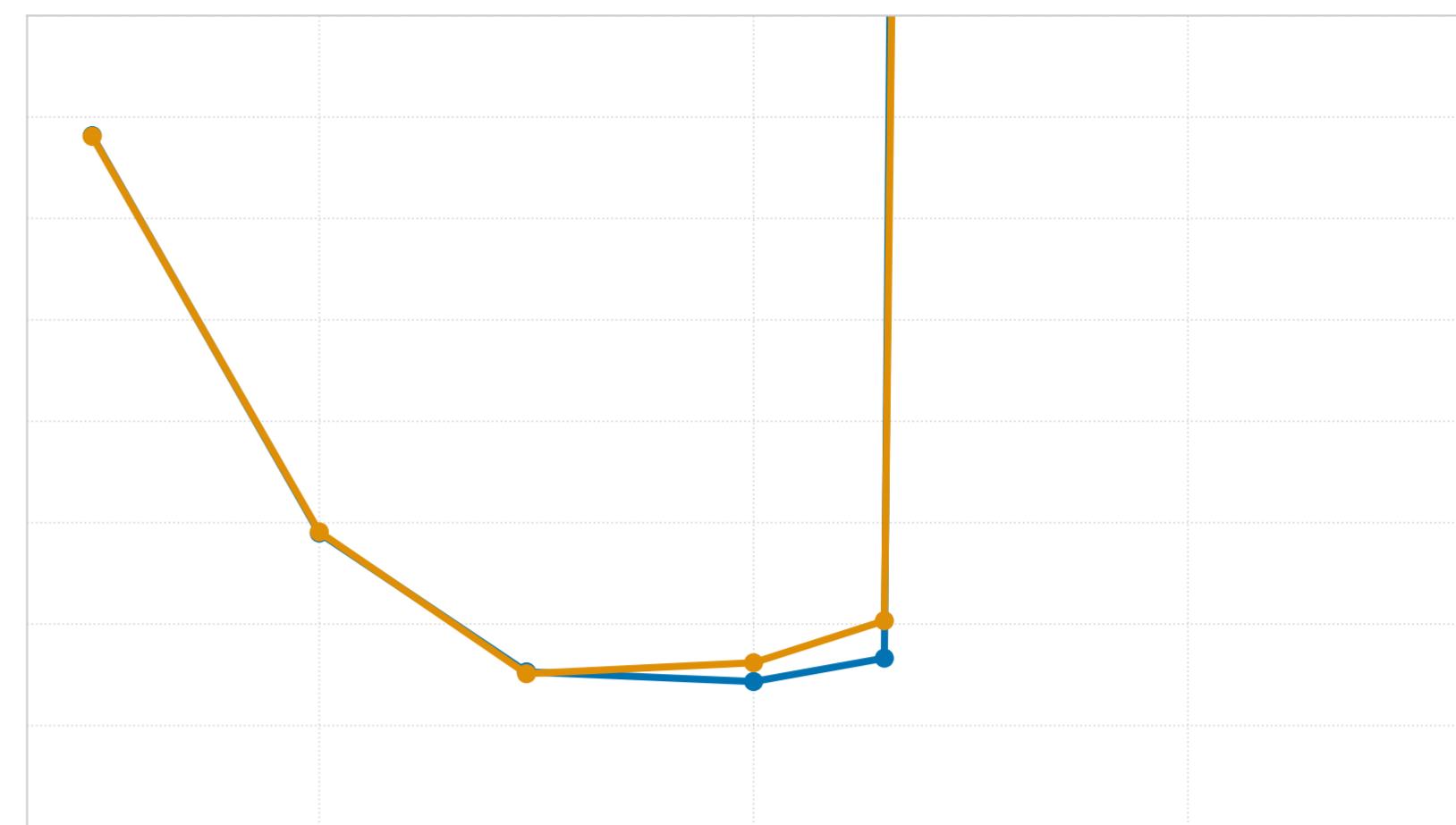
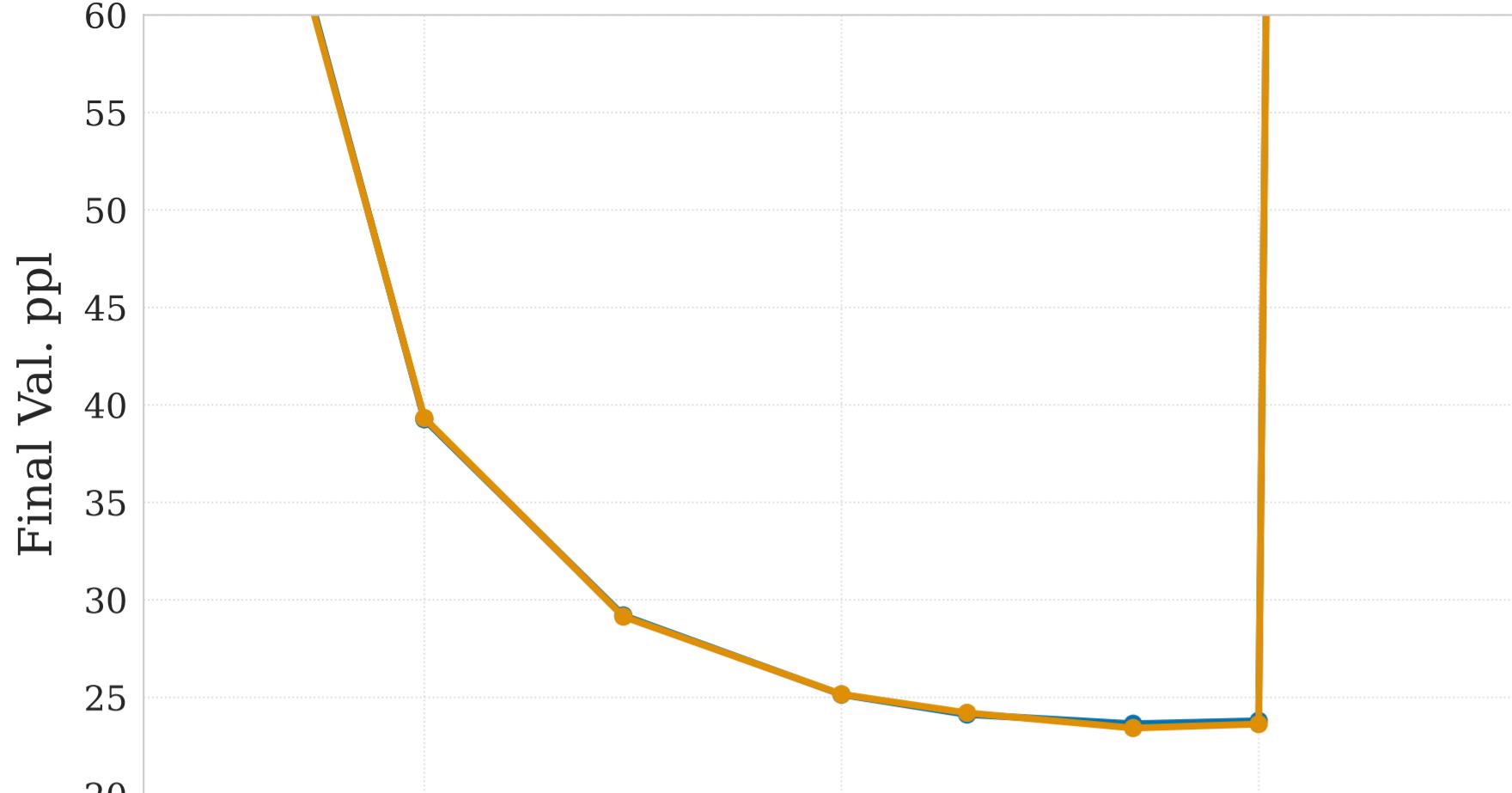


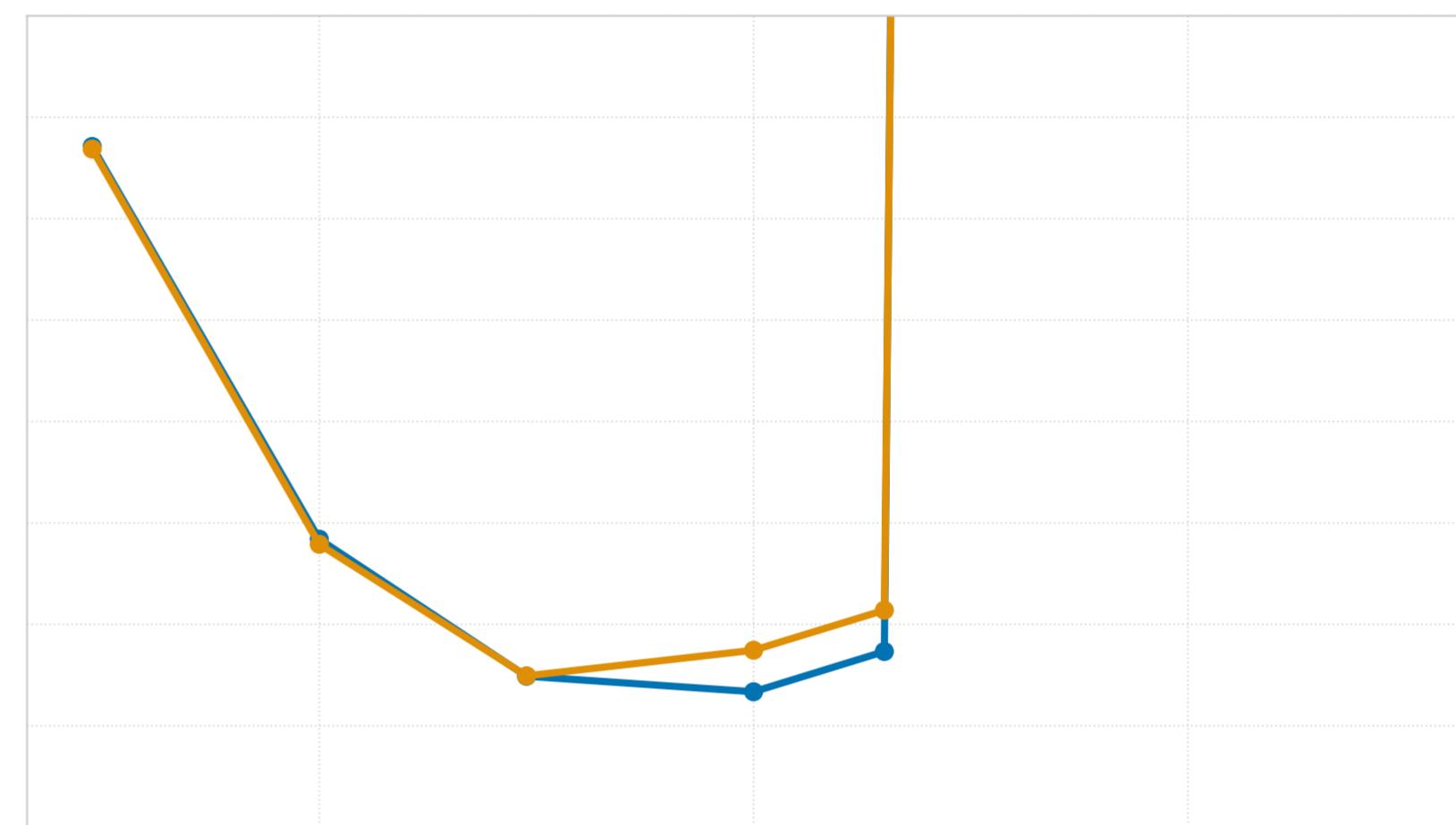
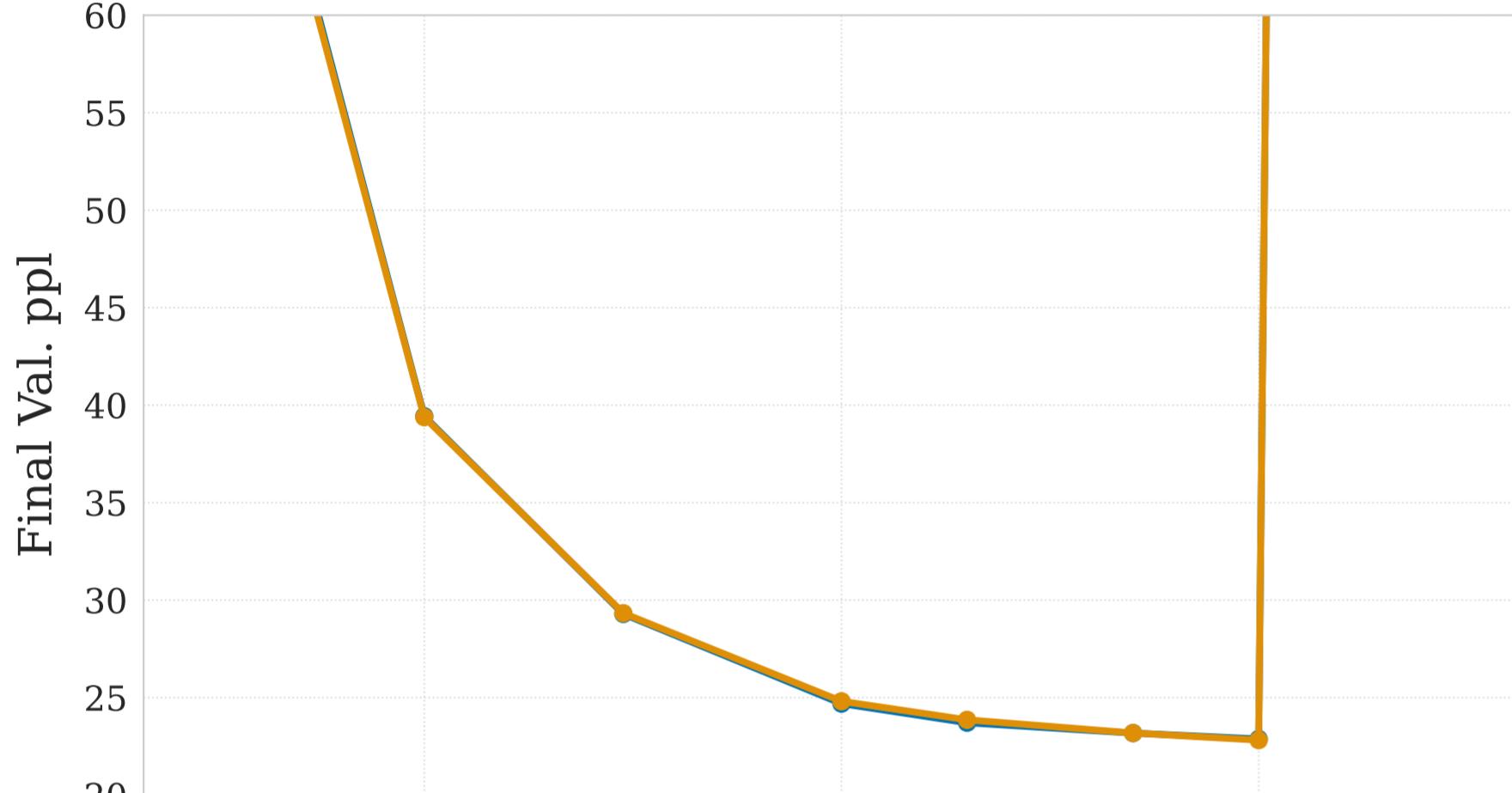
Warmup Cosine
 $\beta_1 = \beta_2 = 0.9$

Constant Learning Rate
 $\beta_1 = \beta_2 = 0.9$



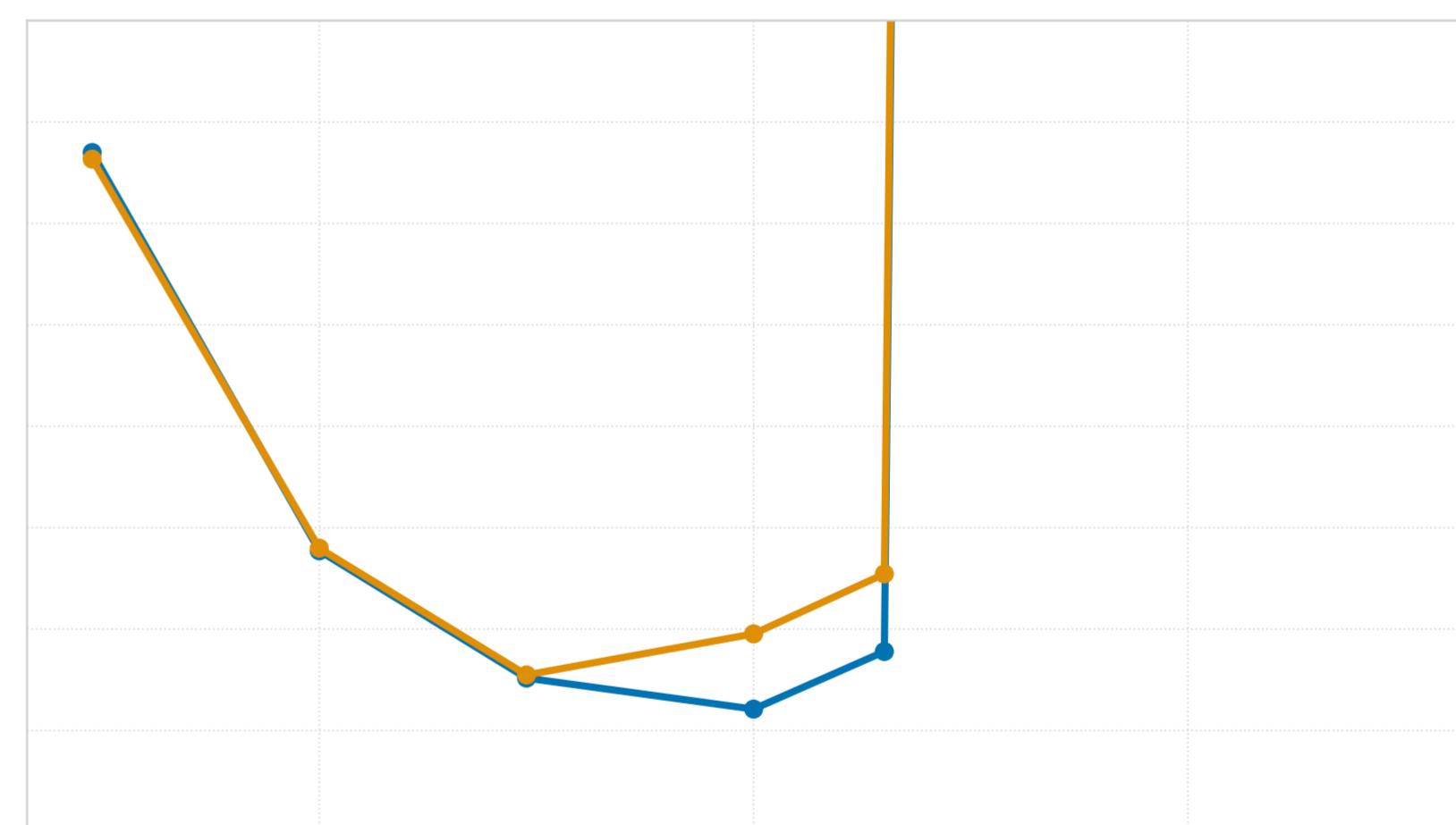
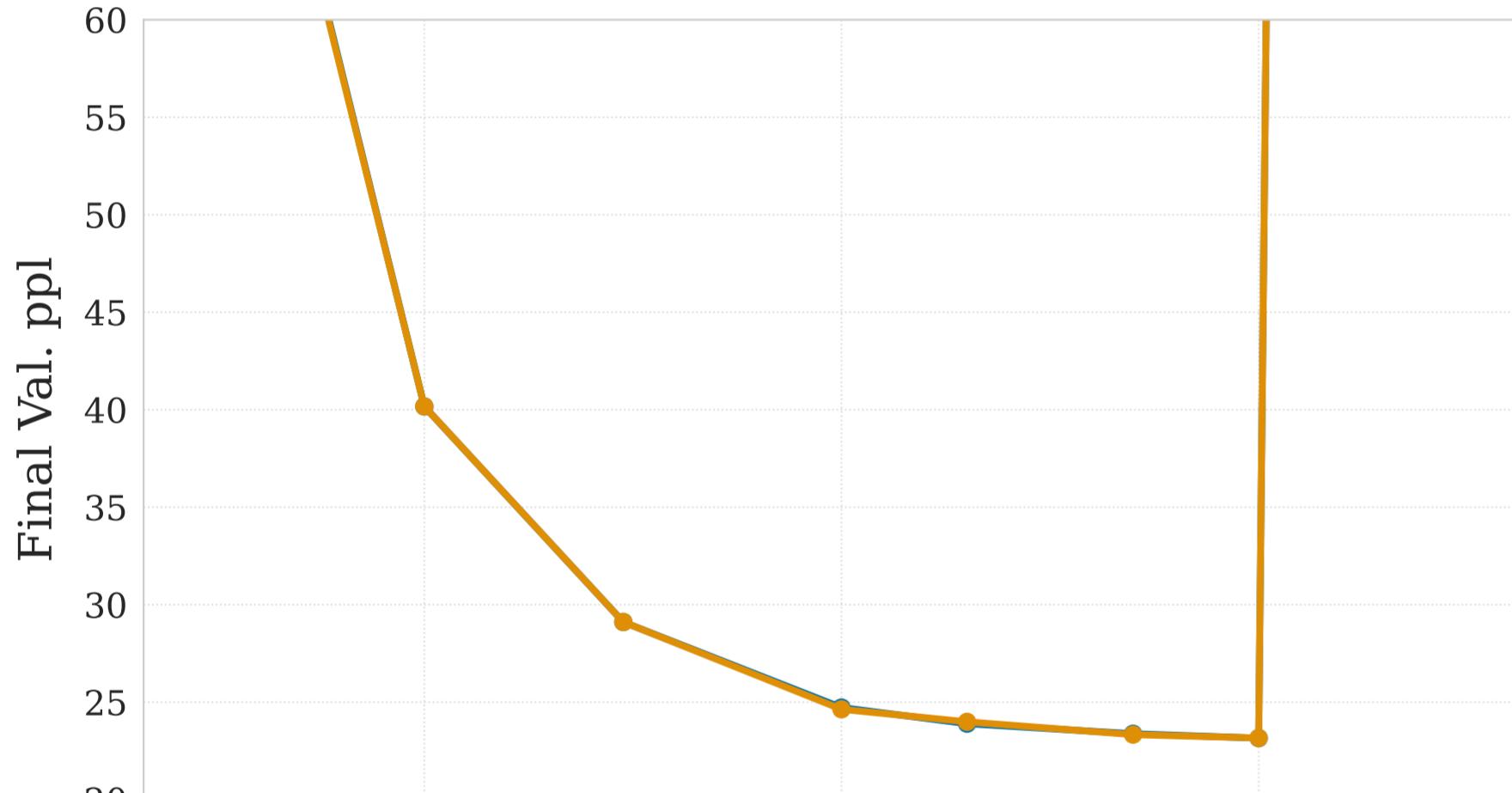
$\beta_1 = \beta_2 = 0.95$

$\beta_1 = \beta_2 = 0.95$



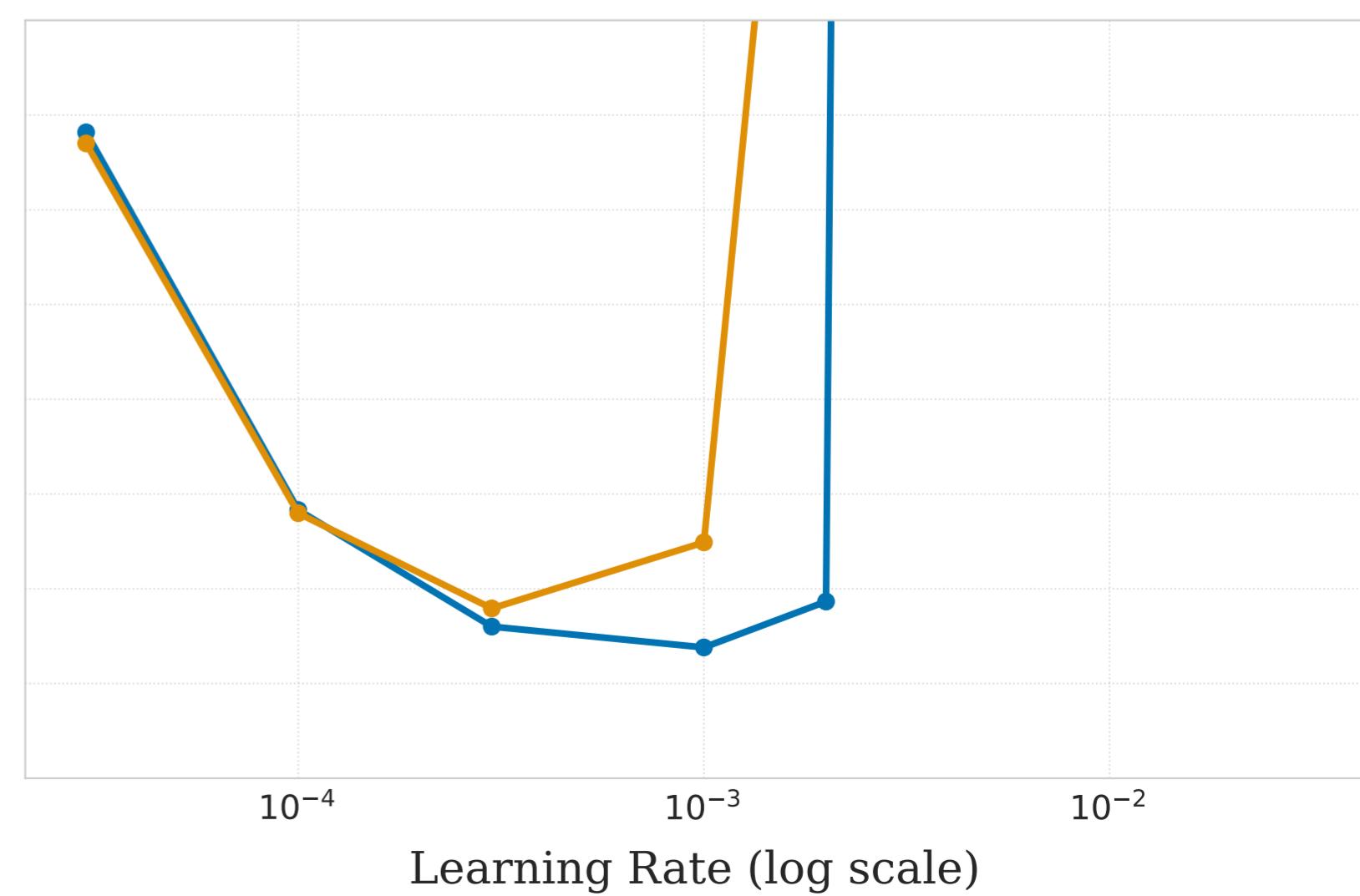
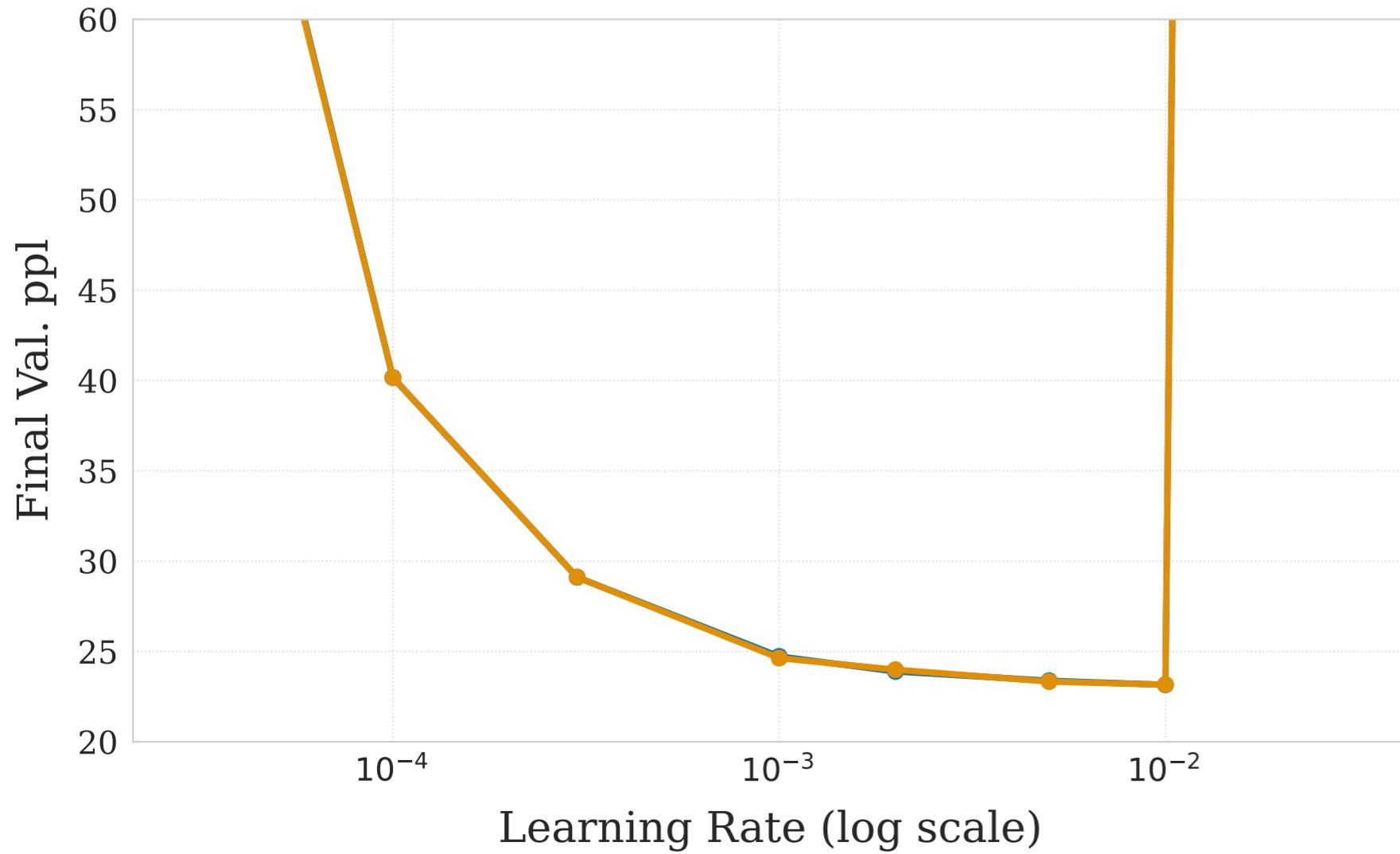
$\beta_1 = \beta_2 = 0.975$

$\beta_1 = \beta_2 = 0.975$



$\beta_1 = \beta_2 = 0.9875$

$\beta_1 = \beta_2 = 0.9875$



Learning Rate (log scale)

— Without Bias Correction

— With Bias Correction