Learning Rate Revisited
Choosing a Learning Rate

- A good default is 1e-3
- But you should use the largest learning rate that trains
- Learning rate linearly related to batch size
Learning Rate Schedules

Decrease the learning rate during training

Step

Linear

Cosine

Cyclic
Optimization Algorithms
SGD + Momentum

- Usually works well
- Learning rate requires tuning

\[
m \leftarrow 0
\]

For N epochs:

For each batch B:

\[
g \leftarrow E_{x, y \sim B} \left[ \nabla_\theta \ell(x, y; \theta) \right]
\]

\[
m \leftarrow \rho m + g
\]

\[
\theta \leftarrow \theta - \gamma m
\]
AdaGrad

- Per-parameter learning rate
- Learning rate decays quickly

\[ v \leftarrow 0 \]

For N epochs:

For each batch B:

\[ g \leftarrow E_{x,y \sim B} \left[ \nabla_\theta \ell(x, y; \theta) \right] \]

\[ v \leftarrow v + g^2 \]

\[ \theta \leftarrow \theta - \gamma \frac{g}{\sqrt{v + \epsilon}} \]

Prevents division by zero

\( v \) grows quickly

\( v \) grows slowly
RMSProp

- Keep the per-parameter learning rate, but don’t decay overall learning rate
- Doesn’t work well with momentum
  - Often momentum = 0
- Good for some RL problems

\[
\begin{align*}
\text{For N epochs:} \\
\text{For each batch B:} \\
&\quad g \leftarrow \mathbb{E}_{x,y \sim B} [\nabla_{\theta} \ell(x, y; \theta)] \\
&\quad v \leftarrow \alpha v + (1 - \alpha) g^2 \\
&\quad m \leftarrow \rho m + \frac{g}{\sqrt{v + \epsilon}} \\
&\quad \theta \leftarrow \theta - \gamma m
\end{align*}
\]
Adam

- Generally good for small networks and small data
- Overfits more than SGD
- Missing some theoretical guarantees

For N epochs:

For each batch $B$:

\[ g \leftarrow E_{x,y \sim B}[\nabla_\theta \ell(x, y; \theta)] \]

\[ m \leftarrow \beta_1 m + (1 - \beta_1) g \]

\[ v \leftarrow \beta_2 v + (1 - \beta_2) g^2 \]

\[ \hat{m} \leftarrow m / (1 - \beta_1^t) \]

\[ \hat{v} \leftarrow v / (1 - \beta_2^t) \]

\[ \hat{\theta} \leftarrow \theta - \gamma \frac{\hat{m}}{\sqrt{\hat{v}} + \epsilon} \]

\[ t \leftarrow t + 1 \]
In Practice

• Large model, large data
  – SGD + Momentum
  – Also try Nesterov momentum
• Small model, small data
  – Adam
• Or, just try both and see which works better