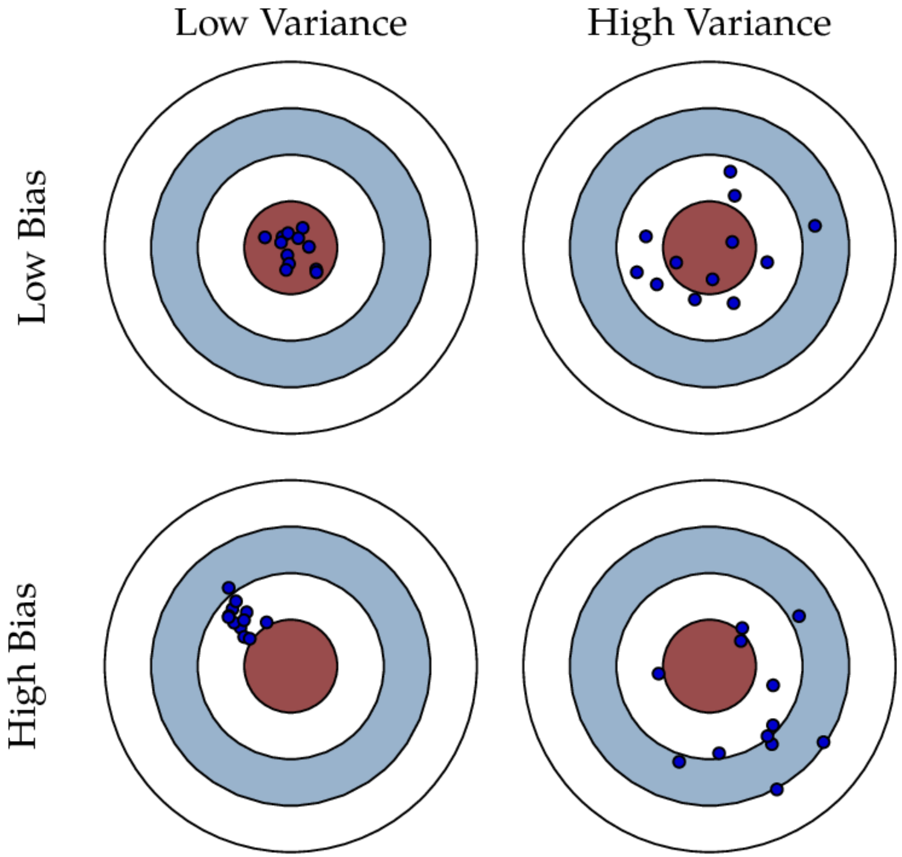
Bias and variance in gradients



Variance reduction

- If stochastic gradients have too high variance they are not usable
- To reduce variance we can apply variance reduction techniques
- The most popular method is to use control variates

Control variates

- We want to reduce the variance in estimating $f(\boldsymbol{x})$
- Assume we have a related function $h(\boldsymbol{x})$
  - For which we know analytically its expectation $\bar{h} = \mathbb{E}_{p_\varphi(\boldsymbol{x})}[h(\boldsymbol{x})]$
  - For instance, $h(\boldsymbol{x})$ can be the second order Taylor expansion $f(\boldsymbol{x})$
- Instead of estimating

$$\mathbb{E}_{p_\varphi(\boldsymbol{x})}[f(\boldsymbol{x})] \approx \hat{f} = \frac{1}{n}\sum_i f\left(\boldsymbol{x}^{(i)}\right), \boldsymbol{x}^{(i)} \sim p_\varphi(\boldsymbol{x})$$

- Substract the baseline and add the analytical expectation

$$\tilde{f} = \mathbb{E}_{p_\varphi(\boldsymbol{x})}[f(\boldsymbol{x}) - \beta h(x)] + \beta\, \bar{h}$$

  - In the limit the expectation will be the same as before, $\mathbb{E}[\tilde{f}] = \mathbb{E}[\hat{f}]$
  - However, the variance lower and reduction optimal for $\beta = -\frac{\mathrm{Cov(f,h)}}{\mathrm{Var(h)}}$

Straight-through gradients

- Often, gradients are hard or impossible to compute
  - For instance, if we have binary stochastic variables $\boldsymbol{z} \sim f(\boldsymbol{x}), \boldsymbol{z} \in \{0, 1\}$
  - If we compute the derivative **on the sample** we would have $\frac{dz}{dx} = 0$
  - $\boldsymbol{z}$ is a constant value (not a function)
- A popular alternative is straight-through gradients
  - We set the gradient is $\frac{dz}{dx} = 1$
  - Another alternative is to set the gradient $\frac{dz}{dx} = \frac{df}{dx}$
- Straight-through gradients introduce bias
  - our estimated gradient is different from the true gradient

Variance reduction in deep networks

- REBAR (Tucker et al.)
  - Low variance, unbiased gradient estimates for discrete latent variables
  - Inspired by REINFORCE and continuous relaxations
  - Removing the bias from the continuous relaxation
- RELAX (Grathwohl et al.)
  - Low variance, unbiased gradient estimates for black box functions
  - Applicable to discrete and continuous settings

Low bias low variance gradients

- Existing methods have troubles with deep Boolean stochastic nets
- Successive straight-through in multiple layers fails
  - Efficient but the bias accumulates over multiple layers
  - Optimization quickly gets stuck and learning stops
- Using unbiased estimates (REBAR, RELAX) is too inefficient
- Expand boolean networks with harmonic analysis (Fourier)
  - Bias and variance is caused by higher order coefficients
  - Manipulates those coefficients to reduce bias and variance
- Can train up to 80 layers instead of 2

Pervez, Cohen and Gavves, Low Bias Low Variance Gradient Estimates for Hierarchical Boolean Stochastic Networks

Summary

- Monte Carlo simulation
- Stochastic gradients
- MC gradient estimators
- Bias and variance in gradients

Reading material:

- All papers mentioned in the slides