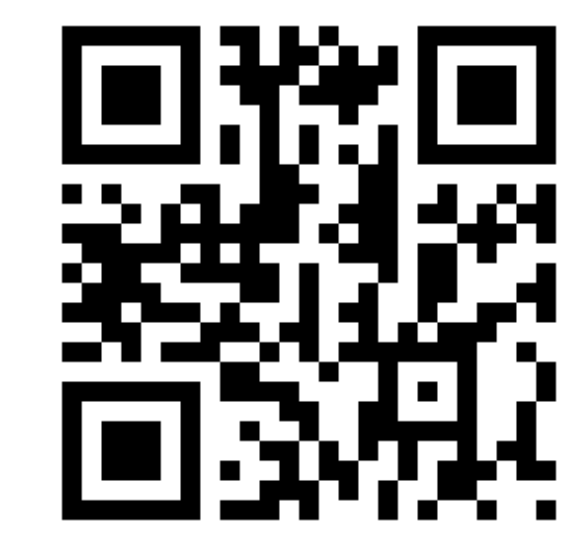
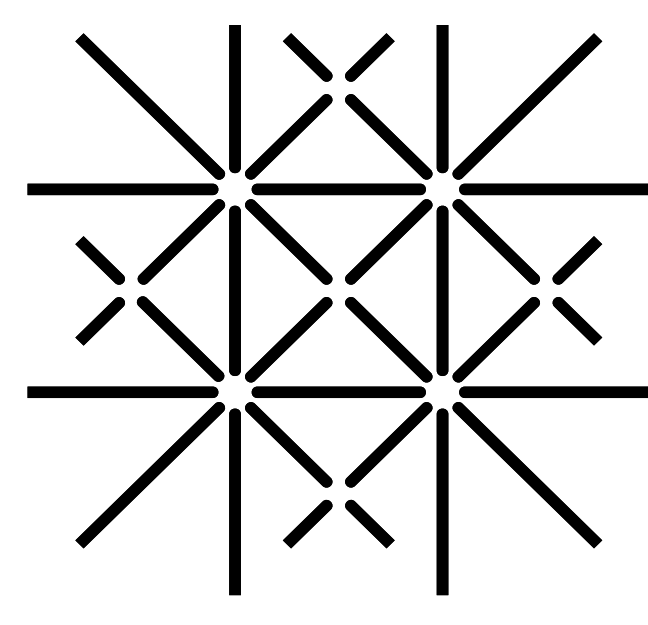
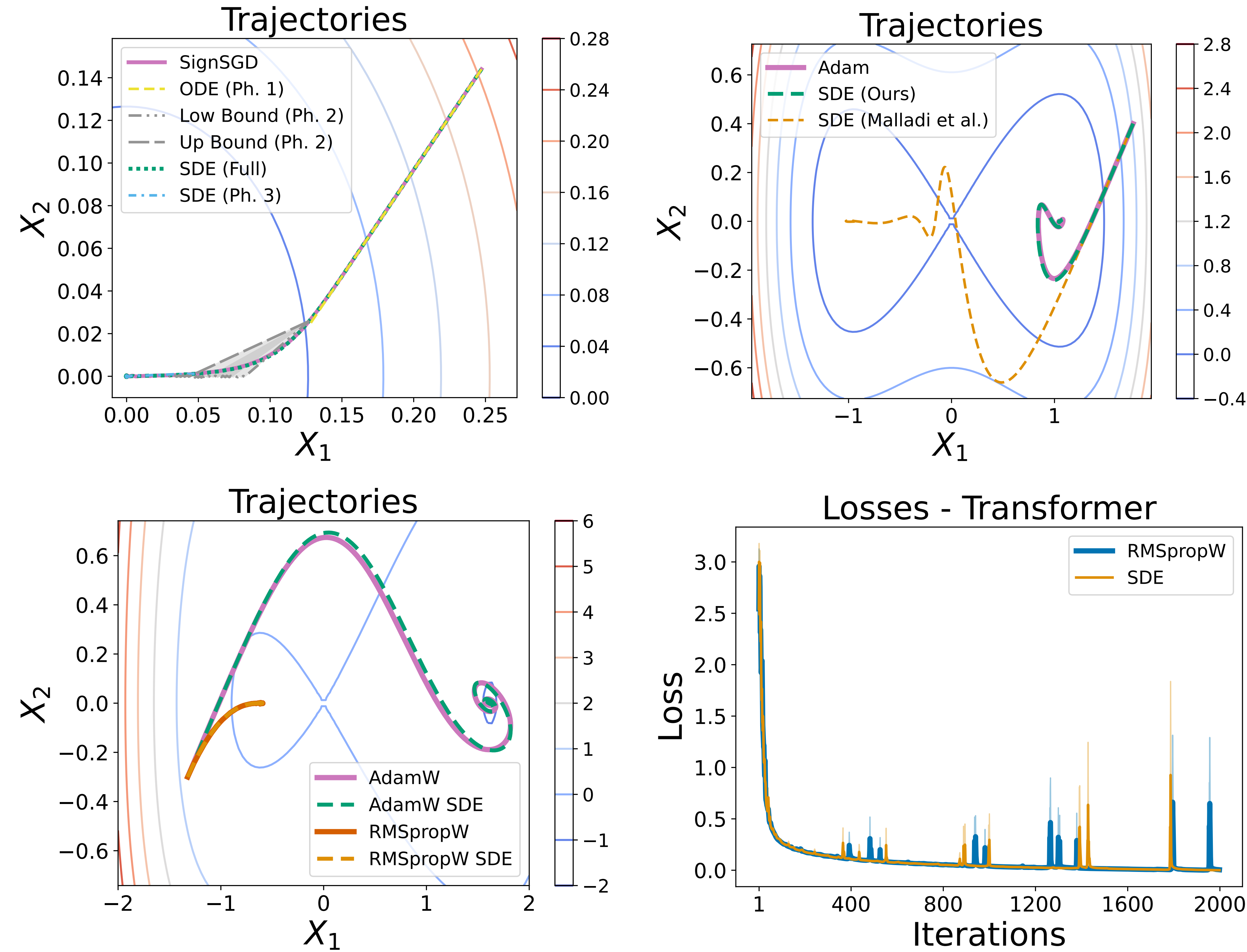




Visual Intuition - SDEs do Track the Optimizers



SDE Definitions

The (simplified) SDE of **SignSGD** is

$$dX_t = -\sqrt{\frac{2}{\pi}} \Sigma^{-\frac{1}{2}} \nabla f(X_t) dt + \sqrt{\eta} \sqrt{I_d - \frac{2}{\pi} \text{diag} \left(\Sigma^{-\frac{1}{2}} \nabla f(X_t) \right)^2} dW_t. \quad (1)$$

The SDE of **AdamW** for $\beta_i = 1 - \eta \rho_i$, $\iota_i(t) = 1 - e^{-\rho_i t}$, and $P_t = \text{diag} \sqrt{V_t} + \epsilon \sqrt{\iota_2(t)} I_d$ is

$$\begin{aligned} dX_t &= -\frac{\sqrt{\iota_2(t)}}{\iota_1(t)} P_t^{-1} (M_t + \eta \rho_1 (\nabla f(X_t) - M_t)) dt - \theta X_t dt, \\ dM_t &= \rho_1 (\nabla f(X_t) - M_t) dt + \sqrt{\eta} \rho_1 \sqrt{\Sigma(X_t)} dW_t, \\ dV_t &= \rho_2 ((\nabla f(X_t))^2 + \text{diag}(\Sigma(X_t)) - V_t) dt. \end{aligned} \quad (2)$$

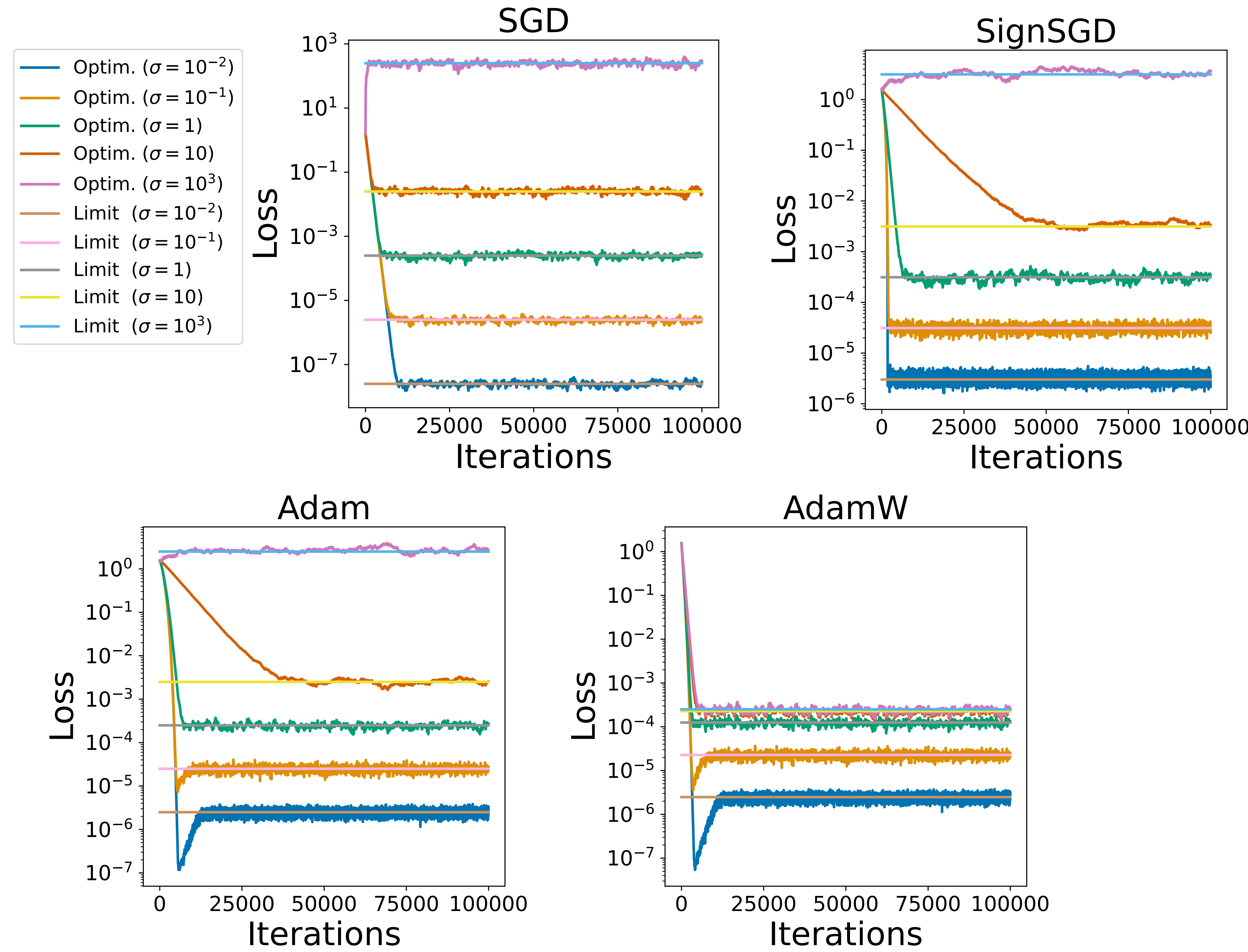
Problem of Interest

1. How do gradient noise and adaptivity interact?
2. What is the role of *decoupled* weight decay?
3. Are there any scaling laws involving weight decay?

Contributions

1. First SDE formulation for SignSGD and AdamW;
2. Adaptivity brings resilience to large noise;
3. **Decoupled** weight decay brings extreme resilience to it;
4. New scaling rules for hyperparameter tuning of AdamW.

Noise Resilience: Empirical Observation



Noise Resilience: Role of Adaptivity and Weight Decay

Assumptions: Let $f: \mathbb{R}^d \rightarrow \mathbb{R}$ be μ -strongly convex, L -smooth, $\Sigma(x) = \sigma^2 I_d$.

Theorem 1 (SGD).

$$\mathbb{E}[f(X_t) - f(X_*)] \leq (f(X_0) - f(X_*)) e^{-2\mu t} + (1 - e^{-2\mu t}) \frac{\eta L d}{4\mu} \times \frac{\sigma^2}{B}. \quad (3)$$

Theorem 2 (SignSGD - Simplified).

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

Theorem 3 (Adam on L^2 -Regularized Loss - Simplified).

$$\mathbb{E}[f(X_t) - f(X_*)] \stackrel{t \rightarrow \infty}{\leq} \frac{\eta L d}{4\mu} \times \frac{\sigma}{\sqrt{B} + \theta \frac{L+\mu}{2\mu L}}. \quad (5)$$

Theorem 4 (AdamW - Simplified).

$$\mathbb{E}[f(X_t) - f(X_*)] \stackrel{t \rightarrow \infty}{\leq} \frac{\eta L d}{4\mu} \times \frac{\sigma}{\sqrt{B} + \sigma \theta \frac{L+\mu}{2\mu L}} \stackrel{\forall \sigma}{<} \infty. \quad (6)$$

Scaling Laws: Batchsize Scaling

If we change the batchsize, how do we adapt the other hyperparameters?

This Paper:

1. $B \rightarrow \delta B$;
2. $\eta \rightarrow \sqrt{\delta} \eta$;
3. $\beta_i \rightarrow 1 - \sqrt{\delta}(1 - \beta_i)$;
4. $\theta \rightarrow \sqrt{\delta} \theta$.

OR

Malladi et al.:

1. $B \rightarrow \delta B$;
2. $\eta \rightarrow \sqrt{\delta} \eta$;
3. $\beta_i \rightarrow 1 - \delta(1 - \beta_i)$;
4. $\theta \rightarrow \sqrt{\delta} \theta$.

Validation on Pythia (160M & 10B Tokens)

