Algorithm 2 Local Training

Input: Encrypted global weight $[\![W^{(t)}]\!]$, local training data D_i $(i \in [1, n])$.

Output: Encrypted local gradients $[\![g_i^{(t)}]\!]$.

if $\mathcal{U}_i \notin \mathcal{U}^*$ **then**

  */*Benign training*/*

  $\mathcal{U}_i$ trains $W^{(t)}$ on local data, and obtains local gradients g_i;

  */*Gradient normalization*/*

  $\mathcal{U}_i$ normalizes individual gradients $g_i^{(t)}$ before sending them to $\mathcal{S}_1$;

else

  */*Model poisoning*/*

  $\mathcal{U}_i$ launches model poisoning, and yields poisonous gradients $g_i^{*(t)}$;

return Encrypted local gradients $[\![g_i{}^{(t)}]\!]$ or $[\![g_i^{*(t)}]\!]$.