Curiously Fast Convergence of some Stochastic Gradient Descent Algorithms

Léon Bottou

January 23, 2009

1 Context

Given a finite set of m examples $z_1, \ldots, z_m$ and a strictly convex differentiable loss function $\ell(z, \theta)$ defined on a parameter vector $\theta \in \mathbb{R}^d$, we are interested in minimizing the cost function

$$\min_{\theta} \quad C(\theta) \;=\; \frac{1}{m} \sum_{i=1}^{m} \ell(z_i, \theta)\,.$$

One way to perform such a minimization is to use a stochastic gradient algorithm. Starting from some initial value $\theta[1]$, iteration t consists in picking an example $z[t]$ and applying the stochastic gradient update

$$\theta[t+1] = \theta[t] - \eta_t \frac{\partial \ell}{\partial \theta} \ell(z[t], \theta[t])\,,$$

where the sequence of positive scalars η_t satisfies the well known Robbins-Monro conditions $\sum_t \eta_t = \infty$ and $\sum_t \eta_t^2 < \infty$. We consider three ways to pick the example $z[t]$ at each iteration:

- *Random* Examples are drawn uniformly from the training set at each iteration.

- *Cycle* Examples are picked sequentially from the randomly shuffled training set, that is, $z[km + t] \;=\; z_{\sigma(t)}$, where σ is a random permutation of $\{1, \ldots, m\}$, and k is a nonnegative integer, and $t \in \{1, \ldots, m\}$.

- *Shuffle* Examples are still picked sequentially but the training set is shuffled before each pass, that is, $z[km + t] \;=\; z_{\sigma_k(t)}$, where the σ_k are random permutations of $\{1, \ldots, m\}$, and k is a nonnegative integer, and $t \in \{1, \ldots, m\}$.

With suitable assumptions on the function ℓ, the *random* case can be treated with well known stochastic approximation results [1, 5]. With gains of the form $\eta_t = c/(t + t0)$ and sufficiently large values of the constant c, one obtains results such as

$$\mathbb{E}\left[C(\theta[t]) - \min_\theta C(\theta)\right] \;\sim\; \frac{1}{t}\,,$$

where the expectation is taken over the random choice of examples at each iteration. Various theoretical works [2, 3, 6] indicate that no choice of η_t can lead to faster convergence rates than t^{-1}.

2 Experiments

We report now empirical results obtained with the three method.

The task is the classification of RCV1 documents belonging to class CCAT [4]. Each of the 781,265 examples is a pair composed of a 47,152 dimensional vector x_i representing a document and a variable $y_i = \pm 1$ representing its appartenance to the class CCAT. The parameter vector θ is also a 47,152 dimensional vector and the loss function is

$$\ell(x, y, \theta) = \log\left(1 + e^{-y\,(\theta.x)}\right)\,.$$

All experiments were achieved using a variant of the `svmsgd2` program and datasets.[1] The only modification consists in implementing our three schemes for selecting examples at each iteration.

Figure 1 shows log-log plots of the evolution of $C(\theta[t])$ as a function of the number of iterations. The slope of the curve indicates the exponent of the convergence of the algorithm.

- The *random* case displays a t^{-1} convergence as predicted by the stochastic approximation theory.
- The *cycle* case displays a $t^{-\alpha}$ convergence with α significantly greater than one. This means that this example selection strategy leads to a faster convergence. The exact value of α changes when we consider different permutations of the examples.
- The *shuffle* case displays a more chaotic convergence. A linear interpolation of the curve leads to an exponent α that is curiously close to

[1] `http://leon.bottou.org/projects/sgd`.

two, suggesting that we have an average t^{-2} convergence. This result is stable when we repeat the experiment with different permutations of the training set.

3 The Question

In light of the theoretical works associated with stochastic approximations, stochastic algorithms that converge faster than t^{-1} are very surprising.

In fact, the stochastic approximation results rely on randomness assumption on the successive choice of examples are independent. Both the *cycle* and the *shuffle* break these assumptions but provide a more even coverage of the training set.

What can we prove for the *cycle* and the *shuffle* cases?

References

[1] A. Benveniste, M. Metivier, and P. Priouret. *Algorithmes adaptatifs et approximations stochastiques*. Masson, 1987.

[2] K. Chung. On a stochastic approximation method. *Annals of Mathematical Statistics*, 25(3):463–484, 1954.

[3] V. Fabian. On asymptotic normality in stochastic approximation. *Annals of Mathematical Statistics*, 39(4):1327–1332, 1968.

[4] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *J. Machine Learning Research*, 5:361–397, 2004.

[5] L. Ljung and T. Söderström. *Theory and Practice of recursive identification*. MIT Press, Cambridge, MA, 1983.

[6] P. Major and P. Revesz. A limit theorem for the robbins-monro approximation. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 27:79–86, 1973.

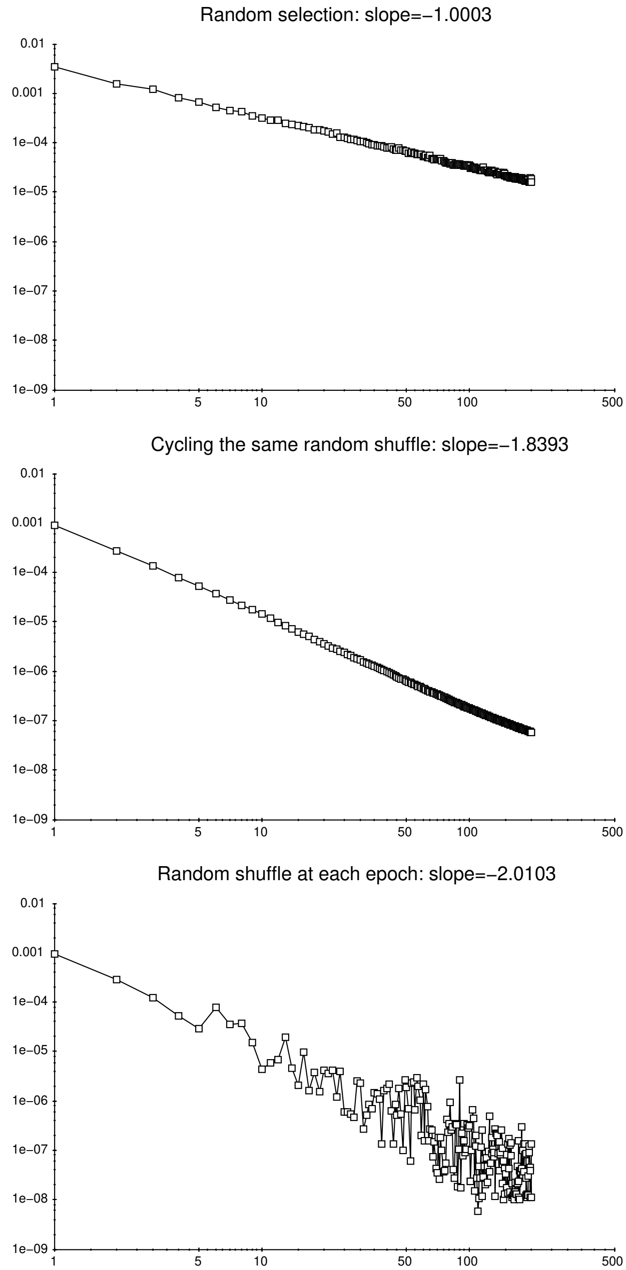

Figure 1: Evolution of $C(\theta[t])$ for our three example selection strategies. The horizontal axe counts the number of epoch. One epoch represents 781,265 iterations, that is, one pass over the training set.