Stochastic algorithms under single spiked models

Emile Richard*

May 5, 2019

Abstract

We study SGD and Adam for estimating a rank one signal planted in matrix or tensor noise. The extreme simplicity of the problem setup allows us to isolate the effects of various factors: signal to noise ratio, density of critical points, stochasticity and initialization. We observe a surprising phenomenon: Adam seems to get stuck in local minima as soon as polynomially many critical points appear (matrix case), while SGD escapes those. However, when the number of critical points degenerates to exponentials (tensor case), then both algorithms get trapped. Theory tells us that at fixed SNR the problem becomes intractable for large d and in our experiments SGD does not escape this. We exhibit the benefits of warm starting in those situations. We conclude that in this class of problems, warm starting cannot be replaced by stochasticity in gradients to find the basin of attraction.

1 Introduction

Reductionism consists of breaking down the study of complex systems and phenomena into their atomic components. While the use of stochastic gradient based algorithms has shown tremendous success at minimizing complicated loss functions arising in deep learning, our understanding of why, when and how this happens is still limited. Statements such as *stochastic gradients escape from isolated critical points along the road to the best basin of attraction*, or *SGD generalizes better because it does not get stuck in steep local minima* still need to be better understood. Can we prove or replicate these phenomena in the simplest instances of the problem?

We study the behavior of stochastic gradient descent (SGD) [RM51] and an adaptive variant (Adam) [KB14] under a class of well studied non-convex problems. The single spiked models were originally designed for studying principal component analysis on matrices [Tao13, CDMF09, FP07] and have also been extended to higher order tensors [MR14]. Adaptive stochastic optimization methods have been gaining popularity in the deep learning community thanks to fast training on some benchmarks. However, it has been observed that despite reaching a low value of the loss function, the solutions found by Adam do not generalize as well as SGD solutions do. An assumption, widely spread and adopted in the community, has been that SGD's randomness helps escaping local critical points [WRS+17]. While the problem has been thoroughly studied theoretically [MR14, HSS15, HSSS16, BAGJ18], our contribution is to propose experimenting with this simple model to challenge claims such as those on randomized gradient algorithms in this very simple setup. It is noteworthy that the landscape of non-global critical points of these toy datasets

*Amazon

are studied [ABAC13, BAMMN17, BAGJ18] and formally linked to the neural nets empirical loss functions [BAMMN17, MM18]. For this problem, the statistical properties of the optimizers are well understood, and in the more challenging tensor situation, also the impact of (spectral) warm start has been discussed [MR14]. We will examine the solutions found by SGD and Adam and compare them with spectral and power methods. This allows to empirically elucidate the existence of multiple regimes: (1) the strong signal regime where all first order methods seem to find good solutions (2) when polynomially many critical points appear, in the matrix case, SGD converges while Adam gets trapped, unless if initialized in the basin of attraction (3) in the presence of exponentially many critical points (the tensor case), all algorithms fail, unless if d is moderately small and the SNR large enough to allow for proper initialization.

2 Single spiked models, and stochastic gradients

2.1 Matrix setup and known results

Even though proving strong results about non-convex loss functions is in general challenging, a class of nonconvex statistical problems is very well studied and relatively well understood. Principal component analysis (PCA) or finding the leading eigenvector of a covariance matrix is a problem of interest in statistics and machine learning. The proof of convergence of power method to the leading principal component and the geometry of critical points of the maximum likelihood problem

$$\text{maximize} \quad \langle \mathbf{u}, \mathbf{A}\mathbf{u} \rangle \quad \text{s.t.} \quad \|\mathbf{u}\|_2 = 1 \ , \qquad \text{Rayleigh quotient for matrix PCA}$$

are well established using eigenvalue decomposition. In addition, more recently, a class of extremely simplified models have shed light on the phase transitions of the problem difficulty as a function of the signal-to-noise ratio in the model. The so-called *single spiked models* consist of considering a symmetric normalized noise matrix to which a rank one signal is added.

$$\mathbf{A} = \lambda \, \mathbf{u_0}\mathbf{u_0}^\mathsf{T} + \mathbf{Z} \ , \quad \mathbf{Z}_{i,j} = \mathbf{Z}_{j,i} \sim \mathcal{N}(0, 1/d) \ \text{ and } \ \|\mathbf{u_0}\|_2 = 1 \qquad \text{Single spiked matrix}$$

It is known [CDMF09] that the spectrum of the noise matrix asymptotically forms a semi-circle situated between -2 and 2. When the signal to noise ratio is weak $\lambda \in [0, 1]$ then the signal dilutes in noise, while the leading principal component pops out of the semi-circle as soon as the signal to noise ratio λ is above the critical value $\lambda > \lambda_c = 1$, in which case the solution of the problem forms asymptotically a cosine value of $\left(1 - \lambda^{-2}\right)^{1/2}$ with the signal and the optimal value of the Rayleigh quotient is $\lambda + \lambda^{-1}$ [Tao13, CDMF09, FP07]. It is proven that the power method allows to obtain the solution after logarithmically many steps, as a function of the problem dimension d.

We will minimize the unconstrained objective function $\ell(u) = -\frac{1}{2}\langle \mathbf{A}, \mathbf{u}\mathbf{u}^\mathsf{T} \rangle + \frac{1}{2}\gamma \|\mathbf{u}\|_2^2$. We set the value of γ to the theoretical asymptotic value of the leading eigenvalue, $\gamma = 2$ for $\lambda < 1$ or $\lambda + \lambda^{-1}$ for larger λ, and will add random normal noise to the gradient for stochasticity: $\nabla_\sigma \ell(\mathbf{u}) = -\mathbf{A}\mathbf{u} + \gamma \mathbf{u} + \sigma \mathbf{z}$ where $\mathbf{z}_i \sim \mathcal{N}(0, 1/d)$. This function has a constant Hessian $\mathsf{H}_\ell(\mathbf{u}) = -\lambda \mathbf{u_0}\mathbf{u_0}^\mathsf{T} - \mathbf{Z} + \gamma \mathbf{I}_d$ which is positive semi-definite as soon as γ is equal or larger than the value of the leading eigenvalue of $\mathbf{A}$.

2.2 Extension to order 3 tensors and a few results

The tensor version of the problem (see [MR14] for notations and more discussion on problem setting)

$$\text{maximize} \quad \langle \mathbf{A}, \mathbf{u}^{\otimes 3} \rangle \ \text{ s. t. } \quad \|\mathbf{u}\|_2 = 1 \qquad \text{Rayleigh quotient tensor PCA}$$

under the tensor single spiked model defined for a symmetric (π is a permutation of 3 elements)

$$\mathbf{A} = \lambda \, \mathbf{u_0}^{\otimes 3} + \mathbf{Z} \ , \quad \mathbf{Z}_{i,j,k} = \mathbf{Z}_{\pi(i,k,j)} \sim \mathcal{N}(0, 1/d) \ \text{ and } \ \|\mathbf{u_0}\|_2 = 1 \qquad \text{Single spiked tensor}$$

presents additional ramifications. Connections between deep learning loss functions and an analogous model in absence of signal ($\lambda = 0$) have first been pointed by [CHM$^+$15] and more recently by [MM18]. Here we discuss the case where signal is present in addition to noise (i.e. $\lambda > 0$) in order to also study the interplay between signal strength λ and problem dimension d. While a requirement for obtaining a solution correlated with the signal is a constant threshold $\lambda > \lambda_c = \mathcal{O}(1)$ is established in [MR14] Theorem 1, the problem's numerical tractability also depends on the value of d in addition to the signal to noise ratio and requiring $\lambda \gtrsim d^{1/4}$. This phenomenon is conjectured to be related to the exponential growth of the number of critical points and their distribution around and orthogonal to the solution [BAMMN17], as opposed to the matrix case where the number of critical points is equal to the number of eigenvalues of a symmetric matrix, d. The stochastic gradients are taken from the objective function $\ell(\mathbf{u}) = -\frac{1}{6}\langle \mathbf{A}, \mathbf{u}^{\otimes 3} \rangle + \frac{1}{2}\gamma \|\mathbf{u}\|_2^2$, and the Hessian, in the single spiked model has a simple expression $\mathsf{H}_\ell(\mathbf{u}) = -\lambda \langle \mathbf{u_0}, \mathbf{u} \rangle \mathbf{u_0}\mathbf{u_0}^\mathsf{T} - \mathbf{Z}\mathbf{u} + \gamma \mathbf{I}_d$ that we use to discuss optimization problem's properties in numerical experiments. We call *spectral initialization* a (first order) algorithm that initiates at the left singular vector of the unfolded tensor $\mathbf{A}$, see `sp.sgd` etc. in numerical experiment.

How does data abundance explain the success of first order methods? On the positive side, we discuss that in this model and considering a large dataset of i.id. samples, weak signals in individual observations accumulate and allow to solve the problem if $n \gtrsim \sqrt{d}$. The counter part to the strong requirement $\lambda \gtrsim d^{1/4}$ (conjectured in [MR14] and proven in [HSS15, HSSS16]) is that accumulation of observations compensate low signal to noise ratio in each individual sample. Formally,

Remark 2.1. Assume n sample of data according to model "Single spiked tensor", with the same signal $\mathbf{u_0}$ and different i.i.d. noises $\mathbf{Z}_q$ are observed:

$$\text{For} \quad q = 1, \cdots, n, \quad \mathbf{A}_{(q)} = \lambda \, \mathbf{u_0}^{\otimes 3} + \mathbf{Z}_{(q)} \ , \quad \mathbf{Z}_{(q)_{i,j,k}} = \mathbf{Z}_{(q)_{\pi(i,k,j)}} \sim \mathcal{N}(0, 1/d) \ \text{ i.i.d. } . \qquad (1)$$

There exists constants c_0, c_1 such that if $n \geq c_0\sqrt{d}$, then, warm started power iteration produces a vector $\mathbf{u}$, such that with high probability $\langle \mathbf{u_0}, \mathbf{u} \rangle > 1 - c_1/\lambda$.

This result is established using Theorems 5 in [MR14] and 6.3 in [HSSS16] and considering the average tensor

$$\bar{\mathbf{A}} = \frac{1}{\sqrt{n}} \sum_{q=1}^{n} \mathbf{A}_{(q)} = \lambda_n \, \mathbf{u_0}^{\otimes 3} + \bar{\mathbf{Z}} \ .$$

Since $\bar{\mathbf{Z}}$ is symmetric and $\bar{\mathbf{Z}}_{i,j,k} \sim \mathcal{N}(0, 1/d)$, the tensor $\bar{\mathbf{A}}$ is sampled from a similar distribution as Single spiked tensor with a SNR $\lambda_n = \sqrt{n}\lambda$. This means that the requirement $\lambda \gtrsim d^{1/4}$ [HSSS16],

in the average tensor case, relaxes to $n \gtrsim \sqrt{d}$. In words, this means that if we are solving a problem with a tensor PCA complexity, and if the number of i.i.d. observations grows quadratically as a function of the problem dimension, we can compute the solution reliably using spectral warm start, even though the original problem looks intractable.

3 Numerical results

Our numerical results report performance of different algorithms at solving simulations of matrix and tensor PCA problems. Under various problem generation parameter choices, we report values of

- `cosine` or $\langle \mathbf{u}, \mathbf{u_0} \rangle$. This is measures the quality of planted (hidden) signal recovery from the noisy observation. Higher values are preferred, and it cannot exceed 1, since both $\mathbf{u_0}$ and $\mathbf{u}$ are normalized. This quantity is to be qualitatively compared with the test error in standard learning problems where the true value of the parameter is not available.
- `Rayleigh` is the value of the log-likelihood objective function that we are maximizing. Higher values are preferred. The theoretical maximum value of this objective is the operator norm of the observed tensor $\mathbf{A}$. This is comparable with (minus) the training loss.

The signal to noise ratio λ is to be compared with the number of observations in a supervised learning problem. The stochasticity of the gradients σ is to be compared with the number of sample points in each minibatch of data: large stochasticity mimics small minibatch situations.

3.1 Matrix PCA results

In Figure 1 we plot the values of the objective function or Rayleigh quotient and the cosine of the ground truth with the solution as a function of the iterates. We replicated these plots at values of the SNR parameter $\lambda < \lambda_c = 1$, at the critical value $\lambda = 1$ and above it for $\lambda = 2$ where the problems is considered to be easy. The learning rate and stochasticity σ were set by generating instances of the problem with different noise matrices. These plots allow to compare the optimization power of different algorithms and also keep track of the quality of the solution found. We can see in these plots that Adam gets stuck around the wrong region very fast. Note that with the value (or larger) of $\gamma = 2$ for $\lambda < 1$ and $\gamma = \lambda + \lambda^{-1}$ for $\lambda \geq 1$ that we can set given the true value of λ, and knowing the concentration around the asymptotics [Tao13, CDMF09, FP07], the objective function is strongly convex so we expect first order methods to show the same convergence rates as the power method. One can also observe that gradient descent corresponds to performing power iteration on a shifted matrix $\mathbf{A} + \alpha \mathbf{I}_d$. Figure 2 shows the value of the objective and the correlation with the ground truth as a function of SNR λ. We can see that SGD is superior to Adam, uniformly along λ, while power method (also a first order method) rivals with SGD. These plots exhibit the instability around and below the critical value $\lambda_c = 1$ while above λ_c the behavior is more stable.

3.2 Tensor PCA results

In the tensor setting we experimented with spectral initialization. Spectral initialization consists of flattening the tensor to a $d \times d^2$ matrix and initializing tensor algorithms with the left singular vector

of the flattened tensor. We observe benefits of spectral initialization at locating the initial point in a basin of attraction that leads to better solutions when λ is large enough. For $\lambda = 1.0$ spectral initialization does not result in better estimates in Figure 3, while for larger values of λ we can see the benefit of warm start. In Figure 4 we also plot values of the gradient and the number of positive eigenvalues of the Hessian along the optimization iterates for $d = 100, \lambda = 2, \gamma = 2$. We observe that spectral initialization located the initial point of the iterations in the basin of attraction where the problem is convex (all eigenvalues of the Hessian are positive). Adam, while starting in this region, fails at finding a solution as good as SGD's. We experimented with the amount of noise added to each gradient evaluation and mapped median values of the estimate and optimization problems for values of λ and σ over 100 instances of the problem generated with different noise matrices and stochastic gradients. Stochasticity does not seem to remedy to the problem difficulty. Numerical experiments suggest that irrespective of the magnitude of the stochastic component added to the gradient, the first order methods, initialized at random, fail at finding the best basin of attraction. In the same setup, spectral initialized first order methods successfully find the solutions.

4 Conclusion

We propose to study algorithms used for minimizing deep learning loss functions, at optimizing a non-convex objective on simple synthetic datasets. Studying simplified problems has the advantage that the problem's properties, and the behavior of the optimizer and the solution, can be studied rigorously. The use of such datasets can help to perform sanity checks on improvement ideas to the algorithms, or to mathematically prove or disprove intuitions. The properties of the toy data sets align with some properties of deep learning loss functions. From the optimization standpoint, the resulting tensor problems may appear to be even harder than deep learning problems. We observe that finding good solutions is hard unless if proper initialization is performed, while the value of stochasticity in gradient estimates seems too narrow and does not appear to compensate for poor initialization heuristics.

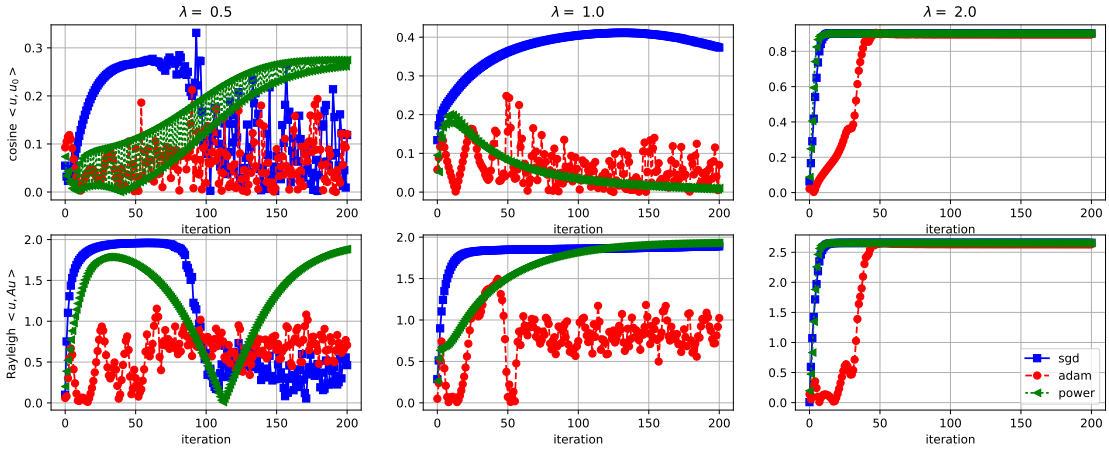

Figure 1: Matrix setup: optimization of the objective function and evolution of the estimation problem along the iterations. Top row: correlation of the solution with the signal, or $\langle \mathbf{u}, \mathbf{u_0} \rangle$, bottom row is the value of the objective function $\langle \mathbf{u}, \mathbf{A}\mathbf{u} \rangle / \|\mathbf{u}\|_2^2$. Each column represents the values of those quantities along iterations of the algorithm for a fixed λ. We can see that at high SNR $\lambda = 2$ all algorithms succeed at optimizing the objective and estimating the parameter, while Adam underperforms SGD and power method at maximizing the objective in low signal regimes $\lambda \leq 0$.

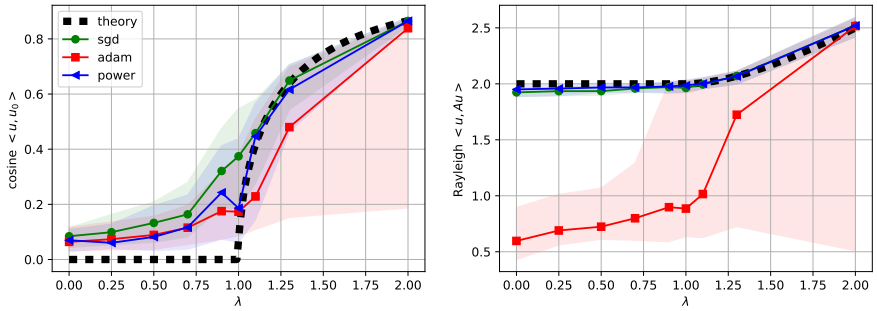

Figure 2: Correlation with the truth (left) and value of the optimization problem (left) in the matrix problem, as a function of the SNR. The certainty surface plotted show variations of solutions found by each method over multiple instances of the generative model. Adam underperforms unless if SNR is very large.

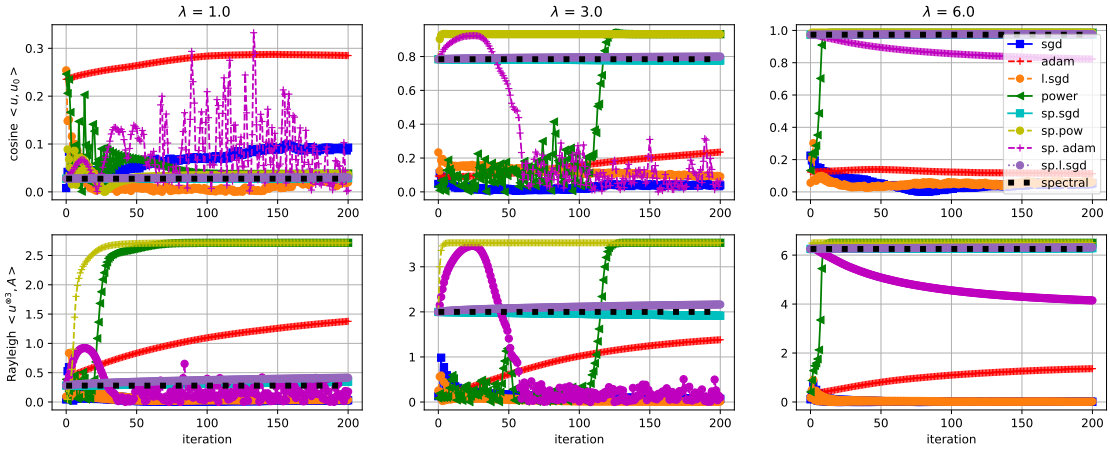

Figure 3: Tensor setup: optimization of the objective function and evolution of the estimation problem along the iterations. Top row: correlation of the solution with the signal (cosine, or $\langle \mathbf{u}, \mathbf{u_0} \rangle$), bottom row is the value of the Rayleigh objective function $\langle \mathbf{u}^{\otimes 3}, \mathbf{A} \rangle / \|\mathbf{u}\|_2^3$. Each column represents the values of those quantities along iterations of the algorithm. The prefix `sp.` refers to spectral initialization and `l.` refers to a decreasing learning weight scheduled in $1/\sqrt{t}$. We observe the value of warm starting as soon as λ is large enough. Even at high SNR $\lambda = 6$, randomly initialized SGD fails while spectrally initialized SGD succeeds. Adam drifts to a non optimal critical point in that regime, even with spectral warm start.

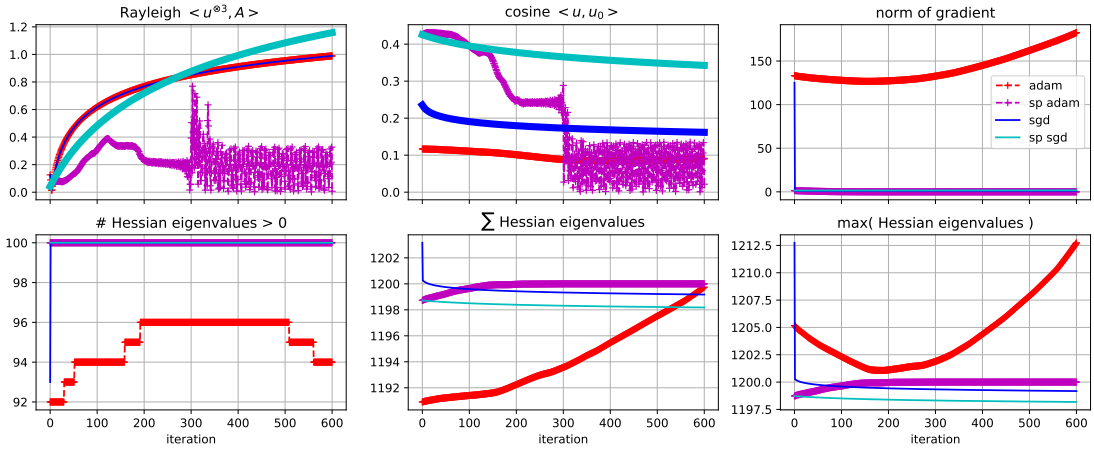

Figure 4: Tensor PCA. Value of the problem, estimate quality, gradient norm, number of positive Hessian eigenvalues, sum of Hessian eigenvalues as a function of iterates. Randomly initialized Adam (in red) seems to find a solution that has negative Hessian eigenvalues (saddle point) and very high maximum eigenvalue (very steep), while warm start seems to locate the initial point inside a region with convex neighborhoods.

References

[ABAC13] Antonio Auffinger, Gerard Ben Arous, and Jiri Cerny, *Random matrices and complexity of spin glasses*, Communications on Pure and Applied Mathematics **66(2)** (2013), 165–201.

[BAGJ18] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath, *Algorithmic thresholds for tensor pca*, arXiv preprint arXiv:1808.00921 (2018).

[BAMMN17] Gerard Ben Arous, Song Mei, Andrea Montanari, and Mihai Nica, *The landscape of the spiked tensor model*, arXiv preprint arXiv:1711.05424 (2017).

[CDMF09] Mireille Capitaine, Catherine Donati-Martin, and Delphine Féral, *The largest eigenvalues of finite rank deformation of large wigner matrices: convergence and nonuniversality of the fluctuations*, The Annals of Probability **37** (2009), no. 1, 1–47.

[CHM+15] Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun, *The loss surfaces of multilayer networks*, Artificial Intelligence and Statistics, 2015, pp. 192–204.

[FP07] Delphine Féral and Sandrine Péché, *The largest eigenvalue of rank one deformation of large wigner matrices*, Communications in mathematical physics **272** (2007), no. 1, 185–228.

[HSS15] Samuel B Hopkins, Jonathan Shi, and David Steurer, *Tensor principal component analysis via sum-of-square proofs*, Conference on Learning Theory, 2015, pp. 956–1006.

[HSSS16] Samuel B Hopkins, Tselil Schramm, Jonathan Shi, and David Steurer, *Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors*, Proceedings of the forty-eighth annual ACM symposium on Theory of Computing, ACM, 2016, pp. 178–191.

[KB14] Diederik P Kingma and Jimmy Ba, *Adam: A method for stochastic optimization*, arXiv preprint arXiv:1412.6980 (2014).

[MM18] Marco Mondelli and Andrea Montanari, *On the connection between learning two-layers neural networks and tensor decomposition*, arXiv preprint arXiv:1802.07301 (2018).

[MR14] Andrea Montanari and Emile Richard, *A statistical model for tensor pca*, Advances in Neural Information Processing Systems, 2014, pp. 2897–2905.

[RM51] Herbert Robbins and Sutton Monro, *A stochastic approximation method*, The annals of mathematical statistics (1951), 400–407.

[Tao13] Terence Tao, *Outliers in the spectrum of iid matrices with bounded rank perturbations*, Probability Theory and Related Fields **155** (2013), no. 1-2, 231–263.

[WRS+17] Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht, *The marginal value of adaptive gradient methods in machine learning*, Advances in Neural Information Processing Systems, 2017, pp. 4148–4158.