1 Nonlinear Minimization By Implicit Gradient Descent

For a treatment of the numerous statistical advantages of implicit gradient descent to traditional gradient descent, we refer the reader to [1]. Suppose one wants to solve a system of equations of the form:

$$\begin{pmatrix} f_1(x) \\ \vdots \\ f_M(x) \end{pmatrix} = 0$$

where $f_l : R^d \to R$ for $1 \leq l \leq M$. Let

$$F(x) = \frac{1}{2}\sum_{l=1}^{M} |f_l(x)|^2$$

$$\nabla F(x) = \sum_{l=1}^{M} f_l(x)\nabla f_l(x)$$

and let $H_F(x)$ denote the Jacobian of ∇F at x, so that:

$$[H_F(x)]_{ij} = \frac{\partial(\nabla F)_i}{\partial x^j} = \frac{\partial}{\partial x^j}\sum_{l=1}^{M} f_l(x)\frac{\partial f_l}{\partial x^i}(x) = \sum_{l=1}^{M}(f_l(x)\frac{\partial f_l}{\partial x^i \partial x^j}(x)) + (\frac{\partial f_l}{\partial x^j}(x)\frac{\partial f_l}{\partial x^i}(x)) \approx$$

$$\approx \sum_{l=1}^{M}(\frac{\partial f_l}{\partial x^j}(x)\frac{\partial f_l}{\partial x^i}(x))$$

the choice to ignore the terms involving the second partial derivatives is often appropriate depending on the smoothness of the feature functions. Choosing to do so drastically reduces the cost of computation, and is recommended if reasonable bounds on the size of the second derivatives can be obtained.

Notice that solving the original system is equivalent to minimizing F. In explicit (or traditional) gradient descent, one would make an initial guess x_0 and then update according to:

$$x_{n+1} = x_n - \delta\nabla F(x_n)$$

The implicit update, on the other hand is:

$$x_{n+1} = x_n - \delta\nabla F(x_{n+1})$$

where one can approximate:

$$\nabla F(x_{n+1}) \approx \nabla F(x_n) + H_F(x_n)(x_{n+1} - x_n)$$

substituting in, rearranging, and solving for x_{n+1} one obtains the update:

$$x_{n+1} = x_n - \delta(I + \delta H_F(x_n))^{-1}\nabla F(x_n)$$

where I denotes the identity matrix. Notice that if the second derivative terms are ignored at each step, the computational cost only marginally exceeds that of traditional gradient descent, as only the first derivatives need to be computed in both cases.

References

[1] Toulis, Panos and Airoldi, E.M, 2016: Implicit stochastic gradient descent. Department of Statistics, Harvard University.