Symmetry Teleportation for Accelerated Optimization

Ildus Sadrtdinov
Alexander Shabalin

https://arxiv.org/pdf/2205.10637.pdf

Symmetry Teleportation

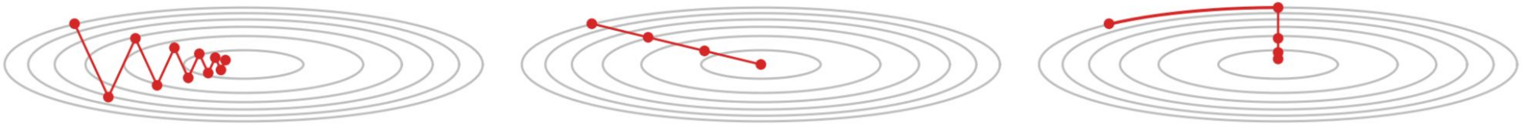


Figure 1: Left to right: gradient descent, second-order methods, proposed method.

Algorithm 1: Symmetry Teleportation

Input: Loss function $\mathcal{L}(\boldsymbol{w})$, learning rate η, number of epochs t_{max}, initialized parameters $\boldsymbol{w}_0$, symmetry group G, teleportation schedule K.

Output: $\boldsymbol{w}_{t_{max}}$.

for $t \leftarrow 0$ **to** $t_{max} - 1$ **do**
  if $t \in K$ **then**
    $g \leftarrow \operatorname{argmax}_{g \in G} \|(\nabla \mathcal{L})|_{g \cdot \boldsymbol{w}_t}\|^2$
    $\boldsymbol{w}_t \leftarrow g \cdot \boldsymbol{w}_t$
  end if
  $\boldsymbol{w}_{t+1} \leftarrow \boldsymbol{w}_t - \eta (\nabla \mathcal{L})|_{\boldsymbol{w}_t}$
end for
return $\boldsymbol{w}_{t_{max}}$

$$G = \left\{ g : \mathcal{L}(w) = \mathcal{L}(g \cdot w) \right\}$$

Example: Booth function

$$\mathcal{L}_b(x_1, x_2) = (x_1 + 2x_2 - 7)^2 + (2x_1 + x_2 - 5)^2 \qquad \mathcal{L}_b(u, v) = u^2 + v^2$$

Bijective change of variables

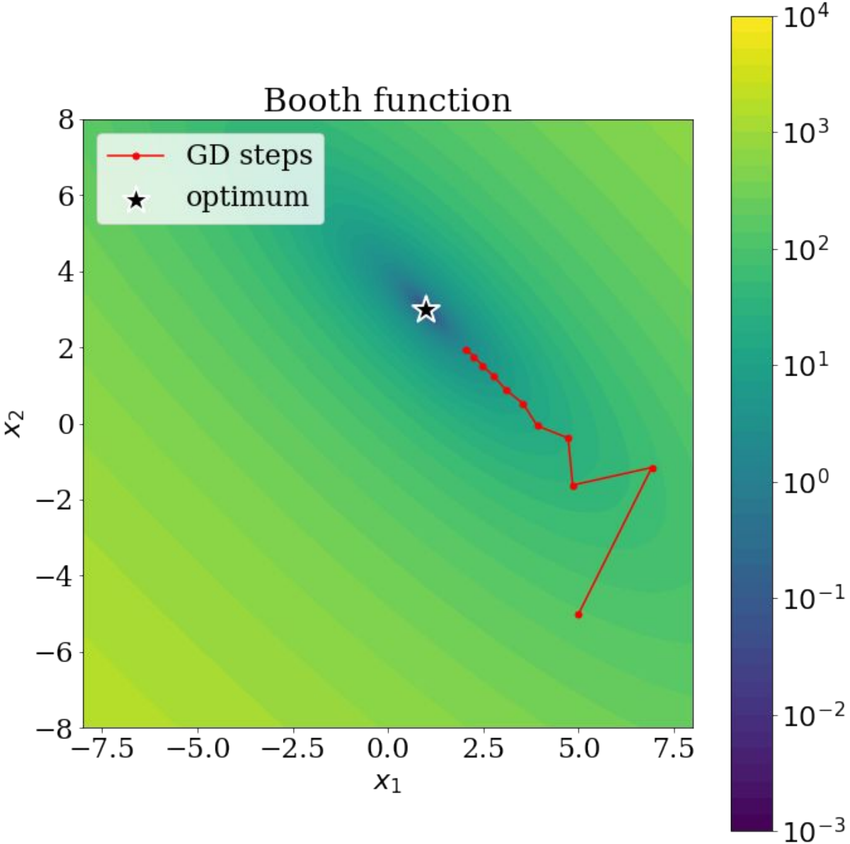
$$(u, v) = h(x_1, x_2) = (x_1 + 2x_2 - 7, 2x_1 + x_2 - 5)$$

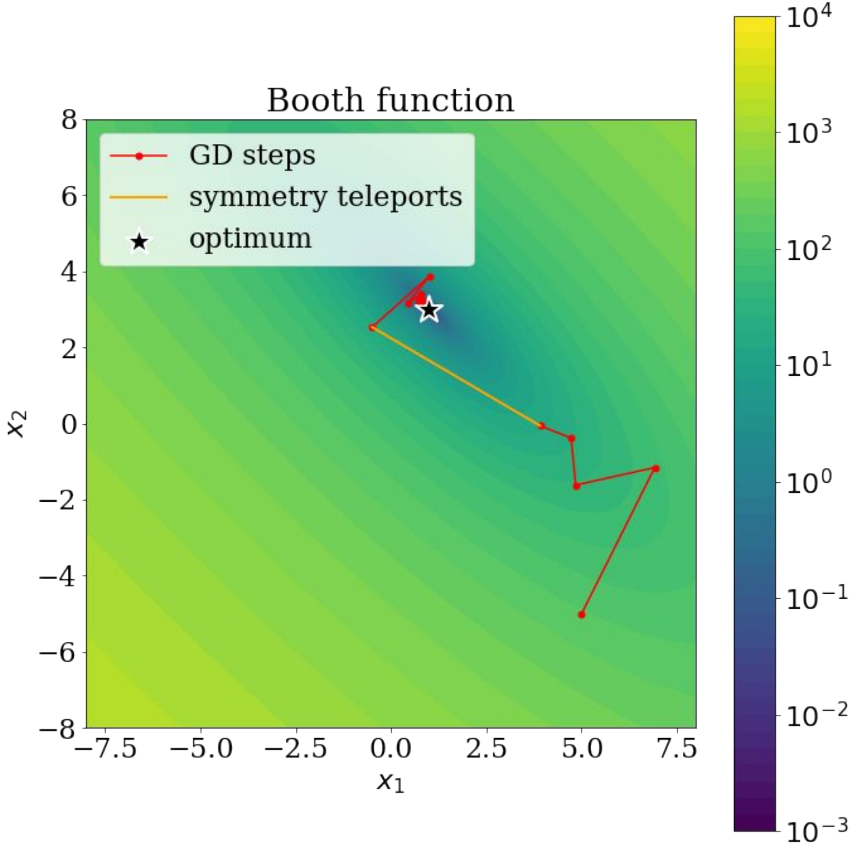
$$(x_1, x_2) = h^{-1}(u, v) = (-\frac{1}{3}u + \frac{2}{3}v + 1, \frac{2}{3}u - \frac{1}{3}v + 3)$$

$$g_\theta \cdot (x_1, x_2) = h^{-1}(R_\theta\, h(x_1, x_2)) \qquad \mathcal{L}_b(x_1, x_2) = \mathcal{L}_b(g_\theta \cdot (x_1, x_2))$$

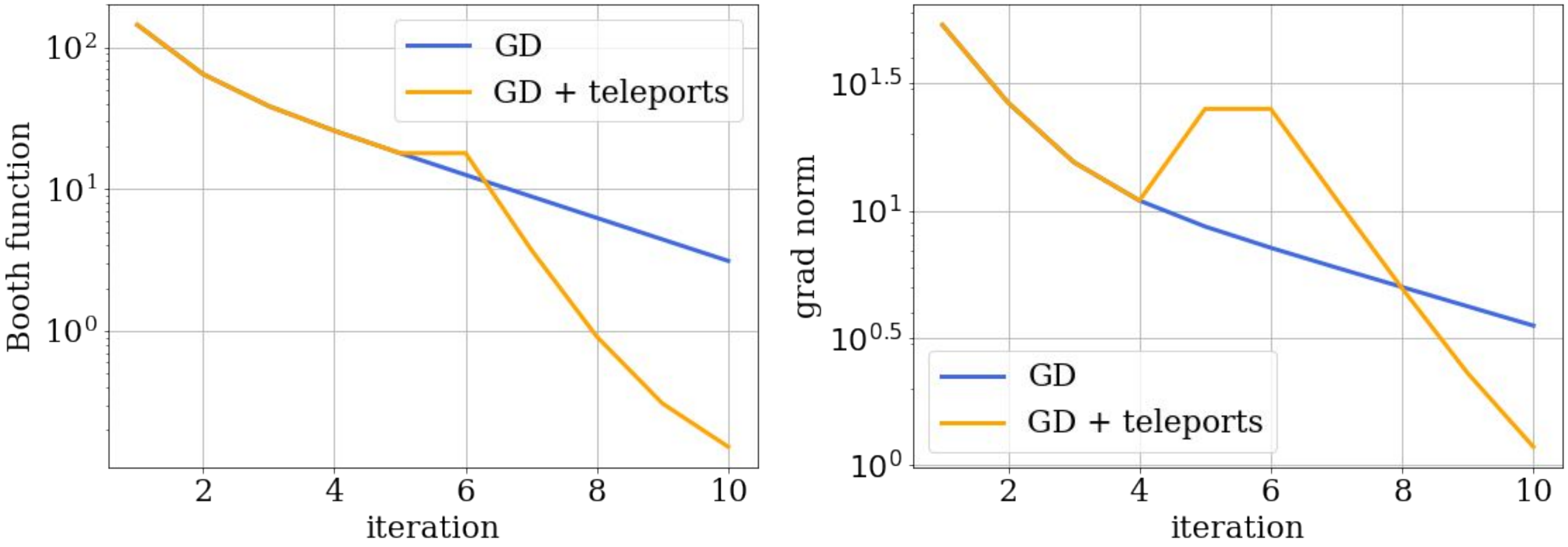
Arbitrary rotation matrix

Example: Booth function





Example: Booth function



Example: Rosenbrock function

$$\mathcal{L}_r(x_1, x_2) = 100(x_1^2 - x_2)^2 + (x_1 - 1)^2 \qquad \mathcal{L}_r(u, v) = u^2 + v^2$$

Bijective change of variables

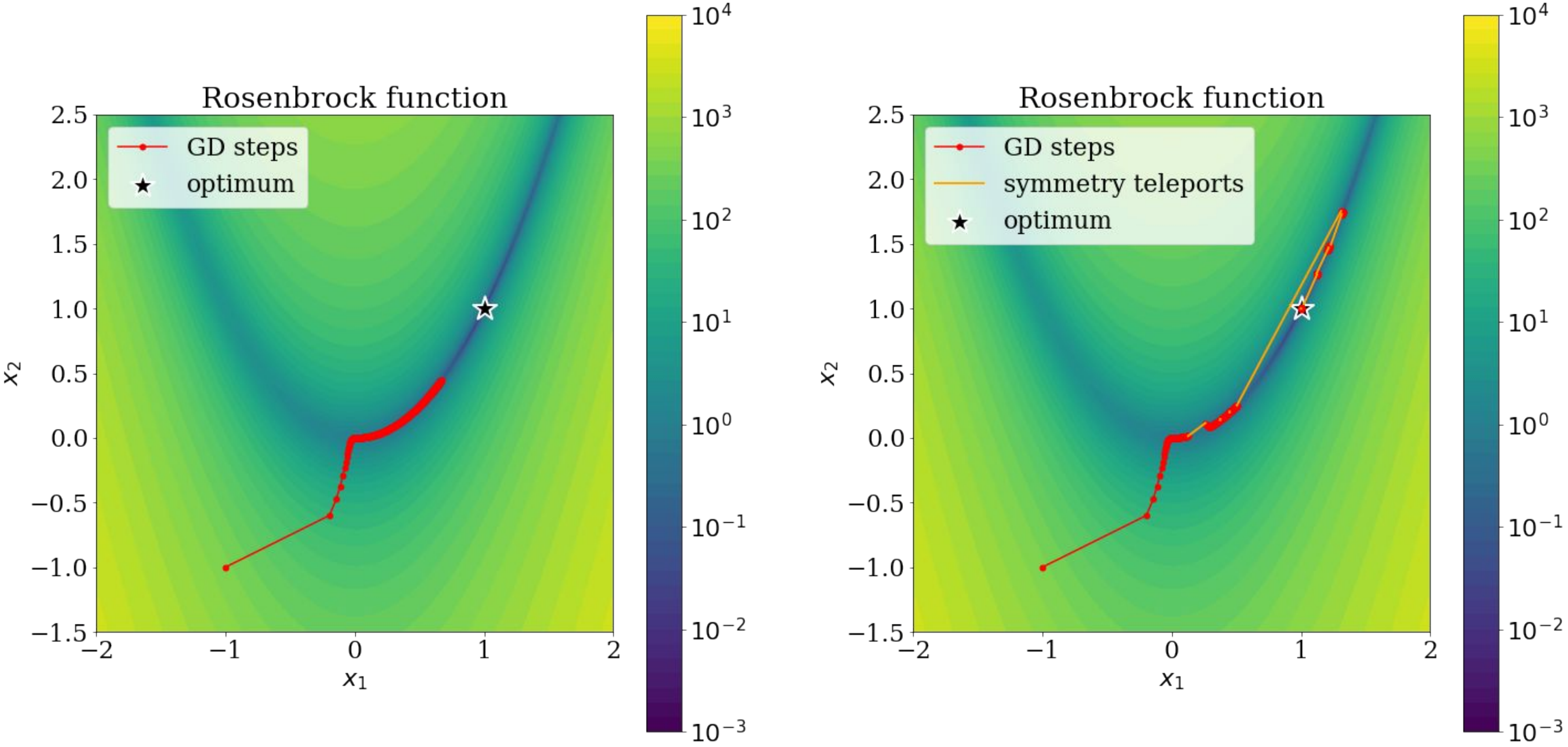
$$(u, v) = h(x_1, x_2) = (10(x_1^2 - x_2), x_1 - 1)$$

$$(x_1, x_2) = h^{-1}(u, v) = (v + 1, (v + 1)^2 - 0.1u)$$

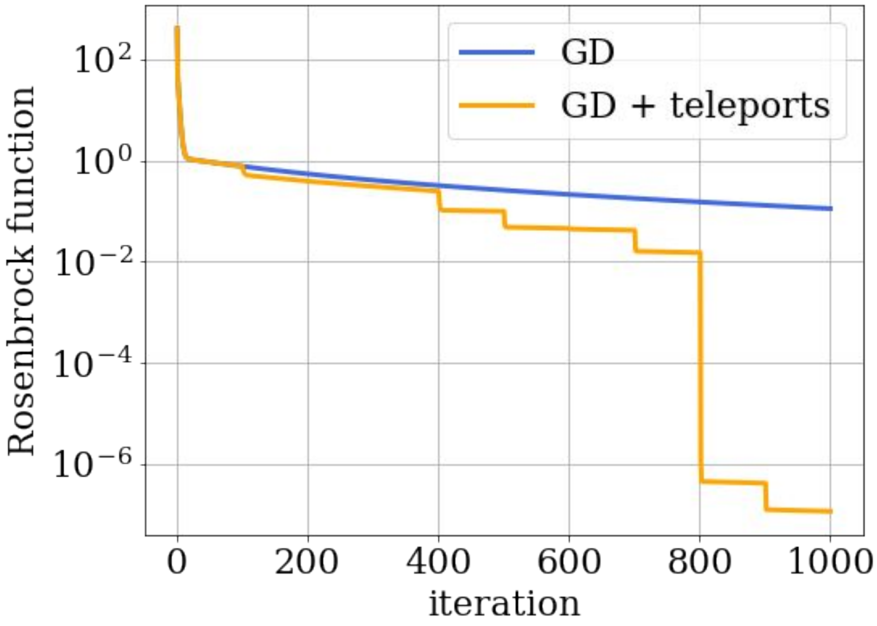
$$g_\theta \cdot (x_1, x_2) = h^{-1}(R_\theta\, h(x_1, x_2)) \qquad \mathcal{L}_r(x_1, x_2) = \mathcal{L}_r(g_\theta \cdot (x_1, x_2))$$

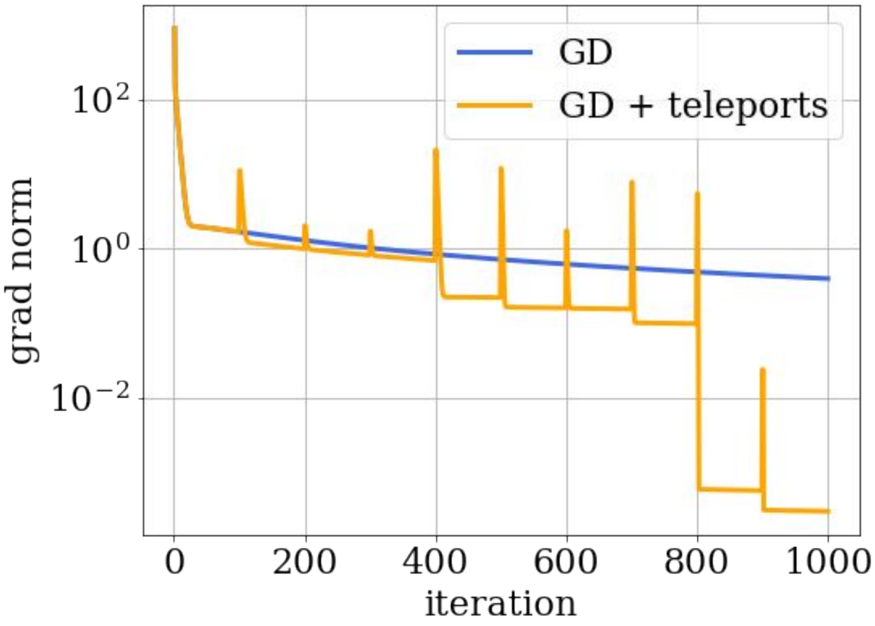
Arbitrary rotation matrix

Example: Rosenbrock function



Example: Rosenbrock function





Example: MNIST

$$MLP(x) = W_3\sigma(W_2\sigma(W_1x))$$

$$g_m \cdot W_k = \begin{cases} W_m g_m^{-1} & k = m \\ \sigma^{-1}\left(g_m \sigma\left(W_{m-1}h_{m-2}\right)\right) h_{m-2}^{-1} & k = m-1 \\ W_k & k \notin \{m, m-1\} \end{cases}$$

$$MLP(x) = \underbrace{W_3 g_3^{-1}}_{W_3'} \sigma\left(\underbrace{\sigma^{-1}(g_3\sigma(W_2h_1))h_1^{-1}}_{W_2'} \underbrace{\sigma\left(W_1x\right)}_{h_1} \right) =$$

Example: MNIST

$$MLP(x) = W_3\sigma(W_2\sigma(W_1x))$$

$$g_m \cdot W_k = \begin{cases} W_m g_m^{-1} & k = m \\ \sigma^{-1}\left(g_m\sigma\left(W_{m-1}h_{m-2}\right)\right)h_{m-2}^{-1} & k = m-1 \\ W_k & k \notin \{m, m-1\} \end{cases}$$

$$MLP(x) = \underbrace{W_3 g_3^{-1}}_{W_3'}\sigma\left(\underbrace{\sigma^{-1}(g_3\sigma(W_2h_1))\cancel{h_1^{-1}}}_{W_2'}\underbrace{\cancel{\sigma(W_1x)}}_{h_1}\right) = \underbrace{W_3 g_3^{-1}}_{W_3'}\sigma\left(\sigma^{-1}(g_3\sigma(W_2h_1))\right) =$$

Example: MNIST

$$MLP(x) = W_3\sigma(W_2\sigma(W_1 x))$$

$$g_m \cdot W_k = \begin{cases} W_m g_m^{-1} & k = m \\ \sigma^{-1}\left(g_m \sigma\left(W_{m-1} h_{m-2}\right)\right) h_{m-2}^{-1} & k = m-1 \\ W_k & k \notin \{m, m-1\} \end{cases}$$

$$MLP(x) = \underbrace{W_3 g_3^{-1}}_{W_3'} \sigma\left(\underbrace{\sigma^{-1}(g_3\sigma(W_2 h_1))\cancel{h_1^{-1}}}_{W_2'} \underbrace{\cancel{\sigma(W_1 x)}}_{h_1} \right) = \underbrace{W_3 g_3^{-1}}_{W_3'} \cancel{\sigma}\left(\cancel{\sigma^{-1}}(g_3\sigma(W_2 h_1)) \right) =$$

$$= W_3 g_3^{-1}(g_3\sigma(W_2 h_1)) = W_3\sigma(W_2 h_1)$$

Example: MNIST

$$MLP(x) = W_3\sigma(W_2\sigma(W_1x))$$

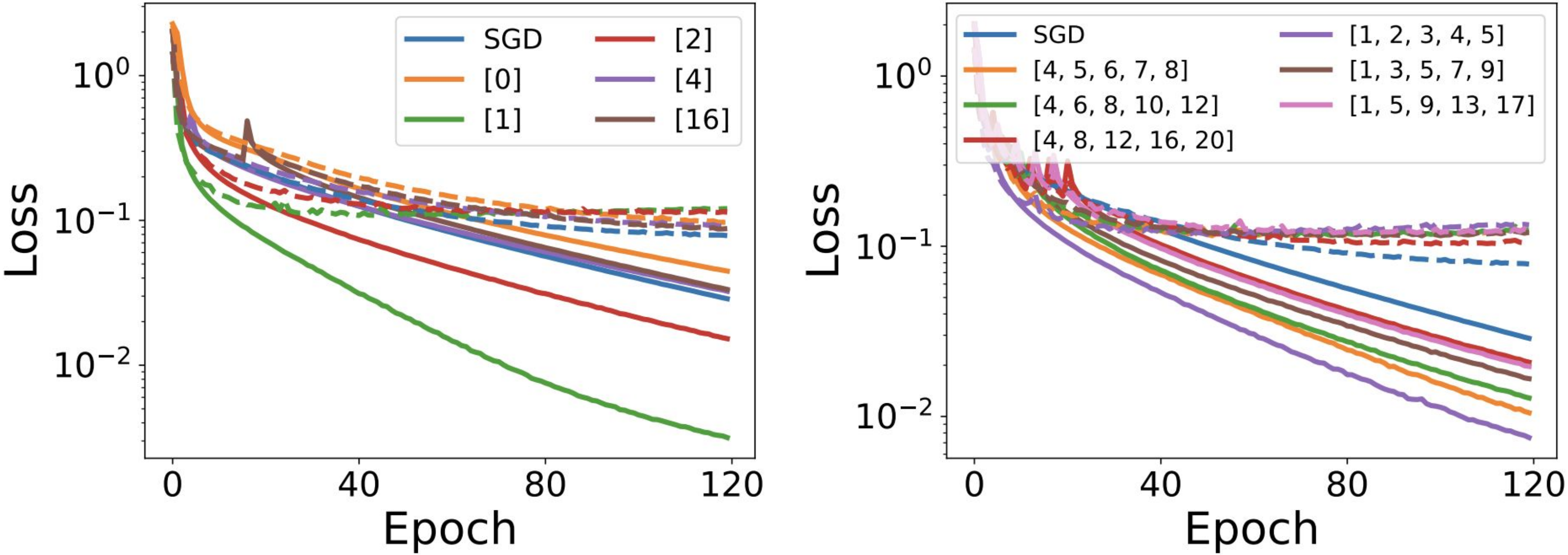
$$g_m \cdot W_k = \begin{cases} W_m g_m^{-1} & k = m \\ \sigma^{-1}\left(g_m\sigma\left(W_{m-1}h_{m-2}\right)\right)h_{m-2}^{-1} & k = m-1 \\ W_k & k \notin \{m, m-1\} \end{cases}$$

$$MLP(x) = \underbrace{W_3g_3^{-1}}_{W_3'}\sigma\left(\underbrace{\sigma^{-1}(g_3\sigma(W_2h_1))\cancel{h_1^{-1}}}_{W_2'}\underbrace{\cancel{\sigma(W_1x)}}_{h_1}\right) = \underbrace{W_3g_3^{-1}}_{W_3'}\cancel{\sigma}\left(\cancel{\sigma^{-1}}(g_3\sigma(W_2h_1))\right) =$$

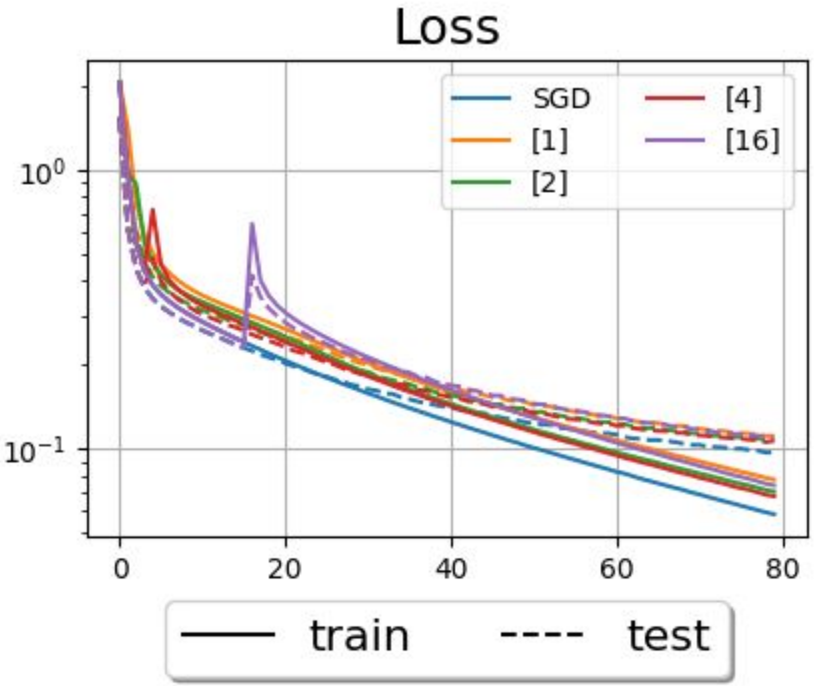
$$= W_3g_3^{-1}(g_3\sigma(W_2h_1)) = W_3\sigma(W_2h_1)$$

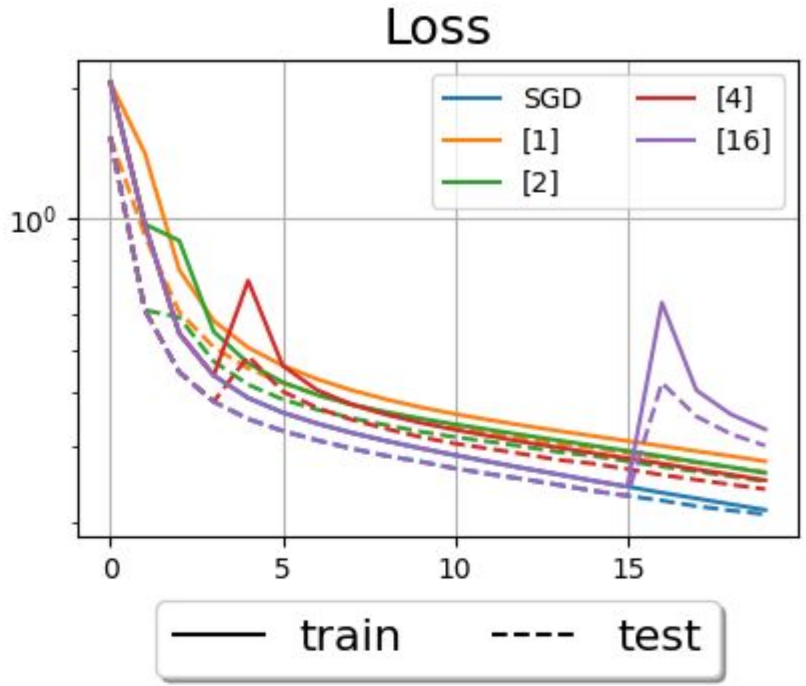
$$g \approx I + \epsilon T, \quad g^{-1} \approx I - \epsilon T, \quad gg^{-1} = I - \epsilon^2T^2 \underset{\epsilon\to 0}{\to} I$$

Example: MNIST (Paper results)

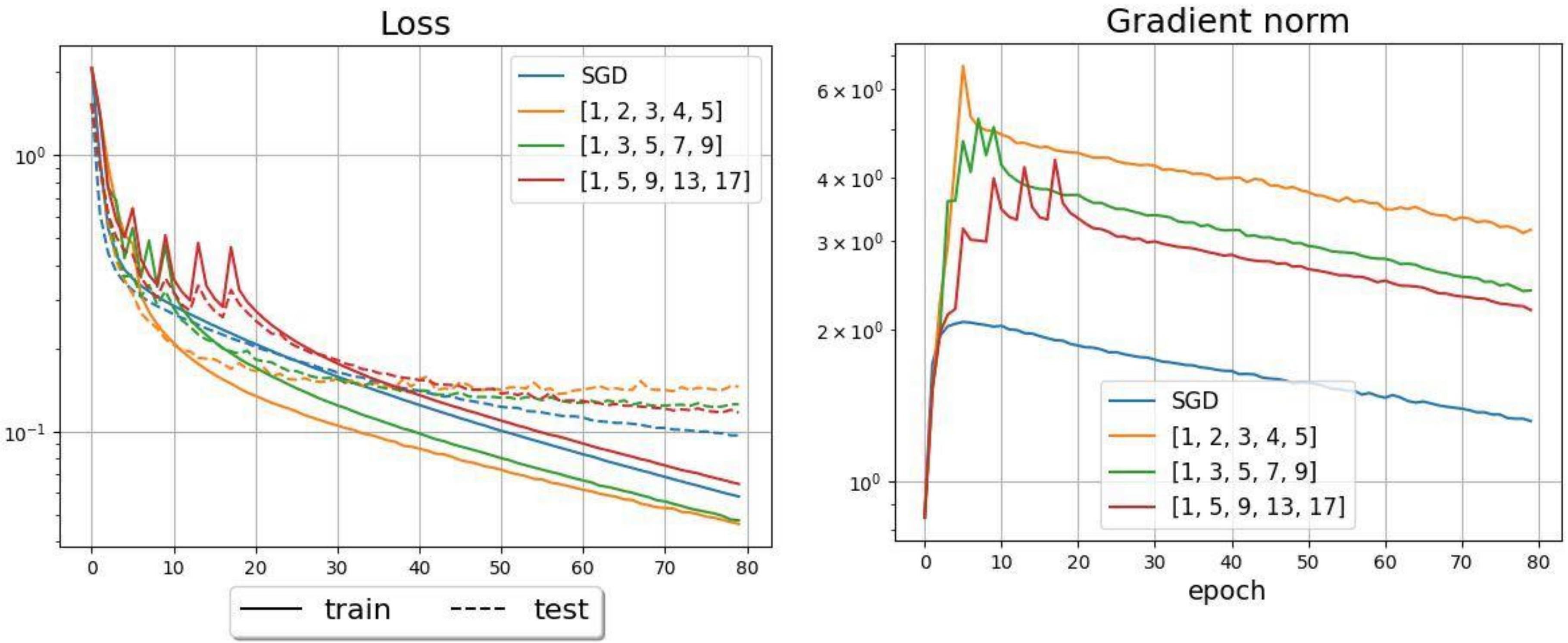


Example: MNIST (One Teleportation)

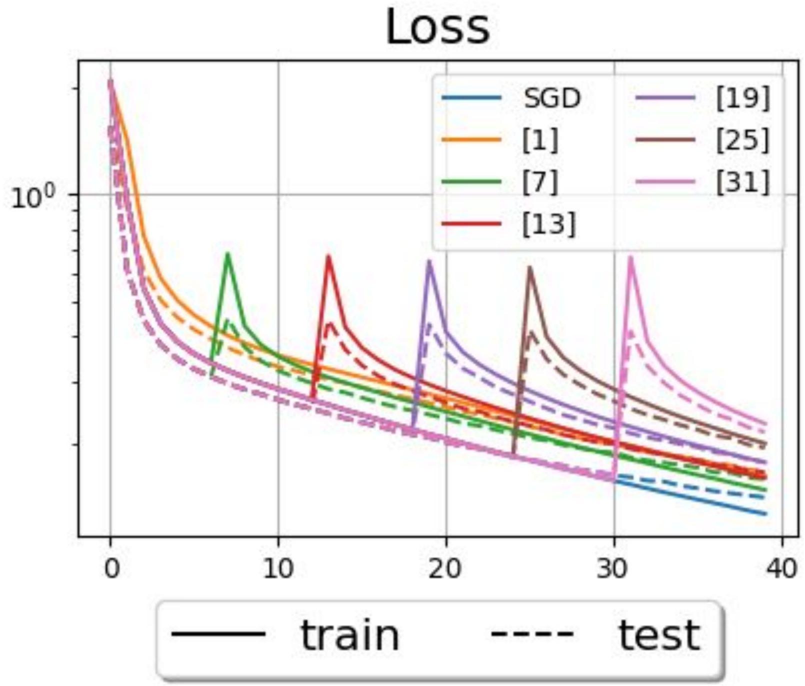


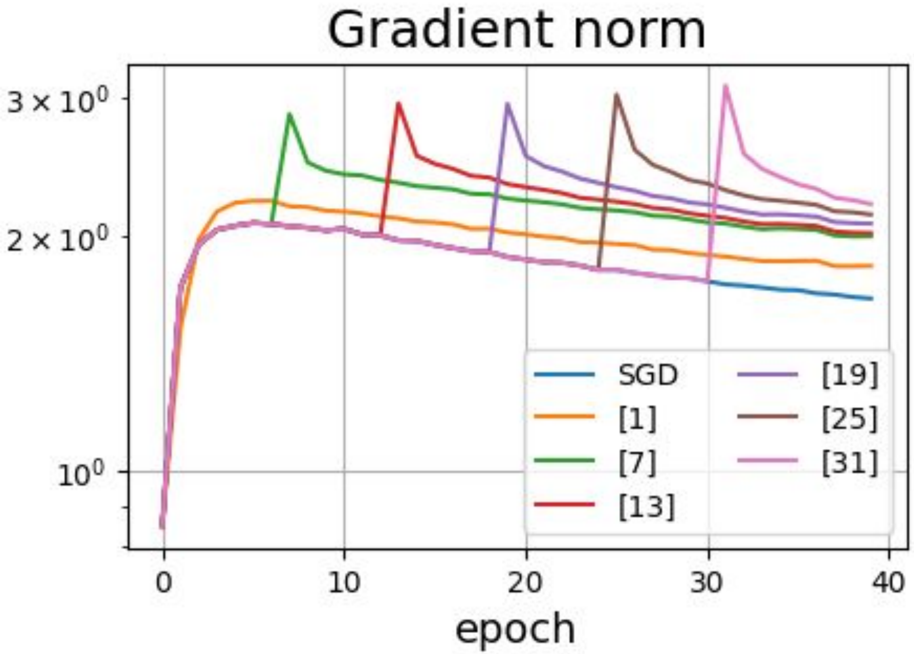


Example: MNIST (Several Teleportations)



Example: MNIST





Example: MNIST (Batch Size 200)