Global Convergence of Stochastic Gradient Descent for Some Non-convex Matrix Problems

Christopher De Sa, Kunle Olukotun, and Chris Ré

cdesa@stanford.edu, kunle@stanford.edu, chrismre@stanford.edu

Departments of Electrical Engineering and Computer Science, Stanford University

Main Idea

We want to analyze SGD for matrix completion.

- ▷ Common problem in machine learning
- ▷ Used in industry by Oracle, MADLib, Twitter, etc

This problem appears in a variety of applications:

- ▷ matrix completion
- ▷ general data analysis
- ▷ subspace tracking
- ▷ PCA
- ▷ optimization
- ▷ recommendation systems.

Previous work: great local convergence results

- ▷ fast convergence if we initialize with SVD
- ▷ SGD known to converge in practice without initialization
- ▷ gap between theory and practice

Our contribution: This widely-used algorithm converges globally, using only random initialization!

- ▷ We also develop intuition for how to set the step size.

Matrix Completion Problem

Goal is to recover a low-rank matrix A using:

$$\begin{aligned}&\text{minimize}\ \ \mathbf{E}\left[\|\tilde{A}-X\|_F^2\right]\\&\text{subject to}\ \ X\in\mathbb{R}^{n\times n}, \mathbf{rank}\,(X)\le p, X\succeq 0,\end{aligned}$$

where $p \in \mathbb{Z}$ and $\tilde{A}$ is an unbiased sample of A. We can simplify this with a quadratic substitution $X = YY^T$ (Burer-Monteiro),

$$\begin{aligned}&\text{minimize}\ \ \mathbf{E}\left[\|\tilde{A}-YY^T\|_F^2\right]\\&\text{subject to}\ \ Y\in\mathbb{R}^{n\times p}\end{aligned}.$$

This leaves us with an unconstrained non-convex problem.

Algorithm Derivation

Stochastic gradient descent on quadratic decomposition:

$$Y_{k+1} = Y_k + \alpha_k\left(\tilde{A}_k - Y_kY_k^T\right)Y_k.$$

By choosing an appropriate Riemannian manifold, we can get

$$Y_{k+1} = \left(I + \eta_k\tilde{A}_k\right)Y_k\left(1 + \eta_kY_k^TY_k\right)^{-1}.$$

and if we ignore the radial component, we get the simple rule

$$Y_{k+1} = \left(I + \eta_k\tilde{A}_k\right)Y_k.$$

Alecton Solution Algorithm

Algorithm Alecton: Solve stochastic matrix problem

Require: $\eta \in \mathbb{R}$, $K \in \mathbb{N}$, $L \in \mathbb{N}$, and a sampling distribution $\mathcal{A}$

▷ ***Angular component (eigenvector) estimation phase***

Select Y_0 uniformly in $\mathbb{R}^{n\times m}$ s.t. $Y_0^TY_0 = I$.

for $k = 0$ **to** $K-1$ **do**

  Select $\tilde{A}_k$ independently from $\mathcal{A}$.

  $Y_{k+1} \leftarrow Y_k + \eta\tilde{A}_kY_k$

end for

$\hat{Y} \leftarrow Y_K\left(Y_K^TY_K\right)^{-\frac{1}{2}}$

▷ ***Radial component (eigenvalue) estimation phase***

$\bar{R} \leftarrow \frac{1}{L}\sum_{l=0}^{L-1}\hat{Y}^T\tilde{A}_l\hat{Y}$

return $\hat{Y}\bar{R}^{\frac{1}{2}}$

Algorithm description:

- ▷ "Angular phase" is equivalent to many algorithms:
  - **–** stochastic gradient descent
  - **–** stochastic power iteration
  - **–** stochastic proximal iteration
- ▷ "Radial phase" is maximum likelihood estimator, given result of angular phase.
- ▷ Both update phases are lightweight, and can be done in constant time if the sample is a single entry of A.

Main Contribution: Convergence Rate

To measure convergence, we let U be the projection matrix onto the column space of the solution X^*, and use the quantity

$$\rho_k = \min_{z\in\mathbb{R}^p}\|UY_kz\|^2/\|Y_kz\|^2.$$

For some $\epsilon > 0$, we say that the algorithm has failed to converge by time K if $\rho_t \le 1-\epsilon$ for all $t \le T$. We denote this event F_T.

We only require a bound on the second moment of the samples: for any $v \in \mathbb{R}^n$ and $w \in \mathbb{R}^n$, we require that, for some σ,

$$\mathbf{E}\left[(v^T\tilde{A}w)^2\right] \le \sigma^2\|v\|^2\|w\|^2.$$

If we choose any parameter $0 \le \chi \le 1$, set our step size

$$\eta = \frac{\Delta\epsilon\chi^2}{9\pi n\sigma^2p^2(p+\epsilon)},$$

where Δ denotes the spectral gap of A, and let

$$T = \frac{52\pi n\sigma^2p^3}{\Delta^2\epsilon\chi^3}\log\left(\frac{9\pi np}{2\chi^2\epsilon}\right) = \tilde{O}\left(\frac{\sigma^2n}{\Delta^2\epsilon}\right)$$

then the probability of failure after T steps is

$$P(F_T) \le \chi.$$

So, Alecton converges in linearithmic time with constant probability. (More details are in the paper.)

Many Applications

Entrywise sampling

- ▷ Each sample is a single entry of A.
- ▷ Entries are chosen independently and with equal weight.
- ▷ We need to impose an incoherence constraint for rapid convergence to be possible. (This is standard.)
- ▷ We can then bound the second moment of the sample with

$$\sigma^2 \le \mu^2(1-\mu^2)\|A\|_F^2.$$

- ▷ Each step very fast: write only mone row of Y.

Trace sampling

- ▷ We are given the value of v^TAw for random vectors v and w.
- ▷ For this sampling scheme, assuming $n > 50$,

$$\sigma^2 \le 20\|A\|_F^2.$$

Subspace sampling

- ▷ A is a projection matrix.
- ▷ For a random v in the column space of A, and random diagonal sampling matrices Q and R with $\mathbf{E}[Q] = \mathbf{E}[R] = I$, we use $\tilde{A} = Qvv^TR$.
- ▷ We can also bound the second moment of the sample here.

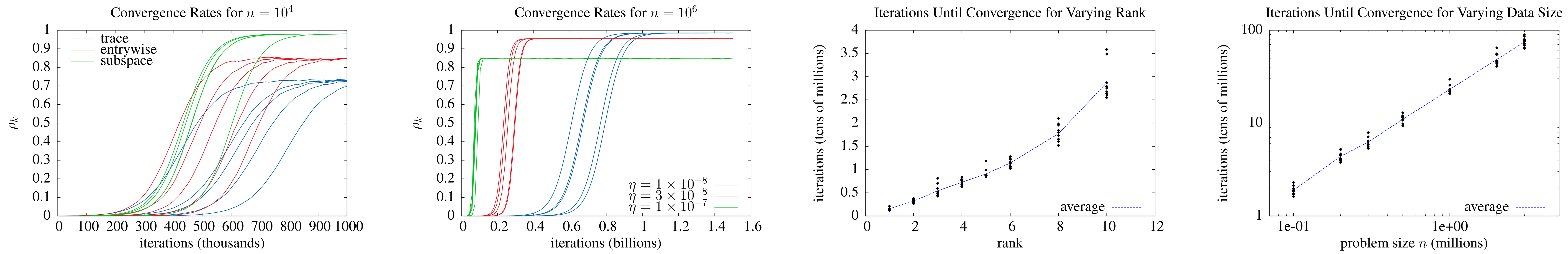
Noisy sampling

- ▷ Easy to handle noisy samples in any application.
- ▷ Can handle both additive and multiplicative noise.

Takeaway point: For all of the above applications, as long as the spectrum of A is fixed as n increases, the number of iterations required for convergence is only

$$T = O(\epsilon^{-1}n\log n).$$

Experiments



These plots show convergence of the angular phase of Alecton on synthetic datasets, varying sampling distribution, step size, problem rank, and problem size.