

Warm-up and Linear Schedule

Gradual Start with Progressive Decay

1 Warm-up Phase

- ↑ Gradual increase from low value
- 🛡️ Prevents large gradient updates from random initialization
- 📊 Typically first 5-10% of training

Schedule Visualization

LR increases during warm-up, then decreases linearly

2 Linear Decay Phase

- ⬇️ Linear decrease from peak to minimum
- 🎯 Remaining 90-95% of training

✨ Key Benefits

- ✓ Stabilizes training in early phase
- ✓ Especially effective with large batch sizes
- ✓ Essential for large-scale model training

⌚ Combined Schedule

Warm-up → Linear or Cosine Decay

🎯 Essential For

BERT

GPT

Vision Transformers