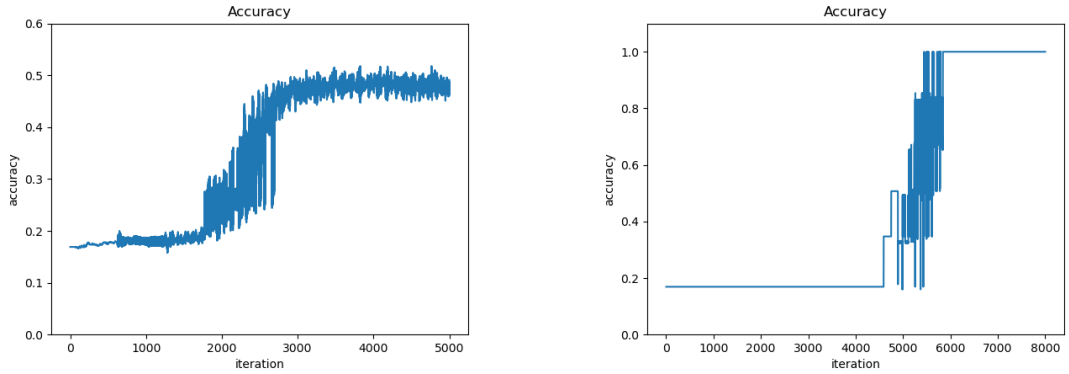
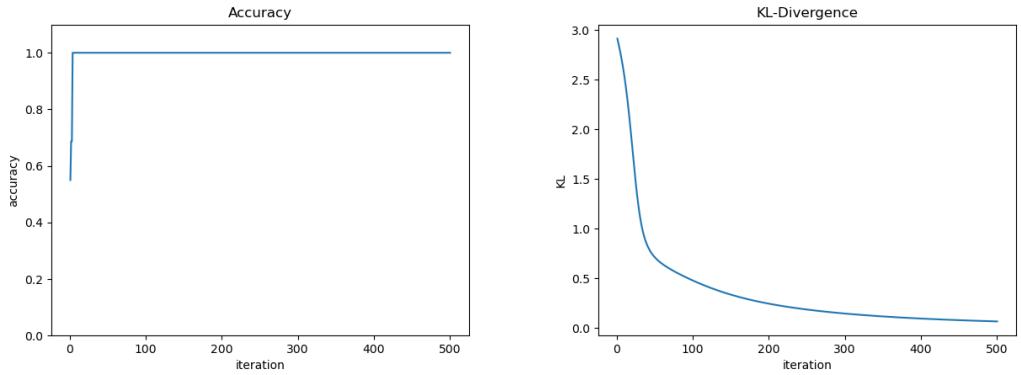
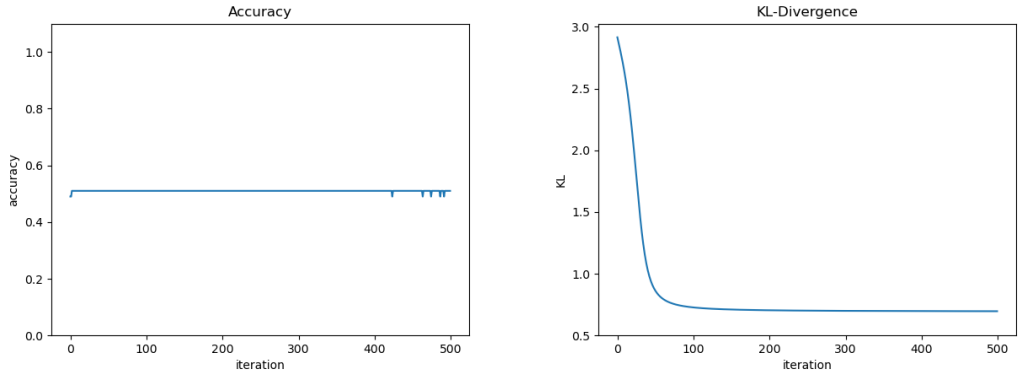


(a) Random walk (b) Deterministic walk

Figure 1: The test accuracy of the experiments conducted using a two-layer transformer for Task 1 and Task 2. The two-layer transformer model stacks two self-attention layers, parameterized by matrices $\boldsymbol{V}_1$, $\boldsymbol{W}_1$ and $\boldsymbol{V}_2$, $\boldsymbol{W}_2$ respectively. All parameters are initialized as independent Gaussian random variables from $N(0, \sigma^2)$ with $\sigma = 0.01$. The learning tasks are the same as experiments in Figure 5 in the paper. The learning rate is set as $\eta = 0.1$. (a) gives the result of learning random walks, and (b) shows the result of learning deterministic walks.

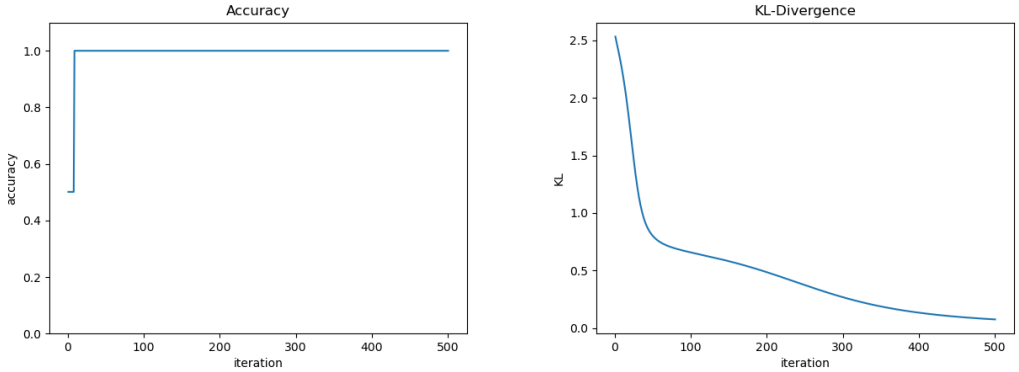


(a) Accuracy (Task 3) (b) KL divergence (Task 3)



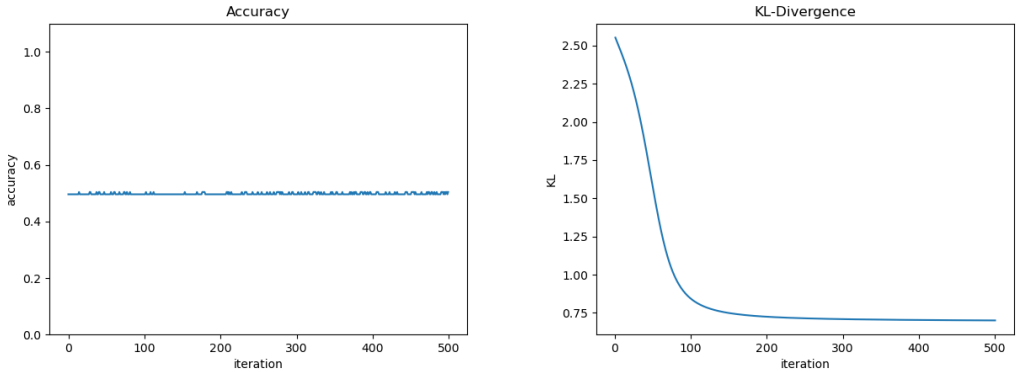
(c) Accuracy (Task 4) (d) KL divergence (Task 4)

Figure 2: The results of experiments conducted using a two-layer transformer for Task 3 and Task 4: (a) and (b) correspond to Task 3; (c) and (d) correspond to Task 4. The two-layer transformer model stacks two self-attention layers, parameterized by matrices $\boldsymbol{V}_1$, $\boldsymbol{W}_1$ and $\boldsymbol{V}_2$, $\boldsymbol{W}_2$ respectively. All parameters are initialized as independent Gaussian random variables from $N(0, \sigma^2)$ with $\sigma = 0.01$. The learning tasks are the same as experiments in Figure 7 in the paper. The learning rate is set as $\eta = 0.1$.



(a) Accuracy (Task 5)

(b) KL divergence (Task 5)



(c) Accuracy (Task 6)

(d) KL divergence (Task 6)

Figure 3: The results of experiments conducted using a two-layer transformer for Task 5 and Task 6: (a) and (b) correspond to Task 5; (c) and (d) correspond to Task 6. The two-layer transformer model stacks two self-attention layers, parameterized by matrices $\boldsymbol{V}_1$, $\boldsymbol{W}_1$ and $\boldsymbol{V}_2$, $\boldsymbol{W}_2$ respectively. All parameters are initialized as independent Gaussian random variables from $N(0, \sigma^2)$ with $\sigma = 0.01$. The learning tasks are described in detail in the response to reviewers. The learning rate is set as $\eta = 0.1$.