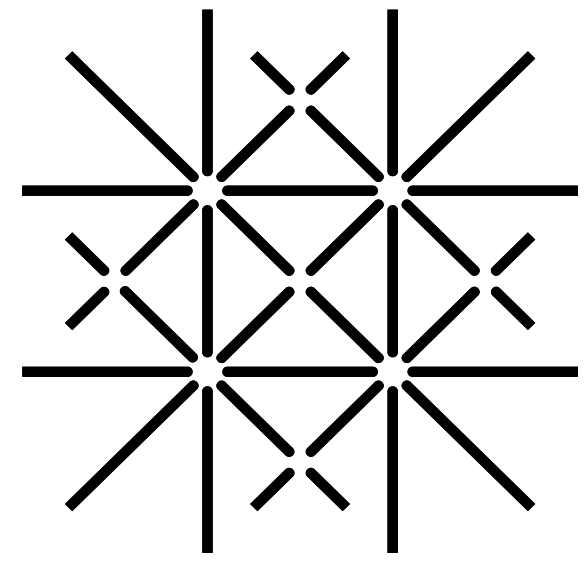
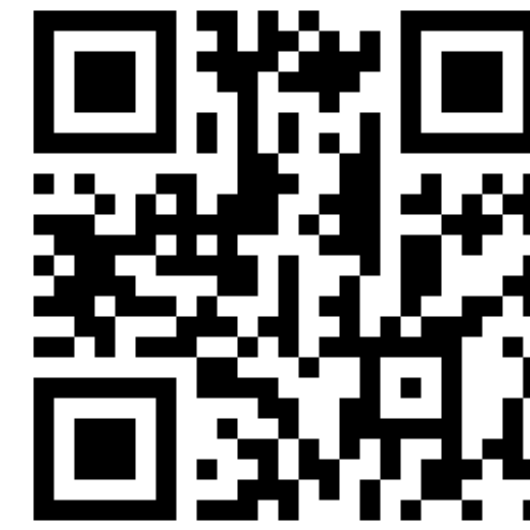




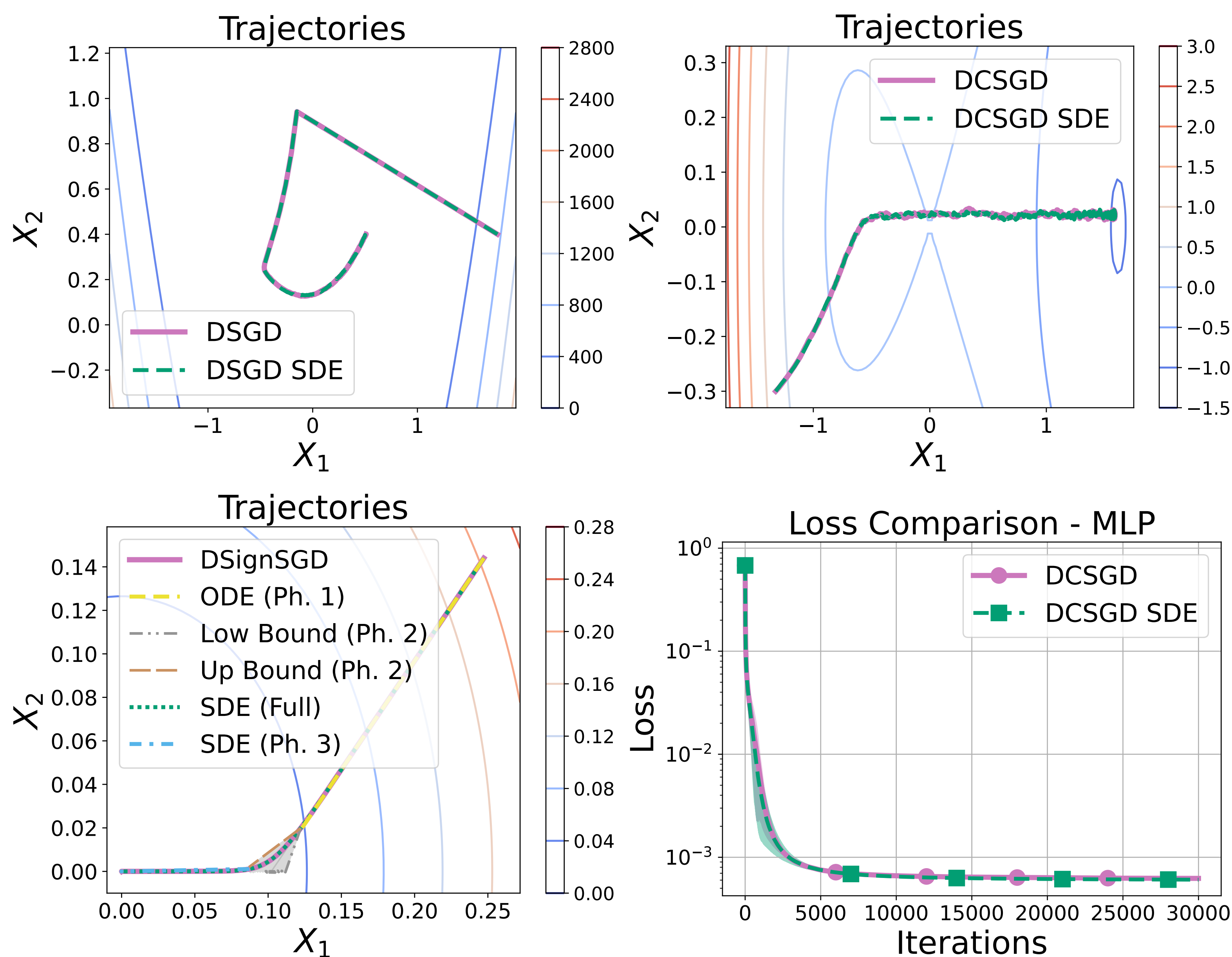
University
of Basel

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

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



Visual Intuition - SDEs do Track the Optimizers



Definitions

Distributed Unbiased Compressed SGD is

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

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

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

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

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

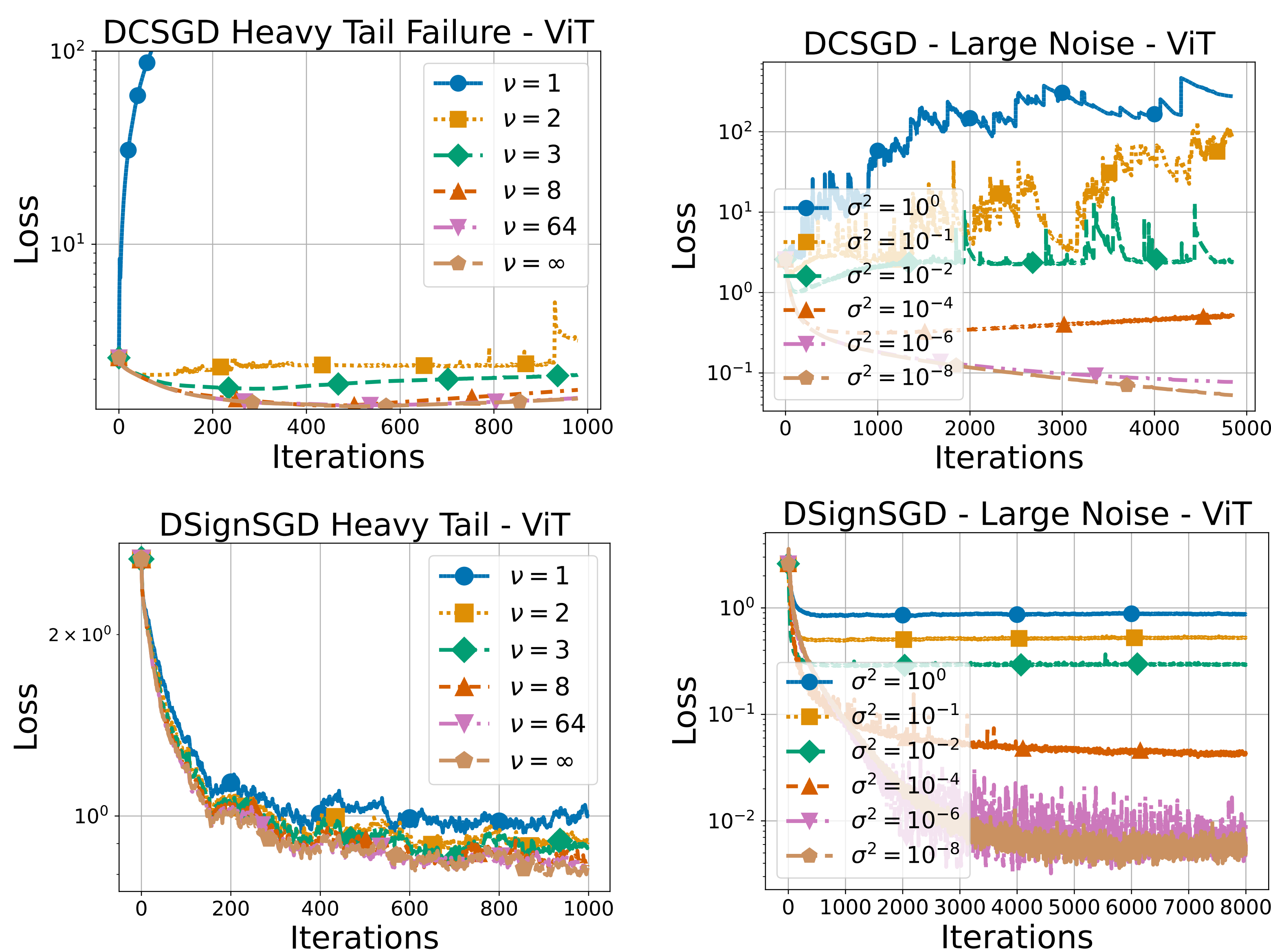
Problem of Interest

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 $\Phi_{\xi_i, \gamma_i}(x) := \mathcal{C}_{\xi_i}(\nabla f_{\gamma_i}(x)) - \nabla f_{\gamma_i}(x)$, the SDE is

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

where

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

Theorem 2 (DSignSGD). The SDE of DSignSGD is

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

where

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

Noise Resilience: Theoretical Justification

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

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

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

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

Scaling Laws: Preserving Performance

1. **Learning Rate:** $\eta \rightarrow \kappa\eta$;
 2. **Batch Size:** $B \rightarrow \delta B$;
 3. **Compression Rate:** $\omega \rightarrow \beta\omega$;
 4. **Client Number:** $N \rightarrow \alpha N$.
- Can we recover **Uncompressed** DSGD(η, B, N)?

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

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

Validation on GPT2-like Model

