

# 1 Parameter Settings

Table 1 lists the hyper-parameters used in all models. For AIM, we use the same hyper-parameters as the baseline models (i.e., batch size, learning rate) except for extra ones.

Table 1: Parameter Settings

Model	Avazu	Criteo
General	bs=2000 lr=1e-3 opt=Adam	bs=2000 lr=1e-3 opt=Adam
FM, AFM	d=40 t=1; h=256 l2_a=0	d=20 t=0.01; h=32 l2_a=0.1
FwFM	d=40; w_init=0.7 w_l1=1e-8; w_l2=1e-7	d=20; w_init=0.7 w_l1=0; w_l2=1e-7
FFM	d=4	d=4
DeepFM, IPNN	d=40; LN=True net=[700× 5, 1]	d=20; LN=True net=[700× 5, 1]
xDeepFM	d=40 net=[700× 5, 1] CIN=[100× 2]	d=20 net=[400× 3, 1] CIN=[100× 4]
AutoFeature	d=40 LN=True net=[500× 5, 1]	d=20 LN=True net=[2048, 1024, 512, 256, 1]
AutoGroup	d=40; lr_h=1e4 $\tau$ =0.01 net=[1024, 512, 256, 1] n_p=[15, 130, 170, 210] high=[100, 80]	d=20; lr_h=10 $\tau$ =0.1 net=[1024, 512, 256, 1] n_p=[35, 390, 300, 500] high=[20, 100]
AutoFIS	d=40; LN=True net=[700× 5, 1] c=0.0005; mu=0.8	d=20; LN=True net=[700× 5, 1] c=0.0005; mu=0.8
AIM	d=40; LN=True net=[700× 5, 1] c_a=0.05; mu_a=0.6 lr2_a=1e-1; c_b=0.005 mu_b=0.55; lr2_b=1e0	d=20; LN=True net=[700× 5, 1] c_a=0.005; mu_a=0.9 lr2_a=1e-1; c_b=0.005 mu_b=0.51; lr2_b=1e-1

\* Note: bs=batch size, opt=optimizer, lr=learning rate, d=embedding size, w\_init = initial value for  $\alpha$ , w\_l1 and w\_l2 are  $l_1$  and  $l_2$  regularization on  $\alpha$ , t=Softmax temperature, l2\_a=  $l_2$  regularization on Attention network, net=MLP structure, LN=layer normalization, BN=batch normalization, lr\_h = learning rate for structural parameters,  $\tau$ = Gumbel-softmax temperature, n\_p= number of feature sets in each order, high = vector size for high order's embedding, c and mu are parameters in GRDA Optimizer, c\_a, mu\_a, lr\_a are GRDA parameters in search interaction-IF stage, c\_b, mu\_b, lr\_b are GRDA parameters in search embed stage.