Stochastic approximation theory for online learning with adaptive updates

P. Sunehag, J. Trumpf and N. Schraudolph
Statistical Machine Learning program, National ICT Australia
Locked bag 8001, 2601 ACT, Australia
email: Peter.Sunehag@nicta.com.au

January 19, 2007

In Machine Learning, optimization problems in which the true objective function $C(w)$ is defined as an expectation $E_z Q(z, w)$, are abundant. In practise, where we have a finite dataset, the empirical objective function $C_n(w) = \frac{1}{n}\sum_{i=1}^{n} Q(z_i, w)$ is optimized instead. Classical optimization techniques compute the entire sum and its gradient for every iteration. As available data sets grow ever larger, such "batch" optimizers therefore become increasingly inefficient. They are also ill-suited for the online (incremental) setting, where partial data must be modeled as it arrives.

Stochastic (online) gradient-based methods, by contrast, work with gradient estimates obtained from small subsamples (mini-batches) of training data. This can greatly reduce computational requirements on large, redundant data sets. Simple Stochastic Gradient Descent has proven more effective than sophisticated second-order batch methods (LeCun et al., 1998). Stochastic Meta-Descent (Schraudolph, 1999, 2002) further accelerates stochastic gradient descent through online adaptation of the update step, multiplying the stochastic gradient with a diagonal scaling matrix. A further step in this is taken by (Schraudolph et al., 2007) with an online version of the Broyden-Fletcher-Goldfarb-Shanno (BFGS) quasi-Newton method where the scaling matrix aims at approximating the inverse Hessian of the objective function.

We have looked at the theoretical foundations for using online updates that include scaling matrices, in particular the fundamental limits for applying the existing quasi-martingale framework (Fisk, 1965) and the super-martingale framework (Robbins and Siegmund, 1971) to establish convergence. (Bottou and LeCun, 2004) have previously presented results, based on (Fisk, 1965), where the scaling matrix is assumed to converge and they have remarked that bounds on the eigenvalues of the scaling matrix is the essential requirement to extend convergence guarantees beyond that. We need such extented results to deal with the online BFGS since we can not guarantee convergence of the scaling matrices for it. The author will present such requirements and their derivation from Robbins and Siegmunds theorem which says that

Theorem 0.1. *Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq ...$ be a sequence of sub σ-fields of $\mathcal{F}$. Let U_t, β_t, ξ_t and ζ_t, $t = 1, 2, ...$ be non-negative $\mathcal{F}_t$-measurable*

random variables such that

$$E(U_{t+1} \mid \mathcal{F}_t) \leq (1+\beta_t)U_t + \xi_t - \zeta_t,\ t = 1, 2, ... \tag{1}$$

Then on the set $\{\sum_t \beta_t < \infty, \sum_t \xi_t < \infty\}$, U_t *converges a.s. to a random variable and* $\sum_t \zeta_t < \infty$ *a.s.*

If we let ω^* be the minimizer of $C(w)$, the equation

$$E_t(\|\omega_{t+1} - \omega^*\|^2) = \|\omega_t - \omega^*\|^2 - 2a_t(\omega_t - \omega^*)^T \nabla_\omega C(\omega_t) + a_t^2 E(\|\nabla_\omega Q(z_t, \omega_t)\|^2) \tag{2}$$

connects the theorem to updates on the form $\omega_{t+1} = \omega_t - a_t \nabla_\omega Q(z_t, \omega_t)$ under the conditions $\sum a_t = \infty$, $\sum a_t^2 < \infty$ and $\inf_{(\tilde{w}-w^*)^T(\tilde{w}-w^*)>\epsilon}(\tilde{\omega} - \omega^*)^T \nabla_\omega C(\tilde{\omega}) > 0$ for all $\epsilon > 0$. The main problem with extending this to updates on the form $\omega_{t+1} = \omega_t - a_t B_t \nabla_\omega Q(z_t, \omega_t)$ where B_t is a positive and symmetric matrix is that the last of the three conditions becomes complicated by having a matrix inserted between $(\tilde{w} - w^*)^T$ and $\nabla_\omega C(\tilde{\omega})$. The maximum possible damage has to be assesed.

Some of the modifications of the BFGS algorithm that Schraudolph et al. (2007) used, relate closely to controlling the eigenvalues of the scaling matrix, and it will be discussed how these and other possible modifications relate to enforcing the derived conditions. The conditions and the corresponding modifications that are required depend to how large a class of objective functions we want to be able to optimize succesfully.

References

Y. LeCun, L. Bottou, G. Orr, and K. Müller. Efficient backprop. In *Neural Networks, Tricks of the trade*, pages 25–52, 1998.

N. N. Schraudolph. Local gain adaptation in stochastic gradient descent. In *Proc. Intl. Conf. Artificial Neural Networks*, pages 569–574, Edinburgh, Scotland, 1999. IEE, London.

N. N. Schraudolph. Fast curvature matrix-vector products for second-order gradient descent. *Neural Computation*, 14(7):1723–1738, 2002.

N. Schraudolph, J. Yu, and S. Günter. A stochastic quasi-newton method for online convex optimization. In *AISTAT*, San, Juan, Puerto Rico, 2007.

D. L. Fisk. Quasi-martingales. *Transactions of the AMS*, 3:369–389, 1965.

H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In J. Rustagi, editor, *Optimizing Methods in Statistics*, pages 233–257, New York, 1971. Academic.

L. Bottou and Y. LeCun. Online learning for very large datasets. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, 2004.