Supplementary material for: Partial-Hessian Strategies for Fast Learning of Nonlinear Embeddings

Max Vladymyrov and Miguel Á. Carreira-Perpiñán
EECS, School of Engineering, University of California, Merced

May 21, 2012

1 Proofs

We give a derivation of the result for the rate of linear convergence in p. 4 of the paper. Consider an objective function f to be minimized using a search direction obtained from $\mathbf{B}_k\mathbf{p}_k = -\nabla f(\mathbf{x}_k)$ where $\mathbf{B}_k = \mathbf{B}(\mathbf{x}_k)$ is a positive definite partial Hessian for $k = 0, 1, 2, \dots$ Under the assumptions of theorem 3.1 in the paper, $\mathbf{x}_k$ converges to a stationary point $\mathbf{x}^*$. Assume that $\mathbf{B}(\mathbf{x})$ is Lipschitz continuous in the region of interest (that is, $\exists L > 0$: $\|\mathbf{B}(\mathbf{x}) - \mathbf{B}(\mathbf{y})\| \le L\|\mathbf{x}-\mathbf{y}\|$ for any two points $\mathbf{x}$, $\mathbf{y}$ in the region) with bounded condition number, and that we take unit steps in the line search. Then:

$$\mathbf{x}_k + \mathbf{p}_k - \mathbf{x}^* = \mathbf{x}_k - \mathbf{x}^* - \mathbf{B}_k^{-1}\nabla f(\mathbf{x}_k) = \mathbf{B}_k^{-1}\left(\mathbf{B}_k(\mathbf{x}_k - \mathbf{x}^*) - (\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}^*))\right)$$

since $\nabla f(\mathbf{x}^*) = \mathbf{0}$. Applying Taylor's theorem (Nocedal and Wright, 2006, th. 2.1) to $\nabla f(\mathbf{x}_k)$ we have

$$\begin{aligned}\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}^*) &= \int_0^1 \nabla^2 f(\mathbf{x}^* + t(\mathbf{x}_k - \mathbf{x}^*))\,(\mathbf{x}_k - \mathbf{x}^*)\,dt \\ &= \int_0^1 \mathbf{B}(\mathbf{x}^* + t(\mathbf{x}_k - \mathbf{x}^*))\,(\mathbf{x}_k - \mathbf{x}^*)\,dt + \int_0^1 \left(\nabla^2 f(\mathbf{x}^* + t(\mathbf{x}_k - \mathbf{x}^*)) - \mathbf{B}(\mathbf{x}^* + t(\mathbf{x}_k - \mathbf{x}^*))\right)(\mathbf{x}_k - \mathbf{x}^*)\,dt.\end{aligned}$$

Hence:

$$\begin{aligned}&\|\mathbf{B}_k(\mathbf{x}_k - \mathbf{x}^*) - (\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}^*))\| = \\ &\left\|\int_0^1 (\mathbf{B}(\mathbf{x}_k) - \mathbf{B}(\mathbf{x}^* + t(\mathbf{x}_k - \mathbf{x}^*)))\,(\mathbf{x}_k - \mathbf{x}^*)\,dt - \int_0^1 \left(\nabla^2 f(\mathbf{x}^* + t(\mathbf{x}_k - \mathbf{x}^*)) - \mathbf{B}(\mathbf{x}^* + t(\mathbf{x}_k - \mathbf{x}^*))\right)(\mathbf{x}_k - \mathbf{x}^*)\,dt\right\| \\ &\le \left\|\int_0^1 (\mathbf{B}(\mathbf{x}_k) - \mathbf{B}(\mathbf{x}^* + t(\mathbf{x}_k - \mathbf{x}^*)))\,(\mathbf{x}_k - \mathbf{x}^*)\,dt\right\| + \left\|\int_0^1 \left(\nabla^2 f(\mathbf{x}^* + t(\mathbf{x}_k - \mathbf{x}^*)) - \mathbf{B}(\mathbf{x}^* + t(\mathbf{x}_k - \mathbf{x}^*))\right)(\mathbf{x}_k - \mathbf{x}^*)\,dt\right\| \\ &\le \int_0^1 \|\mathbf{B}(\mathbf{x}_k) - \mathbf{B}(\mathbf{x}^* + t(\mathbf{x}_k - \mathbf{x}^*))\|\,\|\mathbf{x}_k - \mathbf{x}^*\|\,dt + \int_0^1 \left\|\nabla^2 f(\mathbf{x}^* + t(\mathbf{x}_k - \mathbf{x}^*)) - \mathbf{B}(\mathbf{x}^* + t(\mathbf{x}_k - \mathbf{x}^*))\right\|\,\|\mathbf{x}_k - \mathbf{x}^*\|\,dt \\ &\le \int_0^1 L\,\|\mathbf{x}_k - \mathbf{x}^*\|^2\,t\,dt + \int_0^1 \left\|\nabla^2 f(\mathbf{x}^* + t(\mathbf{x}_k - \mathbf{x}^*)) - \mathbf{B}(\mathbf{x}^* + t(\mathbf{x}_k - \mathbf{x}^*))\right\|\,\|\mathbf{x}_k - \mathbf{x}^*\|\,dt \\ &= \frac{1}{2}L\,\|\mathbf{x}_k - \mathbf{x}^*\|^2 + \left(\int_0^1 \left\|\nabla^2 f(\mathbf{x}^* + t(\mathbf{x}_k - \mathbf{x}^*)) - \mathbf{B}(\mathbf{x}^* + t(\mathbf{x}_k - \mathbf{x}^*))\right\|\,dt\right)\|\mathbf{x}_k - \mathbf{x}^*\|\end{aligned}$$

so the upper bound contains a second-order term (on $\|\mathbf{x}_k - \mathbf{x}^*\|$) and a first-order term. The latter vanishes if using $\mathbf{B}_k = \nabla^2 f(\mathbf{x}_k)$ (i.e., Newton's method) and results in quadratic convergence. Otherwise, convergence is linear and we can estimate its rate as follows:

$$\begin{aligned}\|\mathbf{x}_k + \mathbf{p}_k - \mathbf{x}^*\| &\le \left\|\mathbf{B}_k^{-1}\right\|\,\|\mathbf{B}_k(\mathbf{x}_k - \mathbf{x}^*) - (\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}^*))\| \\ &\le \mathcal{O}(\|\mathbf{x}_k - \mathbf{x}^*\|^2) + \left\|\mathbf{B}_k^{-1}\right\|\left(\int_0^1 \left\|\nabla^2 f(\mathbf{x}^* + t(\mathbf{x}_k - \mathbf{x}^*)) - \mathbf{B}(\mathbf{x}^* + t(\mathbf{x}_k - \mathbf{x}^*))\right\|\,dt\right)\|\mathbf{x}_k - \mathbf{x}^*\|\,.\end{aligned}$$

When $\mathbf{x}_k - \mathbf{x}^*$ is small, the second-order term is negligible and the first-order term becomes approximately $r\,\|\mathbf{x}_k - \mathbf{x}^*\|$ with $r = \left\|\mathbf{B}^{-1}(\mathbf{x}^*)\nabla^2 f(\mathbf{x}^*) - \mathbf{I}\right\|$.

References

J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer-Verlag, New York, second edition, 2006.