Peace*, extracting a vocabulary of 83 characters. The language model is a two-layer LSTM with 128 hidden units each. We used a sequence length of 50 characters and a batch size of 50. Drop-out regularization was applied during training. We trained for 200k steps; the global step size α was multiplied with a factor of 0.1 after 125k steps. TensorFlow code was adapted from https://github.com/sherjilozair/char-rnn-tensorflow.

A.2. Step Size Tuning

Step sizes α (initial step sizes for the experiments with a step size decrease schedule) for each optimizer have been tuned by first finding the maximal stable step size by trial and error and then searching downwards over multiple orders of magnitude, testing $6 \cdot 10^m$, $3 \cdot 10^m$, and $1 \cdot 10^m$ for order of magnitude m. We evaluated loss and accuracy on the full test set (as well as on an equally-sized portion of the training set) at a constant interval and selected the best-performing step size for each method in terms of maximally reached test accuracy. Using the best choice, we replicated the experiment ten times with different random seeds, randomizing the parameter initialization, data set shuffling, drop-out, et cetera. In some rare cases where the accuracies for two different step sizes were very close, we replicated both and then chose the one with the higher maximum mean accuracy.

The following list shows all explored step sizes, with the "winner" in bold face.

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

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

B. Mathematical Details

B.1. The Sign of a Stochastic Gradient

We have stated in the main text that the sign of a stochastic gradient, $s(\theta) = \text{sign}(g(\theta))$, has success probabilities

$$\begin{aligned}\rho_i &:= \mathbf{P}[s(\theta)_i = \text{sign}(\nabla\mathcal{L}(\theta)_i)] \\ &= \frac{1}{2} + \frac{1}{2}\,\text{erf}\left(\frac{|\nabla\mathcal{L}(\theta)_i|}{\sqrt{2}\sigma(\theta)_i}\right)\end{aligned} \tag{27}$$

under the assumption that $g \sim \mathcal{N}(\nabla\mathcal{L}, \Sigma)$. The following Lemma formally proves this statement and Figure 6 provides a pictorial illustration.

Lemma 4. *If $X \sim \mathcal{N}(\mu, \sigma^2)$ then*

$$\mathbf{P}[\text{sign}(X) = \text{sign}(\mu)] = \frac{1}{2}\left(1 + \text{erf}\left(\frac{|\mu|}{\sqrt{2}\sigma}\right)\right). \tag{28}$$

Proof. Define $\rho := \mathbf{P}[\text{sign}(X) = \text{sign}(\mu)]$. The cumulative density function (cdf) of $X \sim \mathcal{N}(\mu, \sigma^2)$ is $\mathbf{P}[X \leq x] = \Phi((x-\mu)/\sigma)$, where $\Phi(z) = 0.5(1 + \text{erf}(z/\sqrt{2}))$ is the cdf of the standard normal distribution. If $\mu < 0$, then

$$\begin{aligned}\rho &= \mathbf{P}[X < 0] = \Phi\left(\frac{0-\mu}{\sigma}\right) \\ &= \frac{1}{2}\left(1 + \text{erf}\left(\frac{-\mu}{\sqrt{2}\sigma}\right)\right).\end{aligned} \tag{29}$$

If $\mu > 0$, then

$$\begin{aligned}\rho &= \mathbf{P}[X > 0] = 1 - \mathbf{P}[X \leq 0] = 1 - \Phi\left(\frac{0-\mu}{\sigma}\right) \\ &= 1 - \frac{1}{2}\left(1 + \text{erf}\left(\frac{-\mu}{\sqrt{2}\sigma}\right)\right) \\ &= \frac{1}{2}\left(1 + \text{erf}\left(\frac{\mu}{\sqrt{2}\sigma}\right)\right),\end{aligned} \tag{30}$$

where the last step used the anti-symmetry of the error function. □

B.2. Analysis on Stochastic QPs

B.2.1. Derivation of $\mathcal{I}_{\text{SGD}}$ and $\mathcal{I}_{\text{SSD}}$

We derive the expressions in Eq. (13), dropping the fixed θ from the notation for readability.

For SGD, we have $\mathbf{E}[g] = \nabla\mathcal{L}$ and $\mathbf{E}[g^T Q g] = \nabla\mathcal{L}^T Q \nabla\mathcal{L} + \text{tr}(Q\mathbf{cov}[g])$, which is a general fact for quadratic forms of random variables. For the stochastic QP the gradient covariance is $\mathbf{cov}[g] = \nu^2 QQ$, thus $\text{tr}(Q\mathbf{cov}[g]) = \nu^2\,\text{tr}(QQQ) = \nu^2 \sum_i \lambda_i^3$. Plugging everything into Eq. (12) yields

$$\mathcal{I}_{\text{SGD}} = \frac{(\nabla\mathcal{L}^T\nabla\mathcal{L})^2}{\nabla\mathcal{L}^T Q \nabla\mathcal{L} + \nu^2 \sum_{i=1}^d \lambda_i^3}. \tag{31}$$

For stochastic sign descent, $s = \text{sign}(g)$, we have $\mathbf{E}[s_i] = (2\rho_i - 1)\,\text{sign}(\nabla\mathcal{L}_i)$ and thus $\nabla\mathcal{L}^T\mathbf{E}[s] = \sum_{i=1}^d \nabla\mathcal{L}_i \mathbf{E}[s_i] = \sum_i (2\rho_i - 1)|\nabla\mathcal{L}_i|$. Regarding the denominator, it is

$$\begin{aligned}s^T Q s &\leq \left|\sum_{i=1}^d q_{ij} s_i s_j\right| \leq \sum_{i=1}^d |q_{ij}||s_i||s_j| \\ &= \sum_{i=1}^d |q_{ij}|,\end{aligned} \tag{32}$$

since $|s_i| = 1$. Further, by definition of $p_{\text{diag}}(Q)$, we have $\sum_{i=1}^d |q_{ij}| = p_{\text{diag}}(Q)^{-1} \sum_{i=1}^d |q_{ii}|$. Since Q is positive definite, its diagonal elements are positive, such that $\sum_{i=1}^d |q_{ii}| = \sum_{i=1}^d q_{ii} = \sum_{i=1}^d \lambda_i$. Plugging everything into Eq. (12) yields

$$\mathcal{I}_{\text{SSD}} \geq \frac{1}{2}\frac{\left(\sum_{i=1}^d (2\rho_i - 1)|\nabla\mathcal{L}(\theta)_i|\right)^2}{\sum_{i=1}^d \lambda_i} p_{\text{diag}}(Q). \tag{33}$$

B.2.2. Properties of $p_{\text{DIAG}}(Q)$

By writing $Q = \sum_k \lambda_k v_k v_k^T$ in its eigendecomposition with orthonormal eigenvectors $v_k \in \mathbb{R}^d$, we find

$$\begin{aligned}\sum_{i,j} |q_{ij}| &= \sum_{i,j}\left|\sum_k \lambda_k v_{k,i} v_{k,j}\right| \leq \sum_{i,j}\sum_k \lambda_k |v_{k,i} v_{k,j}| \\ &= \sum_k \lambda_k \left(\sum_i |v_{k,i}|\right)\left(\sum_j |v_{k,j}|\right) \\ &\leq \sum_k \lambda_k \|v_k\|_1^2.\end{aligned} \tag{34}$$

As mentioned before, $\sum_i |q_{ii}| = \sum_i \lambda_i$. Hence,

$$p_{\text{diag}}(Q) = \frac{\sum_i |q_{ii}|}{\sum_{i,j} |q_{ij}|} = \frac{\sum_i \lambda_i}{\sum_i \lambda_i \|v_i\|_1^2}. \tag{35}$$

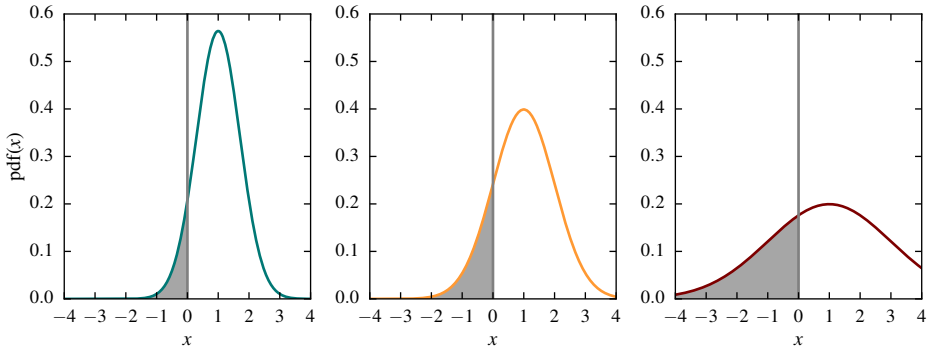

Figure 6. Probability density functions (pdf) of three Gaussian distributions, all with $\mu = 1$, but different variances $\sigma^2 = 0.5$ (left), $\sigma^2 = 1.0$ (middle), $\sigma^2 = 4.0$ (right). The shaded area under the curve corresponds to the probability that a sample from the distribution has the opposite sign than its mean. For the Gaussian distribution, this probability is uniquely determined by the fraction $\sigma/|\mu|$, as shown in Lemma 4.

As we have already seen, the best case arises if the eigenvectors are axis-aligned (diagonal Q), resulting in $\|v_i\|_1 = \|v_i\|_2 = 1$.

A worst case bound originates from the (tight) upper bound $\|w\|_1 \leq \sqrt{d}\|w\|_2$ for any $w \in \mathbb{R}^d$, which results in

$$p_{\text{diag}}(Q) \geq \frac{1}{d}. \tag{36}$$

We can get a rough intuition for the average case from the following consideration: For a d-dimensional random vector $w \sim \mathcal{N}(0, I)$, which corresponds to a random orientation, we have

$$\mathbf{E}[\|w\|_2] \approx \sqrt{d}, \quad \mathbf{E}[\|w\|_1] = d\sqrt{2/\pi}. \tag{37}$$

As a rough approximation, we can thus assume that a randomly-oriented vector will satisfy $\|w\|_1 \approx \sqrt{2d/\pi}\|w\|_2$. Plugging that in for the eigenvectors of Q in Eq. (35) yields an approximate average case value of

$$p_{\text{diag}}(Q) \approx \frac{\pi}{2d} \approx \frac{1.57}{d}. \tag{38}$$

B.3. Variance Adaptation Factors

Proof of Lemma 1. Using $\mathbf{E}[\hat{p}_i] = p_i$ and $\mathbf{E}[\hat{p}_i^2] = p_i^2 + \sigma_i^2$, we get

$$\begin{aligned}
\mathbf{E}[\|\gamma \odot \hat{p} - p\|_2^2] &= \sum_{i=1}^{d} \mathbf{E}[(\gamma_i \hat{p}_i - p_i)^2] \\
&= \sum_{i=1}^{d} \left(\gamma_i^2 \mathbf{E}[\hat{p}_i^2] - 2\gamma_i p_i \mathbf{E}[\hat{p}_i] + p_i^2\right) \\
&= \sum_{i=1}^{d} \left(\gamma_i^2 (p_i^2 + \sigma_i^2) - 2\gamma_i p_i^2 + p_i^2\right).
\end{aligned} \tag{39}$$

Setting the derivative w.r.t. γ_i to zero, we find the optimal choice

$$\gamma_i = \frac{p_i^2}{p_i^2 + \sigma_i^2}. \tag{40}$$

For the second part, using $\mathbf{E}[\text{sign}(\hat{p}_i)] = (2\rho_i - 1)\,\text{sign}(p_i)$ and $\text{sign}(\cdot)^2 = 1$, we get

$$\begin{aligned}
&\mathbf{E}[\|\gamma \odot \text{sign}(\hat{p}) - \text{sign}(p)\|_2^2] \\
&= \sum_{i=1}^{d} \mathbf{E}\left[(\gamma_i \,\text{sign}(\hat{p}_i) - \text{sign}(p_i))^2\right] \\
&= \sum_{i=1}^{d} \left(\gamma_i^2 - 2\gamma_i \,\text{sign}(p_i)\mathbf{E}[\text{sign}(\hat{p}_i)] + 1\right) \\
&= \sum_{i=1}^{d} \left(\gamma_i^2 - 2\gamma_i(2\rho_i - 1) + 1\right)
\end{aligned} \tag{41}$$

and easily find the optimal choice

$$\gamma_i = 2\rho_i - 1. \tag{42}$$

by setting the derivative to zero. □

B.4. Convergence of Idealized SVAG

We prove the convergence results for idealized variance-adapted stochastic gradient descent (Theorem 1). The stochastic optimizer generates a discrete stochastic process $\{\theta_t\}_{t \in \mathbb{N}_0}$. We denote as $\mathbf{E}_t[\cdot] = \mathbf{E}[\cdot|\theta_t]$ the conditional expectation given a realization of that process up to time step t. Recall that $\mathbf{E}[\mathbf{E}_t[\cdot]] = \mathbf{E}[\cdot]$.

We first show the following Lemma.

Lemma 5. *Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be μ-strongly convex and L-smooth. Denote as $\theta_* := \arg\min_{\theta \in \mathbb{R}^d} f(\theta)$ the unique minimizer and $f_* = f(\theta_*)$. Then, for any $\theta \in \mathbb{R}^d$,*

$$\frac{2L^2}{\mu}(f(\theta) - f_*) \geq \|\nabla f(\theta)\|^2 \geq 2\mu(f(\theta) - f_*). \tag{43}$$

Proof. Regarding the first inequality, we use $\nabla f(\theta_*) = 0$ and the Lipschitz continuity of $\nabla f(\cdot)$ to get $\|\nabla f(\theta)\|^2 = \|\nabla f(\theta) - \nabla f(\theta_*)\|^2 \leq L^2\|\theta - \theta_*\|^2$. Using strong convexity, we have $f(\theta) \geq f_* + \nabla f(\theta_*)^T(\theta - \theta_*) + (\mu/2)\|\theta - \theta_*\|^2 = f_* + (\mu/2)\|\theta - \theta_*\|^2$. Plugging the two inequalities together yields the desired inequality.

The second inequality arises from strong convexity, by minimizing both sides of

$$f(\theta') \geq f(\theta) + \nabla f(\theta)^T(\theta' - \theta) + \frac{\mu}{2}\|\theta' - \theta\|^2 \quad (44)$$

w.r.t. θ'. The left-hand side obviously has minimal value f_*. For the right-hand side, we set its derivative, $\nabla f(\theta) + \mu(\theta' - \theta)$, to zero to find the minimizer $\theta' = \theta - \nabla f(\theta)/\mu$. Plugging that back in yields the minimal value $f(\theta) - \|\nabla f(\theta)\|/(2\mu)$. □

Proof of Theorem 1. Using the Lipschitz continuity of ∇f, we can bound $f(\theta + \Delta\theta) \leq f(\theta) + \nabla f(\theta)^T\Delta\theta + \frac{L}{2}\|\Delta\theta\|^2$. Hence,

$$\begin{aligned}
&\mathbf{E}_t[f_{t+1}] \\
&\leq f_t - \alpha\mathbf{E}_t[\nabla f_t^T(\gamma_t \odot g_t)] + \frac{L\alpha^2}{2}\mathbf{E}_t[\|\gamma_t \odot g_t\|^2] \\
&= f_t - \frac{1}{L}\sum_{i=1}^{d}\gamma_{t,i}\nabla f_{t,i}\mathbf{E}_t[g_{t,i}] + \frac{1}{2L}\sum_{i=1}^{d}\gamma_{t,i}^2\mathbf{E}_t[g_{t,i}^2] \\
&= f_t - \frac{1}{L}\sum_{i=1}^{d}\gamma_{t,i}\nabla f_{t,i}^2 + \frac{1}{2L}\sum_{i=1}^{d}\gamma_{t,i}^2(\nabla f_{t,i}^2 + \sigma_{t,i}^2).
\end{aligned} \quad (45)$$

Plugging in the definition $\gamma_{t,i} = \nabla f_{t,i}^2/(\nabla f_{t,i}^2 + \sigma_{t,i}^2)$ and simplifying, we get

$$\mathbf{E}_t[f_{t+1}] \leq f_t - \frac{1}{2L}\sum_{i=1}^{d}\frac{\nabla f_{t,i}^4}{\nabla f_{t,i}^2 + \sigma_{t,i}^2}. \quad (46)$$

This shows that $\mathbf{E}_t[f_{t+1}] \leq f_t$. Defining $e_t := f_t - f_*$, this implies

$$\mathbf{E}[e_{t+1}] = \mathbf{E}[\mathbf{E}_t[e_{t+1}]] \leq \mathbf{E}[e_t] \quad (47)$$

and consequently, by iterating backwards, $\mathbf{E}[e_t] \leq \mathbf{E}[e_0] = e_0$ for all t. Next, using the discrete version of Jensen's inequality[5] we find

$$\sum_{i=1}^{d}\frac{\nabla f_{t,i}^4}{\nabla f_{t,i}^2 + \sigma_{t,i}^2} \geq \frac{\|\nabla f_t\|^4}{\|\nabla f_t\|^2 + \sum_{i=1}^{d}\sigma_{t,i}^2}. \quad (48)$$

[5] Jensen's inequality states that, for a real convex function ϕ, numbers $x_i \in \mathbb{R}$, and positive weights $a_i \in \mathbb{R}_+$ with $\sum_i a_i = 1$, we have $\sum_i a_i\phi(x_i) \geq \phi\left(\sum_i a_i x_i\right)$. We apply it here to the convex function $\phi(x) = 1/x, x > 0$, with $x_i := \frac{\nabla f_{t,i}^2 + \sigma_{t,i}^2}{\nabla f_{t,i}^2}$ and $a_i := \frac{\nabla f_{t,i}^2}{\|\nabla f_t\|^2}$.

Using the assumption $\sum_{i=1}^{d}\sigma_{t,i}^2 \leq c_v|\nabla f_t\|^2 + M_v$ in the denominator, we obtain

$$\frac{\|\nabla f_t\|^4}{\|\nabla f_t\|^2 + \sum_{i=1}^{d}\sigma_{t,i}^2} \geq \frac{\|\nabla f_t\|^4}{(1 + c_v)\|\nabla f_t\|^2 + M_v}. \quad (49)$$

Using Lemma 5, we have

$$\frac{2L^2}{\mu}e_t \geq \|\nabla f_t\|^2 \geq 2\mu e_t \quad (50)$$

and can further bound

$$\begin{aligned}
\frac{\|\nabla f_t\|^4}{(1 + c_v)\|\nabla f_t\|^2 + M_v} &\geq \frac{4\mu^2 e_t^2}{\frac{2(1+c_v)L^2}{\mu}e_t + M_v} \\
&=: \frac{c_1 e_t^2}{c_2 e_t + c_3},
\end{aligned} \quad (51)$$

where the last equality defines the (positive) constants c_1, c_2 and c_3. Combining Eqs. (48), (49) and (51), inserting in (46), and subtracting f_* from both sides, we obtain

$$\mathbf{E}_t[e_{t+1}] \leq e_t - \frac{1}{2L}\frac{c_1 e_t^2}{c_2 e_t + c_3}, \quad (52)$$

and, consequently, by taking expectations on both sides,

$$\begin{aligned}
\mathbf{E}[e_{t+1}] &\leq \mathbf{E}[e_t] - \frac{1}{2L}\mathbf{E}\left[\frac{c_1 e_t^2}{c_2 e_t + c_3}\right] \\
&\leq \mathbf{E}[e_t] - \frac{1}{2L}\frac{c_1\mathbf{E}[e_t]^2}{c_2\mathbf{E}[e_t] + c_3}
\end{aligned} \quad (53)$$

where the last step is due to Jensen's inequality applied to the convex function $\phi(x) = \frac{c_1 x^2}{c_2 x + c_3}$. Using $\mathbf{E}[e_t] \leq e_0$ in the denominator and introducing the shorthand $\bar{e}_t := \mathbf{E}[e_t]$, we get

$$\bar{e}_{t+1} \leq \bar{e}_t - c\bar{e}_t^2 = \bar{e}_t(1 - c\bar{e}_t), \quad (54)$$

with $c := c_1/(2L(c_2 e_0 + c_3)) > 0$. To conclude the proof, we will show that this implies $\bar{e}_t \in \mathcal{O}(\frac{1}{t})$. Without loss of generality, we assume $\bar{e}_{t+1} > 0$ and obtain

$$\begin{aligned}
\bar{e}_{t+1}^{-1} &\geq \bar{e}_t^{-1}(1 - c\bar{e}_t)^{-1} \geq \bar{e}_t^{-1}(1 + c\bar{e}_t) \\
&= \bar{e}_t^{-1} + c,
\end{aligned} \quad (55)$$

where the second step is due to the simple fact that $(1 - x)^{-1} \geq (1 + x)$ for any $x \in [0, 1)$. Summing this inequality over $t = 0, \ldots, T - 1$ yields $\bar{e}_T^{-1} \geq e_0^{-1} + Tc$ and, thus,

$$T\bar{e}_T \leq \left(\frac{1}{Te_0} + c\right)^{-1} \overset{T\to\infty}{\longrightarrow} \frac{1}{c} < \infty, \quad (56)$$

which shows that $\bar{e}_t \in \mathcal{O}(\frac{1}{t})$. □

B.5. Gradient Variance Estimates via Moving Averages

We proof Eq. (20). Iterating the recursive formula for $\tilde{m}_t$ backwards, we get

$$m_t = \sum_{s=0}^{t} \underbrace{\frac{1-\beta_1}{1-\beta_1^{t+1}}\beta_1^{t-s}}_{=:c(\beta_1,t,s)} g_s, \tag{57}$$

with coefficients $c(\beta_1, t, s)$ summing to one by the geometric sum formula, making m_t a convex combination of stochastic gradients. Likewise, $v_t = \sum_{s=0}^{t} c(\beta_2, t, s) g_s^2$ is a convex combination of squared stochastic gradients. Hence,

$$\begin{aligned} \mathbf{E}[m_{t,i}] &= \sum c(\beta, t, s)\mathbf{E}[g_{s,i}], \\ \mathbf{E}[v_{t,i}] &= \sum c(\beta, t, s)\mathbf{E}[g_{s,i}^2]. \end{aligned} \tag{58}$$

Assumption 1 thus necessarily implies $\mathbf{E}[g_{s,i}] \approx \nabla\mathcal{L}_{t,i}$ and $\mathbf{E}[g_{s,i}^2] \approx \nabla\mathcal{L}_{t,i}^2 + \sigma_{t,i}^2$. (This will of course be utterly wrong for gradient observations that are far in the past, but these won't contribute significantly to the moving average.) It follows that

$$\begin{aligned} \mathbf{E}[m_{t,i}^2] &= \mathbf{E}[m_{t,i}]^2 + \mathbf{var}[m_{t,i}] \\ &= \nabla\mathcal{L}_{t,i}^2 + \sum_{s=0}^{t} c(\beta, t, s)^2 \, \mathbf{var}[g_{s,i}] \\ &= \nabla\mathcal{L}_{t,i}^2 + \sigma_{t,i}^2 \sum_{s=0}^{t} c(\beta, t, s)^2, \end{aligned} \tag{59}$$

where the second step is due to the fact that g_s and $g_{s'}$ are stochastically independent for $s \neq s'$. The last term evaluates to

$$\begin{aligned} \rho(\beta, t) &:= \sum_{s=0}^{t} c(\beta, t, s)^2 = \sum_{s=0}^{t} \left(\frac{1-\beta}{1-\beta^{t+1}}\beta^{t-s}\right)^2 \\ &= \frac{(1-\beta)^2}{(1-\beta^{t+1})^2} \sum_{k=0}^{t} (\beta^2)^k \\ &= \frac{(1-\beta)^2}{(1-\beta^{t+1})^2} \frac{1-(\beta^2)^{t+1}}{1-\beta^2} \\ &= \frac{(1-\beta)(1-\beta)}{(1-\beta^{t+1})(1-\beta^{t+1})} \frac{(1-\beta^{t+1})(1+\beta^{t+1})}{(1-\beta)(1+\beta)} \\ &= \frac{(1-\beta)(1+\beta^{t+1})}{(1+\beta)(1-\beta^{t+1})}, \end{aligned} \tag{60}$$

where the fourth step is another application of the geometric sum formula, and the fifth step uses $1-x^2 = (1-x)(1+x)$. Note that

$$\rho(\beta, t) \to \frac{1-\beta}{1+\beta} \quad (t \to \infty), \tag{61}$$

such that $\rho(\beta, t)$ is uniquely defined by β in the long term.

As an interesting side note, the division by $1 - \rho(\beta, t)$ in Eq. (22) is the analogon to Bessel's correction (the use of $n-1$ instead of n in the classical sample variance) for the case where we use moving averages instead of arithmetic means.

B.6. Connection to Generalization

Proof of Lemma 3. Like in the proof of Lemma 3.1 in Wilson et al. (2017), we inductively show that $\theta_t = \lambda_t \operatorname{sign}(X^T y)$ with a scalar λ_t. This trivially holds for $\theta_0 = 0$. Assume that the assertion holds for all $s \leq t$. Then

$$\begin{aligned} \nabla R(\theta_t) &= \frac{1}{n} X^T (X\theta_t - y) \\ &= \frac{1}{n} X^T (\lambda_t X \operatorname{sign}(X^T y) - y) \\ &= \frac{1}{n} X^T (\lambda_t c y - y) = \frac{1}{n}(\lambda_t c - 1) X^T y, \end{aligned} \tag{62}$$

where the first step is the gradient of the objective (Eq. 26), the second step uses the inductive assumption, and the third step uses the assumption $X \operatorname{sign}(X^T y) = cy$. Now, plugging Eq. (62) into the update rule, we find

$$\begin{aligned} \theta_{t+1} &= \theta_t - \alpha \operatorname{sign}(\nabla R(\theta_t)) \\ &= \lambda_t \operatorname{sign}(X^T y) - \alpha \operatorname{sign}((\lambda_t c - 1) X^T y) \\ &= (\lambda_t - \alpha \operatorname{sign}(\lambda_t c - 1)) \operatorname{sign}(X^T y). \end{aligned} \tag{63}$$

Hence, the assertion holds for $t+1$. □

C. Alternative Methods

C.1. SVAG

M-SVAG applies variance adaptation to the update direction m_t, resulting in the variance adaptation factors Eq. 25. We can also update in direction g_t and choose the appropriate estimated variance adaptation factors, resulting in an implementation of SVAG without momentum. We have already derived the necessary variance adaptation factors en route to those for the momentum variant, see Eq. (23) in §4.2. Pseudo-code is provided in Alg. 3. It differs from M-SVAG only in the last two lines.

C.2. Variants of ADAM

This paper interpreted ADAM as variance-adapted M-SSD. The experiments in the main paper used a standard implementation of ADAM as described by Kingma & Ba (2015). However, in the derivation of our implementation of M-SVAG, we have made multiple adjustments regarding the estimation of variance adaptation factors which correspondingly apply to the sign case. Specifically, this concerns:

Algorithm 3 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**
  $\tilde{m} \leftarrow \beta\tilde{m} + (1-\beta)g(\theta)$, $\quad \tilde{v} \leftarrow \beta\tilde{v} + (1-\beta)g(\theta)^2$
  $m \leftarrow (1-\beta^{t+1})^{-1}\tilde{m}$, $\quad v \leftarrow (1-\beta^{t+1})^{-1}\tilde{v}$
  $s \leftarrow (1-\rho(\beta,t))^{-1}(v-m^2)$
  $\gamma \leftarrow m^2/(m^2+s)$
  $\theta \leftarrow \theta - \alpha(\gamma \odot g)$
end for

- The use of the same moving average constant for the first and second moment ($\beta_1 = \beta_2 = \beta$).
- The bias correction in the gradient variance estimate, see Eq. (22).
- The adjustment of the variance adaptation factors for the momentum case, see §4.3.
- The omission of a constant offset ε in the denominator.

Applying these adjustment to the sign case gives rise to a variant of the original ADAM algorithm, which we will refer to as ADAM*. Pseudo-code is provided in Alg. 4. Note that we use the variance adaptation factors $(1+\eta)^{-1/2}$ and *not* the optimal ones derived in §3.1, which would under the Gaussian assumption be $\mathrm{erf}[(\sqrt{2}\eta)^{-1}]$. We initially experimented with both variants and found them to perform almost identically, which is not surprising given how similar the two are (see Fig. 3). We thus stuck with the first option for direct correspondence with the original ADAM and to avoid the cumbersome error function.

In analogy to SVAG versus M-SVAG, we could also define a variance-adapted version stochastic sign descent *without* momentum, i.e., using the base update direction $\mathrm{sign}(g_t)$. We did not explore this further in this work.

Algorithm 4 ADAM*

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

C.3. Experiments

We tested SVAG as well as ADAM* with and without momentum on the problems (P2) and (P3) from the main paper. Results are shown in Figure 7.

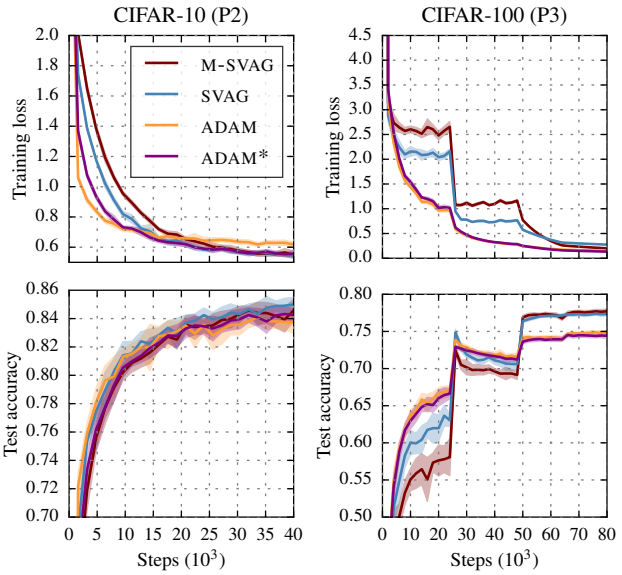

Figure 7. Experimental results for SVAG and ADAM*. The plot is set-up like Fig. 5.

We observe that SVAG performs better than M-SVAG on (P2). On (P3), it makes faster initial progress but later plateaus, leading to slightly worse outcomes in both training loss and test accuracy. SVAG is a viable alternative. In future work, it will be interesting to apply SVAG to problems where SGD outperforms M-SGD.

Next, we compare ADAM* to the original ADAM algorithm. In the CIFAR-100 example (P3) the two methods are on par. On (P2), ADAM is marginally faster in the early stages of the the optimization process. ADAM* quickly catches up and reaches lower minimal training loss values. We conclude that the adjustments to the variance adaptation factors derived in §4 do have a positive effect.

D. Mini-Batch Gradient Variance Estimates

In the main text, we have discussed estimation of gradient variances via moving averages of the past gradient observations. An alternative gradient variance estimate can be obtained locally, within a single mini-batch. The individual gradients $\nabla\ell(\theta; x_k)$ in a mini-batch are iid random variables and $\mathbf{var}[g(\theta)] = |\mathcal{B}|^{-1}\mathbf{var}_{k\sim\mathcal{U}([M])}[\nabla\ell(\theta; x_k)]$. We can thus estimate $g(\theta)$'s variances by computing the sample variance of the $\{\nabla\ell(\theta; x_k)\}_{k\in\mathcal{B}}$, then scaling by $|\mathcal{B}|^{-1}$,

$$\hat{s}^{\mathrm{mb}}(\theta) = \frac{1}{|\mathcal{B}|}\left(\frac{1}{|\mathcal{B}|-1}\sum_{k\in\mathcal{B}}\nabla\ell(\theta; x_k)^2 - g(\theta)^2\right). \quad (64)$$

Several recent papers (Mahsereci & Hennig, 2015; Balles et al., 2017b) have used this variance estimate for other

aspects of stochastic optimization. In contrast to the moving average-based estimators, this is an unbiased estimate of the *local* gradient variance. The (non-trivial) implementation of this estimator for neural networks is described in Balles et al. (2017a).

D.1. M-SVAG with Mini-Batch Estimates

We explored a variant of M-SVAG which use mini-batch gradient variance estimates. The local variance estimation allows for a theoretically more pleasing treatment of the variance of the update direction m_t. Starting from the formulation of m_t in Eq. (57) and considering that g_s and $g_{s'}$ are stochastically independent for $s \neq s'$, we have

$$\mathbf{var}[m_t] = \sum_{s=0}^{t} \left(\frac{1-\beta}{1-\beta^{t+1}} \beta^{t-s} \right)^2 \mathbf{var}[g_s]. \tag{65}$$

Given that we now have access to a true, local, unbiased estimate of $\mathbf{var}[g_s]$, we can estimate $\mathbf{var}[m_t]$ by

$$\bar{s}_t := \sum_{s=0}^{t} \left(\frac{1-\beta}{1-\beta^{t+1}} \beta^{t-s} \right)^2 \hat{s}^{\text{mb}}(\theta_s). \tag{66}$$

It turns out that we can track this quantity with another exponential moving average: It is $\bar{s}_t = \rho(\beta, t) r_t$ with

$$\tilde{r}_t = \beta^2 \tilde{r}_{t-1} + (1-\beta^2)\hat{s}_t^{\text{mb}}, \quad r_t = \frac{\tilde{r}_t}{1-(\beta^2)^{t+1}}. \tag{67}$$

This can be shown by iterating Eq. (67) backwards and comparing coefficients with Eq. (66). The resulting mini-batch variant of M-SVAG is presented in Algorithm 5.

Note that mini-batch gradient variance estimates could likewise be used for the alternative methods discussed in §C. We do not explore this further in this paper.

D.2. Experiments

We tested the mini-batch variant of M-SVAG on the problems (P1) and (P2) from the main text and compared it to the moving average version. Results are shown in Figure 8. The two algorithms have almost identical performance.

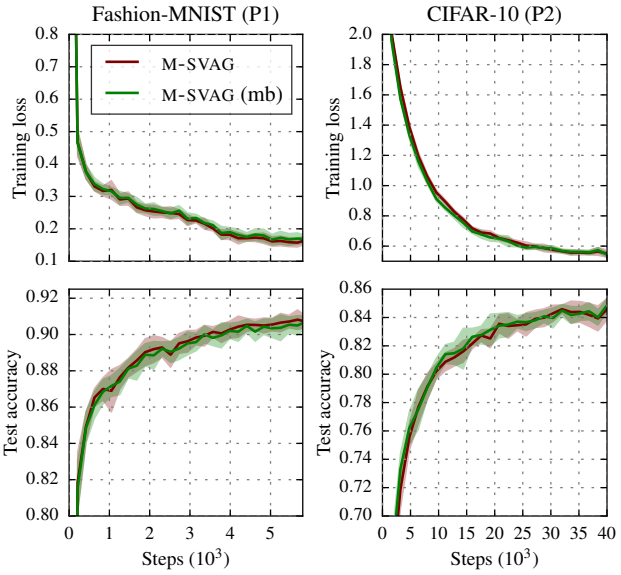

Figure 8. Experimental results for the mini-batch variant of M-SVAG (marked "mb" in the legend). The plot is set-up like Fig. 5.

Algorithm 5 M-SVAG with mini-batch variance estimate

Input: $\theta_0 \in \mathbb{R}^d$, $\alpha > 0$, $\beta \in [0,1]$, $T \in \mathbb{N}$
Initialize $\theta \leftarrow \theta_0$, $\tilde{m} \leftarrow 0$, $\tilde{r} \leftarrow 0$
for $t = 0, \dots, T-1$ **do**
  Compute mini-batch gradient $g(\theta)$ and variance $\hat{s}^{\text{mb}}(\theta)$
  $\tilde{m} \leftarrow \beta\tilde{m} + (1-\beta)g(\theta)$, $\quad \tilde{r} \leftarrow \beta^2\tilde{r} + (1-\beta^2)\hat{s}^{\text{mb}}(\theta)$
  $m \leftarrow (1-\beta^{t+1})^{-1}\tilde{m}$, $\quad r \leftarrow (1-\beta^{2(t+1)})^{-1}\tilde{r}$
  $\gamma \leftarrow m^2/(m^2 + \rho(\beta,t)r)$
  $\theta \leftarrow \theta - \alpha(\gamma \odot m)$
end for