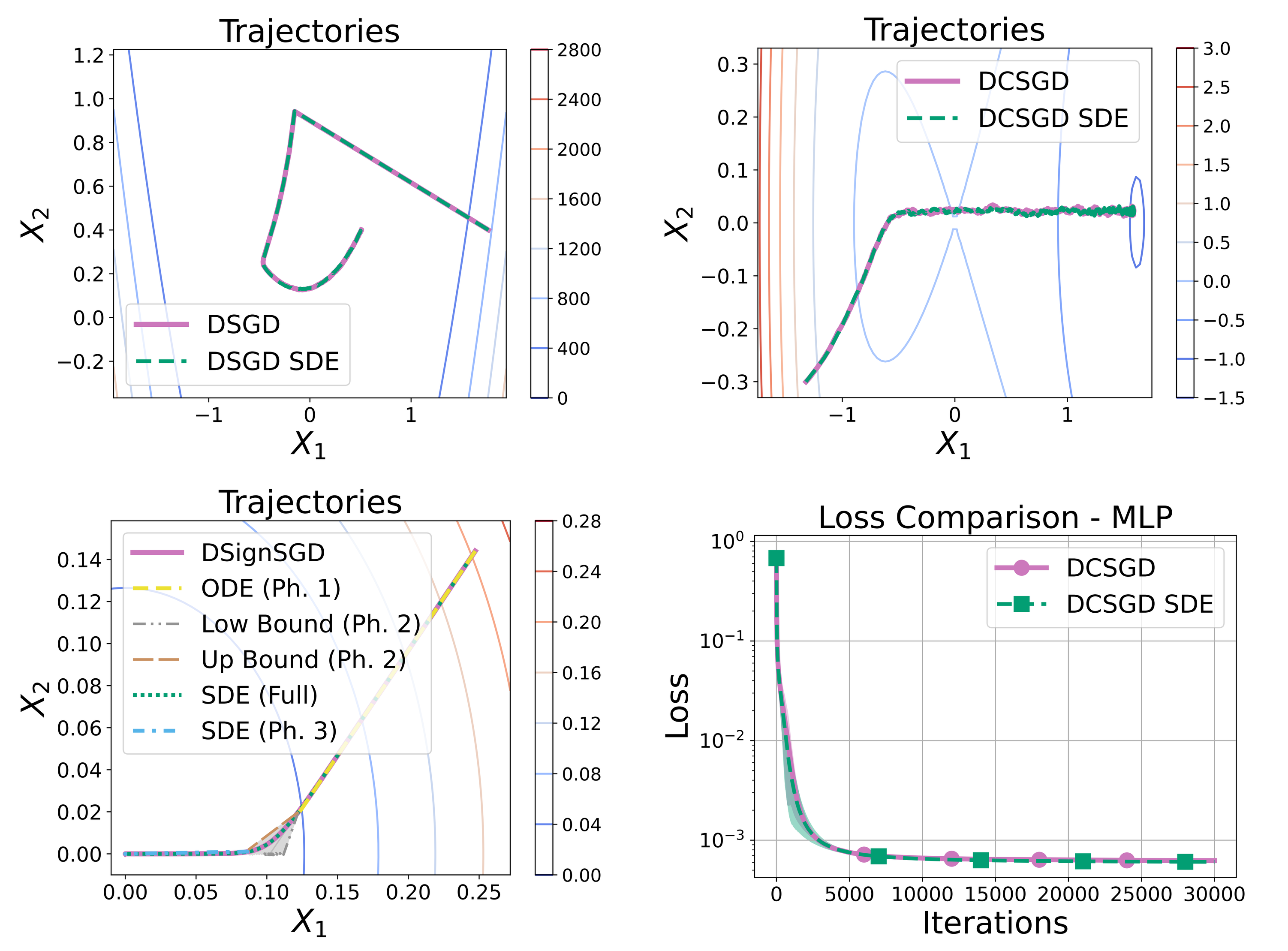
Unbiased and Sign Compression in Distributed Learning: Comparing Noise Resilience via SDEs

University of Basel

Enea Monzio Compagnoni, Rustem Islamov, Frank Norbert Proske, Aurelien Lucchi

UNIVERSITAS OSLOENSIS MDCCCXI

Visual Intuition - SDEs do Track the Optimizers



Definitions

Distributed Unbiased Compressed SGD is

$$x_{k+1} = x_k - \frac{\eta}{N}\sum_{i=1}^{N} \mathcal{C}_{\xi_i}\left(\nabla f_{\gamma_i}(x_k)\right), \tag{1}$$

where the stochastic compressors $\mathcal{C}_{\xi_i}$ are independent and

1. $\mathbb{E}_{\xi_i}\left[\mathcal{C}_{\xi_i}(x)\right] = x$;
2. $\mathbb{E}_{\xi_i}\left[\|\mathcal{C}_{\xi_i}(x) - x\|_2^2\right] \leq \omega_i \|x\|_2^2$ for some compression rates $\omega_i \geq 0$.

Distributed SignSGD is a *biased* compression method with update rule

$$x_{k+1} = x_k - \frac{\eta}{N}\sum_{i=1}^{N} \text{sign}(\nabla f_{\gamma_i}(x_k)). \tag{2}$$

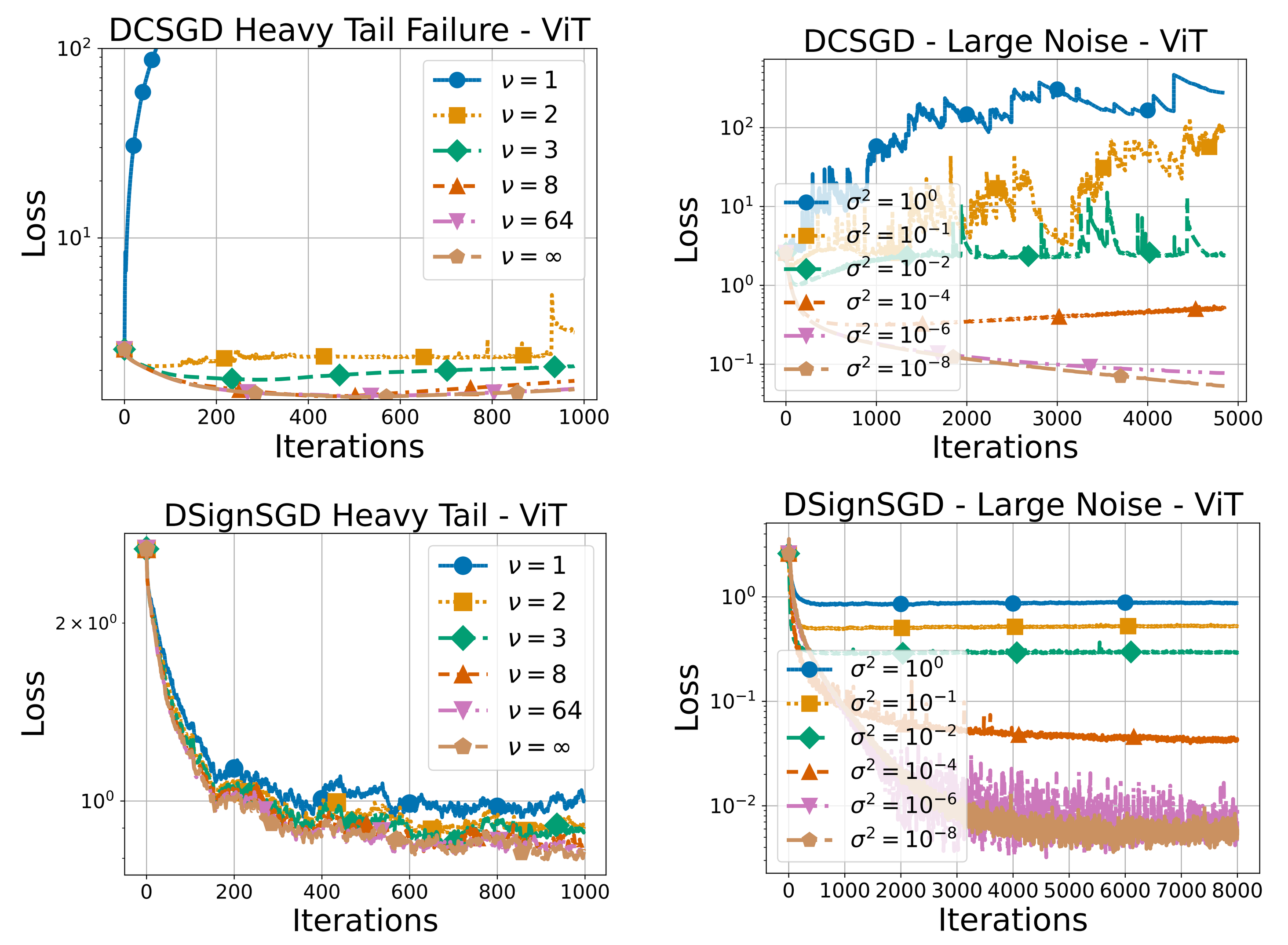
Problem of Interest

1. How do gradient noise and compression interact?
2. Which method is more resilient to large, possibly heavy-tailed noise?
3. Are there any scaling laws designed for Distributed Learning?

Contributions

1. First SDE formulation for DCSGD and DSignSGD.
2. DCSGD is highly sensitive to heavy-tailed noise, while DSignSGD is robust;
3. New scaling rules for hyperparameter tuning;
4. Empirical validation across MLP, ResNet, ViT, and GPT2.

Noise Resilience: Empirical Observation



SDEs

Theorem 1 (DCSGD). *For $\Phi_{\xi_i,\gamma_i}(x) := \mathcal{C}_{\xi_i}\left(\nabla f_{\gamma_i}(x)\right) - \nabla f_{\gamma_i}(x)$, the SDE of DCSGD is*

$$dX_t = -\nabla f(X_t)dt + \sqrt{\frac{\eta}{N}}\sqrt{\tilde{\Sigma}(X_t)}dW_t, \tag{3}$$

where

$$\tilde{\Sigma}(x) = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbb{E}_{\xi_i\gamma_i}\left[\Phi_{\xi_i,\gamma_i}(x)\Phi_{\xi_i,\gamma_i}(x)^\top\right] + \Sigma_i(x)\right). \tag{4}$$

Theorem 2 (DSignSGD). *The SDE of DSignSGD is*

$$dX_t = -\frac{2}{N}\sum_{i=1}^{N}\Xi_\nu\left(\Sigma_i^{-\frac{1}{2}}\nabla f(X_t)\right)dt + \sqrt{\frac{\eta}{N}}\sqrt{\tilde{\Sigma}(X_t)}dW_t. \tag{5}$$

where

$$\tilde{\Sigma}(X_t) := I_d - \frac{4}{N}\sum_{i=1}^{N}\left(\Xi_\nu\left(\Sigma_i^{-\frac{1}{2}}\nabla f(X_t)\right)\right)^2. \tag{6}$$

Noise Resilience: Theoretical Justification

Theorem 3 (DCSGD). *If f is μ-PL, L-smooth, Tr$(\Sigma_i(x)) < d\sigma^2$, and $\Delta := 1 - \frac{\eta L^2 \omega}{2\mu N}$, then*

$$\mathbb{E}\left[f(X_t) - f(X_*)\right] \leq (f(X_0) - f(X_*))e^{-2\mu\Delta t} + \left(1 - e^{-2\mu\Delta t}\right)\frac{\eta L d}{4\mu N} \times \frac{\sigma^2}{B} \times \frac{1+\omega}{1 - \frac{\eta L^2\omega}{2\mu N}}.$$

Theorem 4 (DSignSGD). *If the gradient noise is $Z \sim \sigma t_\nu(0, I_d)$ for $\Delta := \frac{\ell_\nu\sqrt{B}}{\sigma}$,*

$$\mathbb{E}\left[f(X_t) - f(X_*)\right] \leq (f(X_0) - f(X_*))e^{-2\mu\Delta t} + \left(1 - e^{-2\mu\Delta t}\right)\frac{\eta L d}{4\mu N} \times \frac{\sigma}{\sqrt{B}} \times \frac{1}{\ell_\nu}.$$

Scaling Laws: Preserving Performance

1. **Learning Rate:** $\eta \to \kappa\eta$;
2. **Batch Size:** $B \to \delta B$;
3. **Compression Rate:** $\omega \to \beta\omega$;
4. **Client Number:** $N \to \alpha N$.

? $\longrightarrow$ Can we recover **Uncompressed** DSGD(η, B, N)?

Scaling Rule	Implication
$\alpha = 1 + \beta\omega$	CR $\uparrow \implies$ Agents $\uparrow$
$\alpha = \kappa(1+\omega)$	LR $\uparrow \implies$ Agents $\uparrow$
$\alpha = \frac{1+\omega}{\delta}$	BS $\downarrow \implies$ Agents $\uparrow$
$\kappa = \frac{1}{1+\beta\omega}$	CR $\uparrow \implies$ LR $\downarrow$
$\delta = 1 + \beta\omega$	CR $\uparrow \implies$ BS $\uparrow$
$\kappa = \frac{\delta}{1+\omega}$	BS $\uparrow \implies$ LR $\uparrow$

For DSignSGD, it is enough to ensure that $\frac{\kappa}{\alpha\sqrt{\delta}}$ to preserve the performance.

Validation on GPT2-like Model