

The simplest version of REINFORCE can be derived by simply differentiating the expected cost:

$$\mathbb{E}_z[J(\mathbf{y})] = \sum_{\mathbf{y}} J(\mathbf{y})p(\mathbf{y}) \quad (20.59)$$

$$\frac{\partial \mathbb{E}[J(\mathbf{y})]}{\partial \boldsymbol{\omega}} = \sum_{\mathbf{y}} J(\mathbf{y}) \frac{\partial p(\mathbf{y})}{\partial \boldsymbol{\omega}} \quad (20.60)$$

$$= \sum_{\mathbf{y}} J(\mathbf{y})p(\mathbf{y}) \frac{\partial \log p(\mathbf{y})}{\partial \boldsymbol{\omega}} \quad (20.61)$$

$$\approx \frac{1}{m} \sum_{\mathbf{y}^{(i)} \sim p(\mathbf{y}), i=1}^m J(\mathbf{y}^{(i)}) \frac{\partial \log p(\mathbf{y}^{(i)})}{\partial \boldsymbol{\omega}}. \quad (20.62)$$

Equation 20.60 relies on the assumption that J does not reference ω directly. It is trivial to extend the approach to relax this assumption. Equation 20.61 exploits the derivative rule for the logarithm, $\frac{\partial \log p(\mathbf{y})}{\partial \boldsymbol{\omega}} = \frac{1}{p(\mathbf{y})} \frac{\partial p(\mathbf{y})}{\partial \boldsymbol{\omega}}$. Equation 20.62 gives an unbiased Monte Carlo estimator of the gradient.

Anywhere we write $p(\mathbf{y})$ in this section, one could equally write $p(\mathbf{y} \mid \mathbf{x})$. This is because $p(\mathbf{y})$ is parametrized by $\boldsymbol{\omega}$, and $\boldsymbol{\omega}$ contains both $\boldsymbol{\theta}$ and \mathbf{x} , if \mathbf{x} is present.

One issue with the above simple REINFORCE estimator is that it has a very high variance, so that many samples of \mathbf{y} need to be drawn to obtain a good estimator of the gradient, or equivalently, if only one sample is drawn, SGD will converge very slowly and will require a smaller learning rate. It is possible to considerably reduce the variance of that estimator by using **variance reduction** methods (Wilson, 1984; L'Ecuyer, 1994). The idea is to modify the estimator so that its expected value remains unchanged but its variance get reduced. In the context of REINFORCE, the proposed variance reduction methods involve the computation of a **baseline** that is used to offset $J(\mathbf{y})$. Note that any offset $b(\boldsymbol{\omega})$ that does not depend on \mathbf{y} would not change the expectation of the estimated gradient because

$$E_{p(\mathbf{y})} \left[\frac{\partial \log p(\mathbf{y})}{\partial \boldsymbol{\omega}} \right] = \sum_{\mathbf{y}} p(\mathbf{y}) \frac{\partial \log p(\mathbf{y})}{\partial \boldsymbol{\omega}} \quad (20.63)$$

$$= \sum_{\mathbf{y}} \frac{\partial p(\mathbf{y})}{\partial \boldsymbol{\omega}} \quad (20.64)$$

$$= \frac{\partial}{\partial \boldsymbol{\omega}} \sum_{\mathbf{y}} p(\mathbf{y}) = \frac{\partial}{\partial \boldsymbol{\omega}} 1 = 0, \quad (20.65)$$