NATIONAL ICT AUSTRALIA LIMITED

Convergence Theory for Stochastic Quasi-Newton Methods

Foundations for using online BFGS

P. Sunehag, J. Trumpf, N. Schraudolph and L. Bottou (NEC labs, Princeton)

Introduction

At AISTATS 2007, Schraudolph et. al. present an online (L)-BFGS algorithm whose scaling matrices which do not always converge. Bottou has proved convergence for quasi-Newton methods with converging scaling matrices using strong stochastic Lyapunov functions. We here use weak stochastic Lyapunov functions to do the same for methods with scaling matrices which are only assumed to have uniformly bounded spectrum.

Optimization with "batch" methods

- Machine learning poses optimization problems with loss functions of the form $C(w) = E_{\mathbf{z}}L(\mathbf{z}, w)$. In reality an empirical average $C_n(w) = \sum_{i=1}^{n} L(\mathbf{z}_i, w)$ is minimized. "Batch" optimizers have to calculate the entire sum for each evalutation of $C_n(w)$ and $\nabla_w C_n(w)$.
- As data sets grow larger the "batch" methods becomes increasingly inefficient and they are ill-suited for the online setting.

Stochastic (online) gradient-based methods

- Stochastic gradient methods work with gradient estimates obtained from subsamples of the training data.
- On large redundant data sets simple stochastic gradient descent (SGD) typically outperfoms second order "batch" methods by orders of magnitude.
- $w_{t+1} = w_t - a_t Y_t$ where $E(Y_t) = \nabla_w C(w)$ and $a_t > 0$ defines SGD.
- For online Quasi-Newton methods like online (L)-BFGS, Natural Gradient and Kalman filters $w_{t+1} = w_t - a_t B_t Y_t$ where B_t is a positive scaling matrix.
- Stochastic Meta-Descent (SMD) uses diagonal scaling. For SGD B_t is always the identity matrix I.

Stochastic Approximation Theory

- The field of stochastic approximation was founded in 1951 by Herbert Robbins and Sutton Monro. It has influenced statistics, control, optimization and online learning. They wanted a root of a function $M(w)$ given a sequence $y_t = M(w_t) + \varepsilon_t$.
- The Robbins-Monro procedure is defined by $w_{t+1} = w_t - a_t y_t$ and w_t converges to the unique root θ that M is assumed to have if: $\sum a_t^2 < \infty$, $\sum a_t = \infty$, $M(w) > 0$ for $w > \theta$, $M(w) < 0$ if $w < \theta$. They also had a regularity condition for M and a noise model with uniformly bounded variation.
- That $|M(w)| \le C(|w-\theta|+1)$ and $\inf_{|w-\theta|>\delta} M(w)(w-\theta) > 0$ for all $\delta > 0$ is a sufficient regularity condition was proved by Blum in 1954.

Super-Martingales

In 1971, Robbins and Siegmund proved a super-martingale convergence theorem that implies convergence for the multivariate version of the Robbins-Monro procedure and therefore also for multivariate stochastic gradient descent. It is only neccesary to add a transpose to the last condition which results in the multivariate condition $\inf_{|w-\theta|>\delta} M(w)^T(w-\theta) > 0$ for all $\delta > 0$.

The Robbins-Siegmund Theorem

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

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

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.

Stochastic Lyapunov functions

- The Theorem above enables us to bring Lyapunov methods into the stochastic setting. To prove convergence for the multivariate Robbins-Monro procedure we choose $U(w) = \|w-\theta\|_2^2$ as the strong Lyapunov function and $U_t = U(w_t)$.
- $E_t(\|w_{t+1}-\theta\|^2) = \|w_t-\theta\|^2 - 2a_t(w-\theta)^T M(w_t) + a_t^2 E_t(\|Y_t\|^2)$.
- Assuming that $E_t(\|Y_t\|^2)$ is uniformly bounded is sufficent (but not necessary) to conclude that $\sum_t a_t^2 E_t(\|Y_t\|^2) < \infty$. The Robbins-Siegmund theorem then guarantees that $\|w_t-\theta\|^2$ converges almost surely and that $\sum_{t=1}^{\infty} a_t(w_t-\theta)^T M(w_t) < \infty$ almost surely.
- Since $\sum_{t=1}^{\infty} a_t = \infty$ and $\inf_{\|w-\theta\|>\delta}\{(w-\theta)^T M(w)\} > 0$ for all $\delta > 0$ we can conclude that $w_t \to \theta$ as $t \to \infty$ almost surely.

Stochastic approximation theory with scaling matrices

- Consider updates $w_{t+1} = w_t - a_t B_t Y_t$ where B_t is a random positive scaling matrix.
- If we try exactly the same method as above we will have the problem that we can not be certain that $(w_t-\theta)^T B_t M(w_t)$ is positive.
- That B is positive implies that $w^T B w > 0 \ge \lambda_{min} w^T w$ where λ_{min} is the smallest eigenvalue of B. It does **NOT** imply that $w^t B y \ge 0$ whenever $w^T y \ge 0$.
- Solution: Let $U(w) = C(w)$ if C is a cost function with $\nabla_w C(w) = M(w)$.
- This is the weak Lyapunov method.

A Convergence Theorem

Let $C(w)$ be a twice differentiable strictly positive cost function defined on $\mathbb{R}^n$. We study updates that depend on independent realizations $\mathbf{z}_t$ of a random variable $\mathbf{z}$.

$$w_{t+1} = w_t - a_t B_t Y(\mathbf{z}_t, w_t). \tag{2}$$

w_t converges to $\theta = \operatorname{argmin} C(w)$ almost surely under the following conditions:

- C.1. $E_{\mathbf{z}}(Y(\mathbf{z}, w_t)) = \nabla_w C(w)$ for all w.
- C.2. $\|\nabla_w^2 C(w)\| \le 2K$.
- C.3. $\inf_{C(w)-\inf C>\delta} \|\nabla_w C(w)\| > 0$ for all $\delta > 0$.
- C.4. $E_{\mathbf{z}}(\|Y(\mathbf{z}, w)\|^2) \le A + BC(w)$ for all w.
- C.5. B_t is positive for all t and all the eigenvalues are larger than m and smaller than M where $0 < m \le M < \infty$.
- C.6. $\sum a_t^2 < \infty$ and $\sum a_t = \infty$.

Proof sketch

- Since C is twice differentiable and has bounded Hessian (C.2) we can use Taylor expansion and the upper eigenvalue bound (C.5) to prove that

$$C(w_{t+1}) = C(w_t - a_t B_t Y_t) \le C(w_t) - a_t(\nabla_w C(w_t))^T B_t Y_t + KM^2 a_t^2 \|Y_t\|^2 \tag{3}$$

which implies, using (C.1) and (C.4) that

$$E_t(C(w_{t+1})) \le C(w_t) - a_t(\nabla_w C(w_t))^T B_t(\nabla_w C(w_t)) + KM^2 a_t^2(A + BC(w_t)). \tag{4}$$

- If we let $U_t = C(w_t)$ and merge the terms containing U_t it follows that

$$E_t(U_{t+1}) \le U_t(1 + a_t^2 BKM^2) - m a_t \|\nabla_w C(w_t)\|^2 + AKM^2 a_t^2. \tag{5}$$

- Since $\sum_t a_t^2 < \infty$ (C.6), the Robbins-Siegmund theorem can now be applied.
- It follows that $\sum_t a_t \|\nabla_w C(w_t)\|^2 < \infty$. Since $\sum a_t = \infty$ (C.6) it must be true that $\|(\nabla_w C(w_t))\|^2 \to 0$. (C.3) implies that $C(w_t) \to C(\theta) = \inf_w C(w)$ as $t \to \infty$.

Realization

Given B_t we can define modifications $\tilde{B}_t$ which satisfies given eigenvalue bounds.

- If you have the scaling matrices of the form $B_t = Q_t^T D_t Q_t$ where D_t is diagonal, Q_t is orthogonal and the diagonal entries of D_t are $d_{t,j}$, then we can define $\tilde{D}_t$ by letting its diagonal entries be $\tilde{d}_{t,j} = \max(m, \min(d_{t,j}, M))$ and then $\tilde{B}_t = Q_t^T \tilde{D}_t Q_t$.
- In the Online BFGS formula there is a constant $\lambda > 0$ that regularizes the updates in the sense that B_t is approximating $(H + \lambda I)^{-1}$ where I is the identity matrix instead of the inverse of the Hessian H. It forces the eigenvalues of B_t to be less than λ^{-1}. We can add a further modification $\tilde{B}_t = B_t + \gamma I$ where $0 < \gamma \le \lambda^{-1}$. The eigenvalues of $\tilde{B}_t$ lie in $[\gamma, \lambda^{-1}]$.