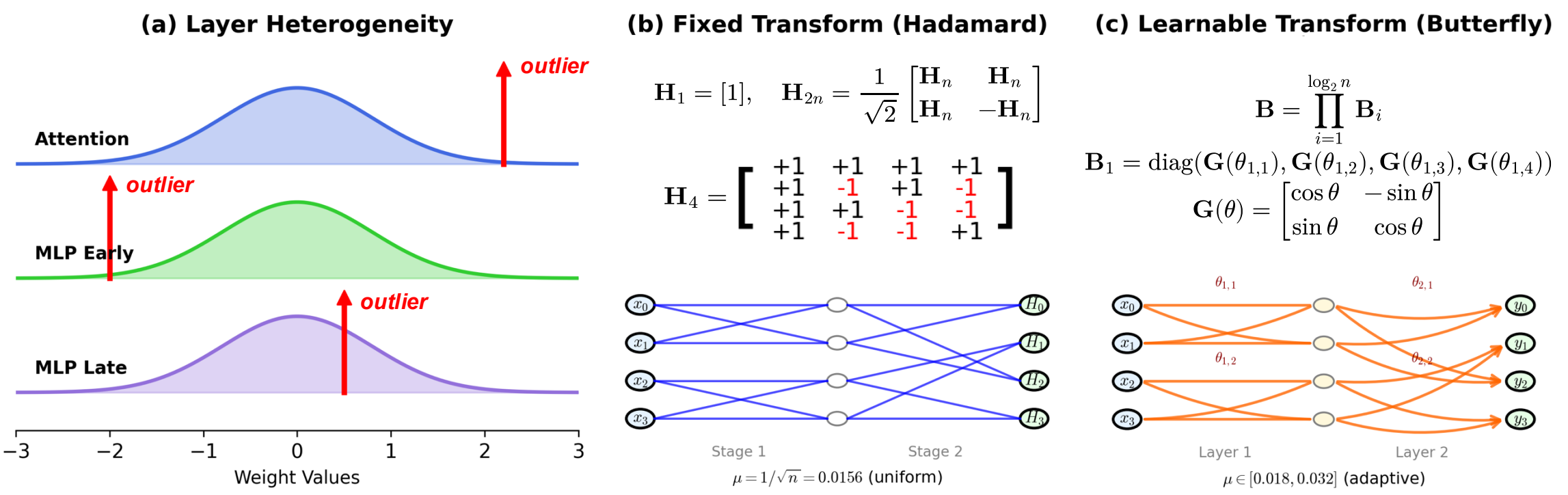
(a) Layer Heterogeneity
outlier
Attention
outlier
MLP Early
outlier
MLP Late
−3
−2
−1
0
1
2
3
Weight Values
(b) Fixed Transform (Hadamard)
$\mathbf{H}_1 = [1], \quad \mathbf{H}_{2n} = \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{H}_n & \mathbf{H}_n \\ \mathbf{H}_n & -\mathbf{H}_n \end{bmatrix}$
$\mathbf{H}_4 = \begin{bmatrix} +1 & +1 & +1 & +1 \\ +1 & -1 & +1 & -1 \\ +1 & +1 & -1 & -1 \\ +1 & -1 & -1 & +1 \end{bmatrix}$
x_0
x_1
x_2
x_3
H_0
H_1
H_2
H_3
Stage 1
Stage 2
$\mu = 1/\sqrt{n} = 0.0156$ (uniform)
(c) Learnable Transform (Butterfly)
$\mathbf{B} = \prod_{i=1}^{\log_2 n} \mathbf{B}_i$
$\mathbf{B}_1 = \mathrm{diag}(\mathbf{G}(\theta_{1,1}), \mathbf{G}(\theta_{1,2}), \mathbf{G}(\theta_{1,3}), \mathbf{G}(\theta_{1,4}))$
$\mathbf{G}(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$
$\theta_{1,1}$
$\theta_{2,1}$
$\theta_{1,2}$
$\theta_{2,2}$
x_0
x_1
x_2
x_3
y_0
y_1
y_2
y_3
Layer 1
Layer 2
$\mu \in [0.018, 0.032]$ (adaptive)