PSEUDOCODE FOR MINI-BATCH BASED STOCHASTIC GRADIENT DESCENT

Given data $\mathbf{x}_i, \mathbf{y}_i$, i=1…N, our loss function $L(f(\mathbf{x}_i;\boldsymbol{\theta}),\mathbf{y}_i)$, with parameters $\boldsymbol{\theta}$, a parameter regularization term $R(\boldsymbol{\theta})$ weighted by λ the pseudo code below can be used to optimize the average loss - also known as the empirical risk, plus the regularization term, i.e.

$$1/N \sum_{i=1}^{N} L(f(\mathbf{x}_i;\boldsymbol{\theta}),\mathbf{y}_i) + \lambda R(\boldsymbol{\theta}).$$

Given sets of indicies I_k for the assignment of examples into K mini-batches where the number of examples in each mini-batch is B_k, a learning rate η_t that potentially depends on the iteration, epoch or time t, and a gradient vector $\mathbf{g}$; we could formulate some fairly general pseudocode for mini-batch based stochastic gradient descent updates $\Delta\boldsymbol{\theta}$ using momentum as follows:

$\boldsymbol{\theta} = \boldsymbol{\theta}_o$ // initialize parameters
$\Delta\boldsymbol{\theta} = \mathbf{0}$
$\mathrm{t} = 0$
while converged == FALSE
  $\{\mathrm{I}_1,\ldots,\mathrm{I}_K\} = \mathrm{shuffle(X)}$ // create K mini-batches
  for $k = 1\ldots K$

$$\mathbf{g} = \frac{1}{B_k} \sum_{i \in \mathrm{I}_k} \left[\frac{\partial}{\partial \boldsymbol{\theta}} L(f(\mathbf{x}_i;\boldsymbol{\theta}),\mathbf{y}_i) \right] + \frac{B_k}{N} \lambda \frac{\partial}{\partial \boldsymbol{\theta}} R(\boldsymbol{\theta})$$

$$\Delta\boldsymbol{\theta} \leftarrow -\eta_t \mathbf{g} + \alpha \Delta\boldsymbol{\theta}$$

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \Delta\boldsymbol{\theta}$$

  end
  $\mathrm{t} = \mathrm{t} + 1$
end

It is common to perform the mini-batch creation before entering the while loop above; however, in some cases the shuffle within the while loop can lead to a better optimization.