Optimizers Demo

SGD, Momentum, Adam, AdamW

From-scratch NumPy implementations

This report demonstrates optimizer behavior through 7 examples:

1. Optimizer trajectories on Beale function (2D contour)
2. Adam convergence on a quadratic (loss curve + trajectory)
3. Adam vs AdamW vs Adam+L2 weight distributions
4. Learning rate schedules (warmup+cosine, cosine, step decay)
5. Momentum effect in a ravine (SGD vs Momentum vs Nesterov)
6. Ill-conditioned quadratic (condition number = 100)
7. Bias correction effect in early Adam steps

Key Findings

1. SGD struggles on non-convex and ill-conditioned problems due to oscillation.

2. Momentum dampens oscillations by accumulating velocity in consistent gradient directions.

3. Nesterov momentum provides slightly faster convergence via lookahead gradient evaluation.

4. Adam adapts per-parameter learning rates, handling diverse gradient scales automatically.

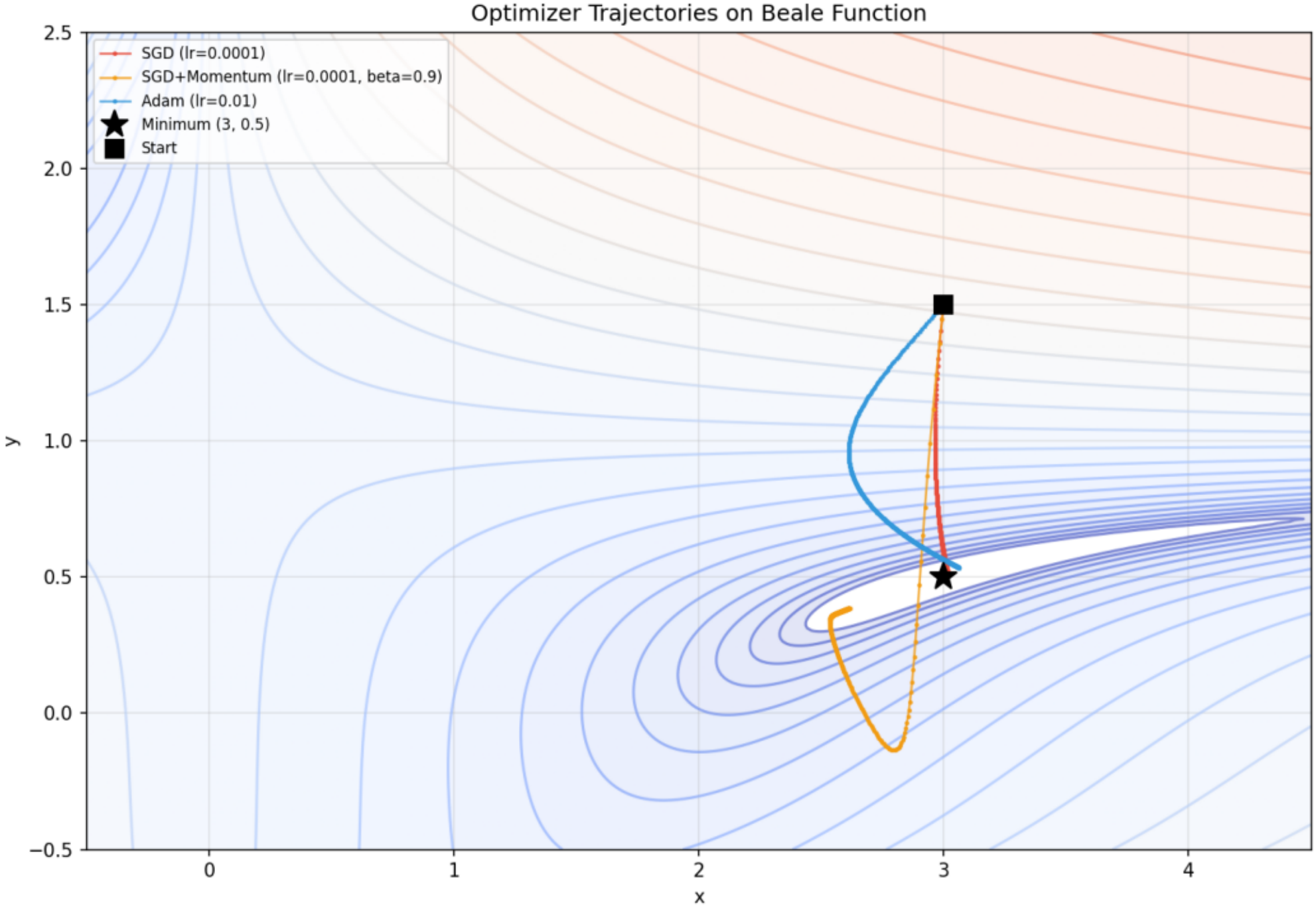
5. AdamW applies weight decay directly to parameters (decoupled), producing tighter weight distributions.

6. Adam+L2 folds weight decay into gradients (coupled), resulting in non-uniform effective regularization.

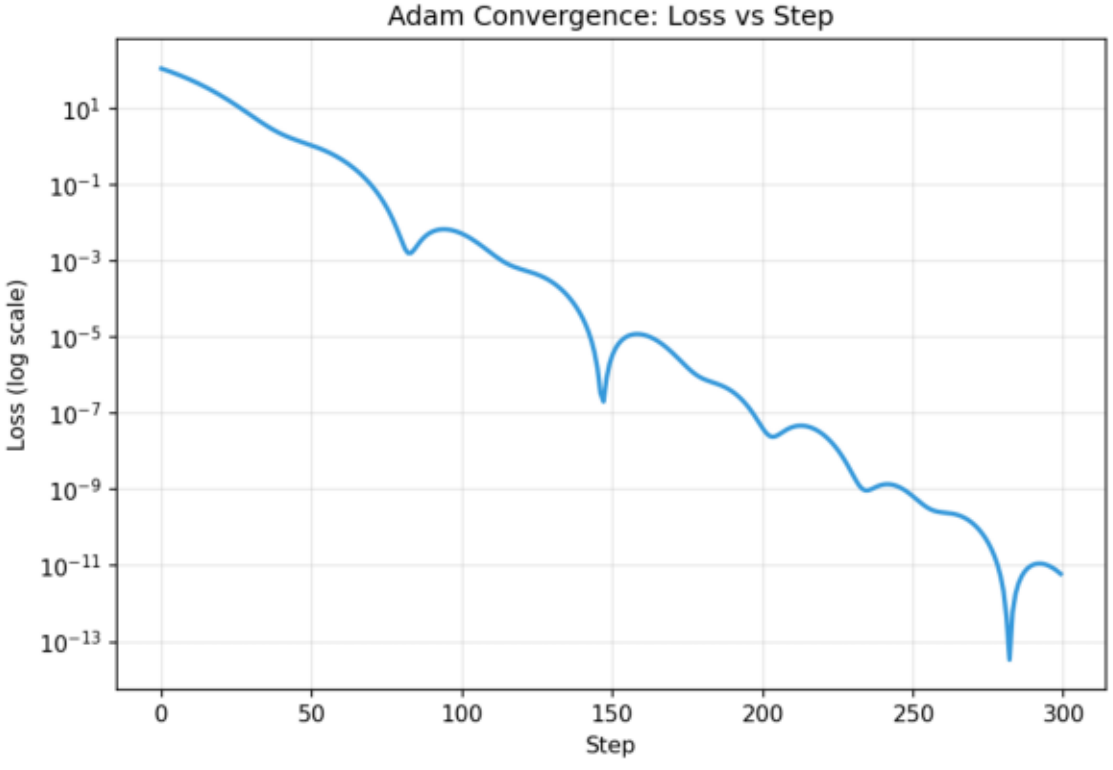
7. Bias correction is critical in early training steps -- without it, moment estimates are biased toward zero.

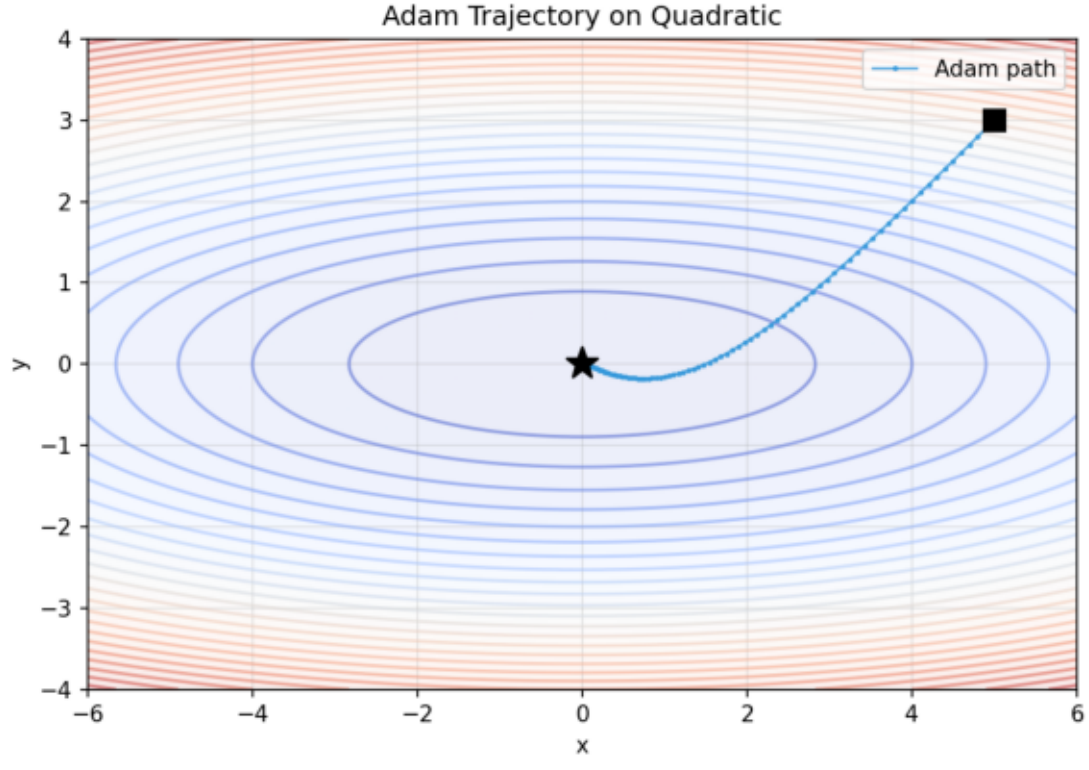
8. Warmup + cosine decay is the standard LR schedule for LLM training (GPT, LLaMA, Mistral).

01 Beale Trajectories



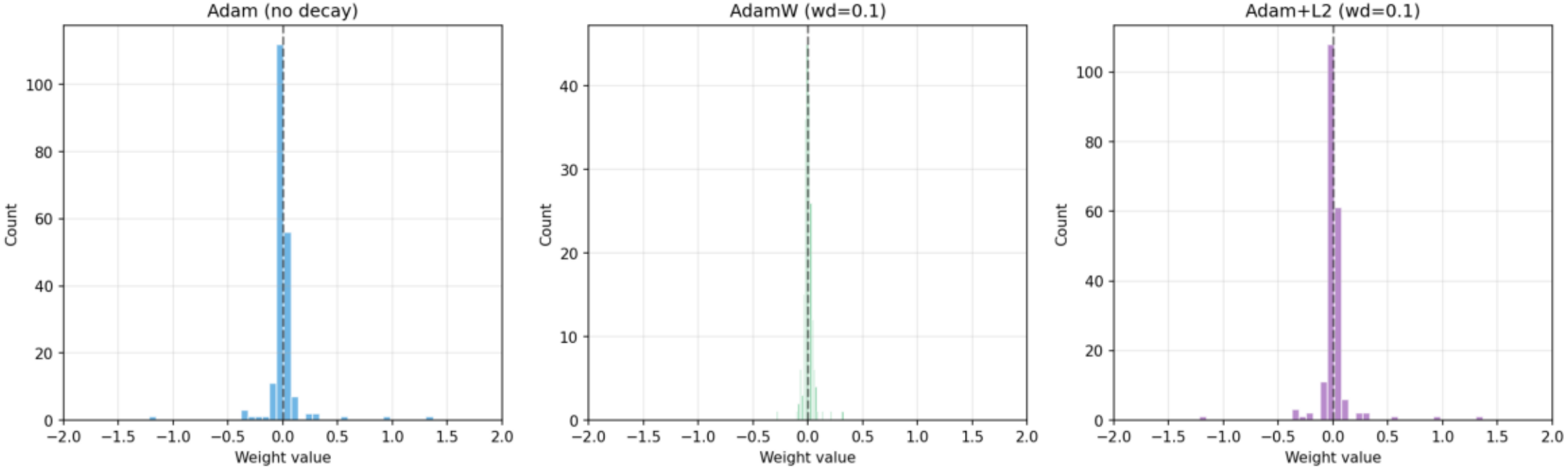
02 Adam Quadratic



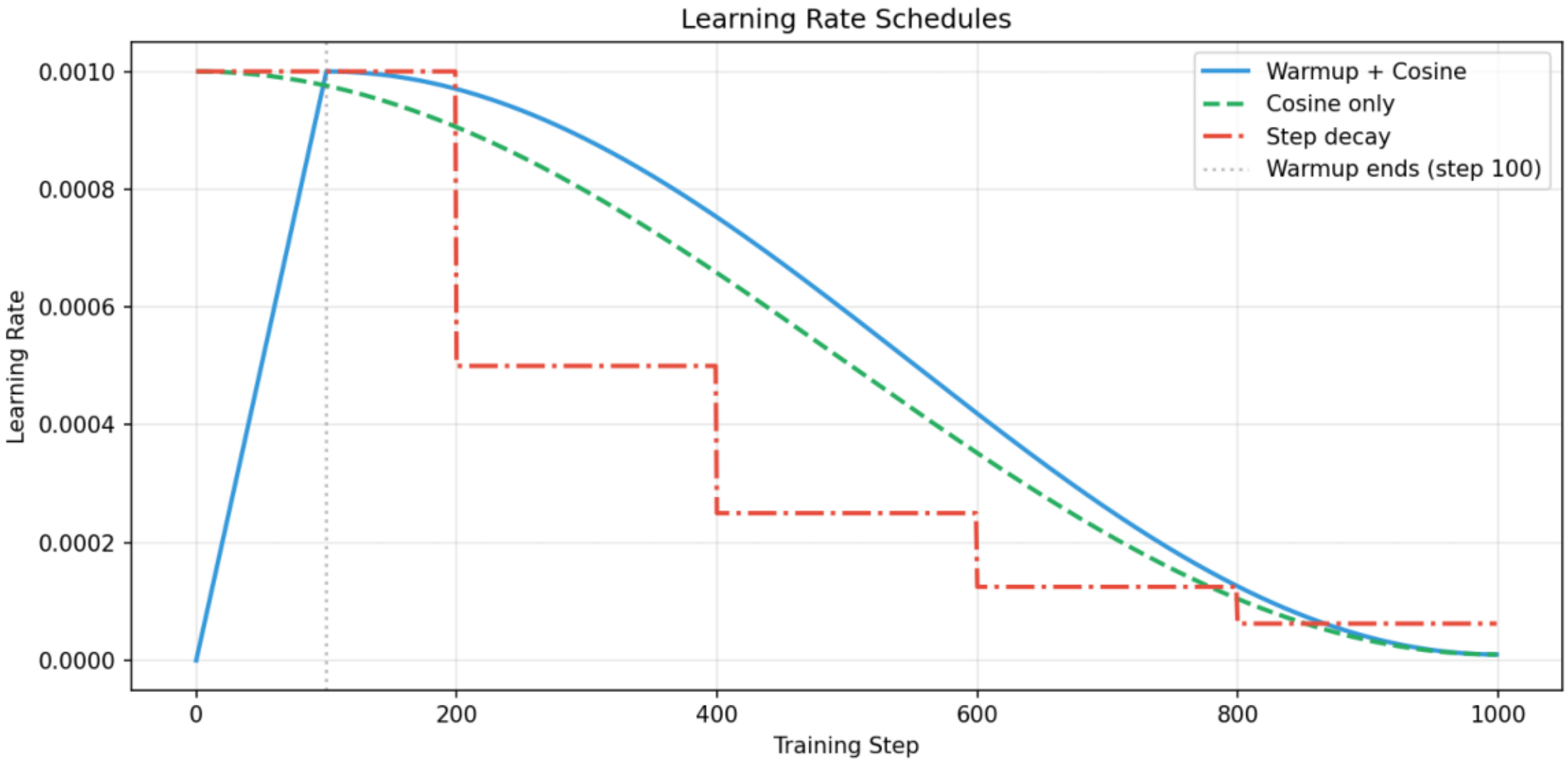


03 Adam Vs Adamw Weights

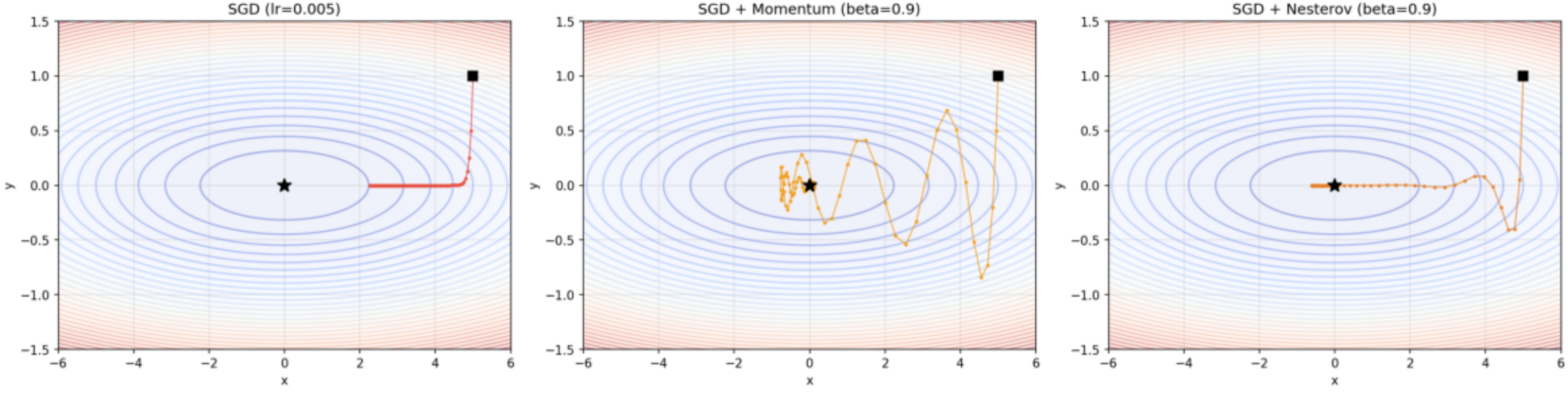
Weight Distributions After 1000 Steps of Noisy Gradient Descent



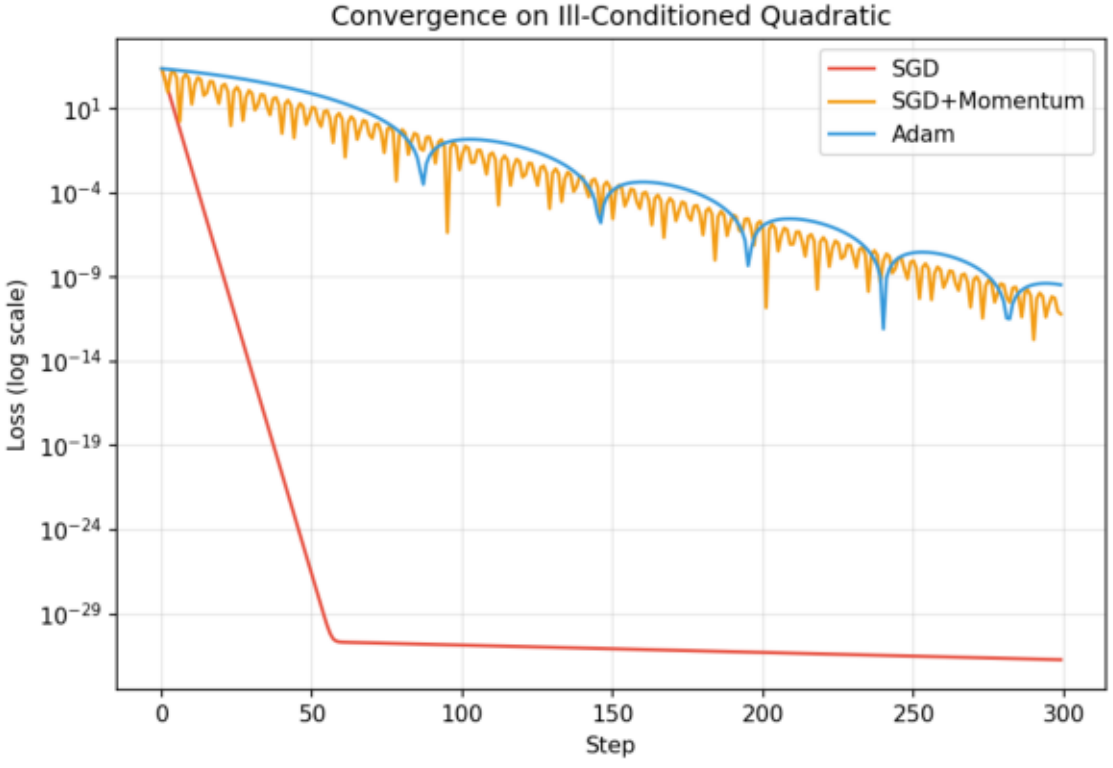
04 Lr Schedules

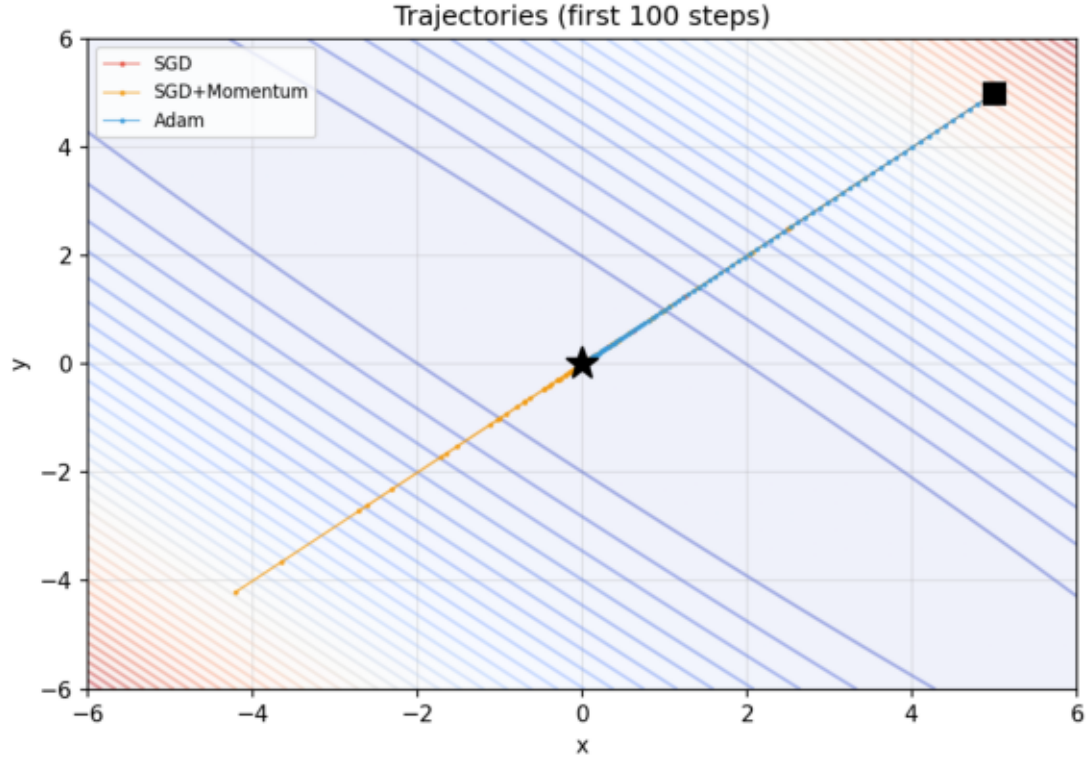


05 Momentum Ravine

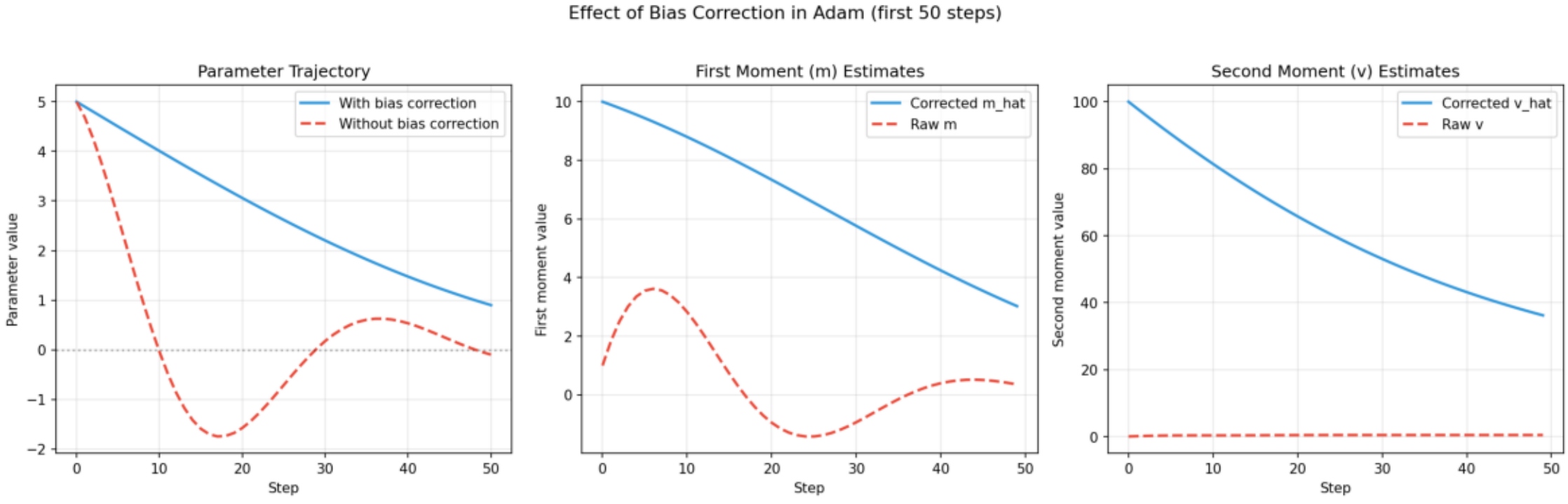


06 Illconditioned Quadratic





07 Bias Correction



07B Bias Correction Zoom

Bias Correction Zoomed: Steps 0-9