UNIVERSALITY IN HALTING TIME

Levent Sagun
Mathematics Department
New York University
sagun@cims.nyu.edu

Thomas Trogdon
Mathematics Department
University of California, Irvine
ttrogdon@math.uci.edu

Yann LeCun
Computer Science Department
New York University
yann@cs.nyu.edu

ABSTRACT

The authors present empirical distributions for the halting time (measured by the number of iterations to reach a given accuracy) of optimization algorithms applied to two random systems: spin glasses and deep learning. Given an algorithm, which we take to be both the optimization routine and the form of the random landscape, the fluctuations of the halting time follow a distribution that, after centering and scaling, remains unchanged even when the distribution on the landscape is changed. We observe two main classes, a Gumbel-like distribution that appears in Google searches, human decision times, QR factorization and spin glasses, and a Gaussian-like distribution that appears in conjugate gradient method, deep network with MNIST input data and deep network with random input data. This empirical evidence suggests presence of a class of distributions for which the halting time is independent of the underlying distribution under some conditions.

1 INTRODUCTION

In this paper we discuss both the presence and application of universality in optimization algorithms. More precisely, in order to optimize an energy functional when the functional itself and the initial guess are random, we consider the following iterative algorithms: conjugate gradient for solving a linear system, gradient descent for spin glasses, and stochastic gradient descent for deep learning.

A bounded, piecewise differentiable random field (See Adler & Taylor (2009) for an account on the connection of random fields and geometry), where the randomness is non-degenerate, yields a landscape with many saddle points and local minima. Given such a landscape and a moving particle that takes steps to reach a low-energy level, an essential quantity is the time the particle takes until it stops which we call the *halting time*. Many useful bounds on the halting time are known for convex cases, where the stopping condition produces a halting time that is, essentially, the time to find the minimum. In non-convex cases, however, the particle knows only the information that can be calculated locally. And a locally measurable stopping condition, such as the norm of the gradient at the present point, or the difference in altitude with respect to the previous step, can lead the algorithm to locate a local minimum. This feature allows the halting time to be calculated in a broad range of non-convex, high-dimensional problems. A prototypical example of such a random field is the class of polynomials with random coefficients. Spin glasses and deep learning cost functions are then special cases of such fields that yield different landscapes. Polynomials with random coefficients are not only a broad class of functions, but also they are hard to study mathematically in any generality. Therefore, in order to capture essential features of such problems, we focus on their subclasses that are well studied (spin glasses) and practically relevant (deep learning cost functions).

The halting time in such landscapes, when normalized to mean zero and variance one (subtracting the mean and dividing by the standard deviation), appears to follow a distribution that is independent of the input data, in other words it follows a universal distribution: the fluctuations are universal. In statistical mechanics, the term "universality" is used to refer to a class of systems which, on a certain macroscopic scale, behave statistically the same while having different statistics on a microscopic scale. An example of such a law is the central limit theorem, which states that the sums of observations tend to follow the same distribution independent of the distribution of the individual observations, as long as contribution from individual observations is reasonably small. It may fail to hold, if the microscopic behavior is not independent, does not have a finite second-moment, or if we consider something different than the sum. This work's focus is an attempt to put forward the cases where we *see* universality. But in this spirit, we show a degenerate case in which halting time fails to follow a universal law.

A rather surprising example of halting time universality is in the cases of observed human decision times and `Google`™ query times. In Bakhtin & Correll (2012) the time it takes a person make a decision in the presence of visual stimulus is shown to have universal fluctuations. The theoretically predicted curve in this experiment follows a Gumbel-like distribution. In addition, we randomly sampled words from two different dictionaries and submitted search queries. The time it takes Google to present the results are recorded. The normalized search times closely follow the same Gumbel-like curve.

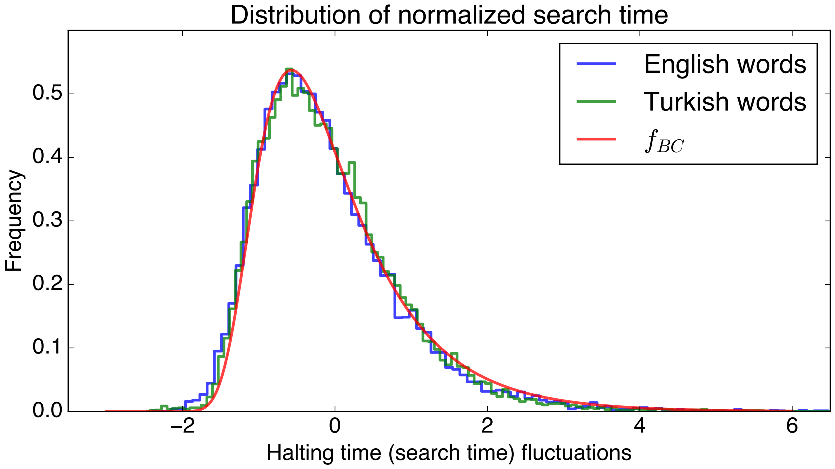


Figure 1: Search times of randomly selected words from two ensembles is compared with the curve f_{BC} in Bakhtin & Correll (2012) that is estimated from the decision times in an experiment conducted on humans. It is evident that more observations have yet to be made in identifying the underlying principles of the algorithms that are increasingly part of our life.

In the cases we observe, we find two main universality classes: (1) A Gumbel-like distribution that appears in Google searches, human decision times, QR factorization and spin glasses, and (2) a Gaussian-like distribution that appears in conjugate gradient algorithm and deep learning. To the best of our knowledge, our work along with the accompanying references in this introduction are the first ones to address the question of observing and classifying the distribution of the halting time.

1.1 Definition of universality

Definition 1.1. *An algorithm $\mathbb{A}$ consists of both a random cost function $F(\mathbf{x}, w)$ where $\mathbf{x}$ is a given random input and an optimization routine that seeks to minimize F with respect to w.*

To each algorithm we attach a precise ϵ-dependent halting criteria for the algorithm. The halting time, which is a random variable, is the time it takes to meet this criteria. Within each algorithm there must be an intrinsic notion of dimension which we denote by N. The halting time $T_{\epsilon,N,\mathbb{A},E}$ depends on ϵ, N, the choice of algorithm $\mathbb{A}$, and the ensemble E (or probability distribution). We use the empirical distribution of $T_{\epsilon,N,\mathbb{A},E}$ to provide heuristics for understanding the qualitative performance of the algorithms.

The presence of universality in an algorithm is the observation that for sufficiently large N and $\epsilon = \epsilon(N)$, the halting time random variable satisfies

$$\tau_{\epsilon,N,\mathbb{A},E} := \frac{T_{\epsilon,N,\mathbb{A},E} - \mathbb{E}[T_{\epsilon,N,\mathbb{A},E}]}{\sqrt{\mathrm{Var}(T_{\epsilon,N,\mathbb{A},E})}} \approx \tau^*_{\mathbb{A}}, \tag{1}$$

where $\tau^*_{\mathbb{A}}$ is a continuous random variable that depends only on the algorithm. The random variable $\tau_{\epsilon,N,\mathbb{A},E}$ is referred to as the *fluctuations* and when such an approximation appears to be valid we say that N and ϵ (and any other external parameters) are in the *scaling region*. For example, in Section 1.2, $\mathbb{A}$ is the QR eigenvalue algorithm, N is the size of the matrix, ϵ is a small tolerance and E is given by a distribution on complex Hermitian (or real symmetric) matrices.

Some remarks must be made:

- A statement like (1) is known to hold rigorously for a few algorithms (see Deift & Trogdon (2016; 2017)) but in practice, it is verified experimentally. This was first done in Pfrang et al. (2014) and expanded in Deift et al. (2014) for a total of 8 different algorithms.
- The random variable $\tau_{\mathbb{A}}^*$ depends fundamentally on the functional form of F. And we only expect (1) to hold for a restricted class of ensembles E.
- $T_{\epsilon,N,\mathbb{A},E}$ is an integer-valued random variable. For it to become a continuous distribution limit must be taken. This is the only reason N must be large — in practice, the approximation in (1) is seen even for small to moderate N.

Universality in this sense is a measure of stability in an algorithm. For example, it is known from the work of Kostlan (1988) that halting time for the power method to compute the largest eigenvalue (in modulus) of symmetric Gaussian matrices has infinite expectation and hence this type of universality is *not* believed to be present. One could use this to conclude that the power method is a naïve method for these matrices. Yet, it is known that the power method is much more efficient on positive-definite matrices where universality can be shown Deift & Trogdon (2017). Therefore, we have evidence that the presence of universality is a desirable feature of a numerical method.

1.2 Example: Demonstration of universality in the QR algorithm

To give some context, we discuss the universality in the solution of the eigenvalue problem with the classical QR algorithm. Historically, this was first noticed in Pfrang et al. (2014). In this example the fundamental object is the QR factorization $(Q, R) = \mathrm{QR}(A)$ where $A = QR$, Q is orthogonal (or unitary) and R is upper-triangular with positive diagonal entries. The QR algorithm applied to a Hermitian $N \times N$ matrix A is given by the iteration

$$\begin{aligned} A_0 &:= A, \\ (Q_j, R_j) &:= \mathrm{QR}(A_j), \\ A_{j+1} &:= R_j Q_j. \end{aligned}$$

Generically, $A_j \to D$ as $j \to \infty$ where D is a diagonal matrix whose diagonal entries are the eigenvalues of A. The halting time in Pfrang et al. (2014) was set to be the time of first deflation, $T_{\epsilon,N,\mathbb{A},E}(A)$, as:

$$\begin{aligned} \min\{j :& \sqrt{N(N-k)} \|A_j(k+1:N, 1:k)\|_\infty < \epsilon \\ & \text{for some } 1 \leq k \leq N-1\}. \end{aligned}$$

Here $\|A\|_\infty$ refers to the maximum entry of a matrix A in absolute value and the notation $A(i : j, k : l)$ refers to the submatrix of A consisting of entries only in rows $i, i+1, \ldots, j$ and in columns $k, k+1, \ldots, l$. Thus the halting time for the QR algorithm is the time at which at least one off-diagonal block is appropriately small. Next, we have to discuss choices for the randomness, or ensemble E, by choosing different distributions on the entries of A. Four such choices for ensembles are, Bernoulli ensemble (BE), Gaussian orthogonal ensemble (GOE), Gaussian unitary ensemble (GUE), Quartic unitary ensemble (QUE):

BE A is real-symmetric with iid Bernoulli ± 1 entries on and below the diagonal.

GOE A is real-symmetric with iid standard normal entries below the diagonal. The entries on the diagonal are iid normal with mean zero and variance two.

GUE A is complex-Hermitian with iid standard complex normal entries below the diagonal. The entries on the diagonal are iid complex normal mean zero and with variance two.

QUE A is complex-Hermitian with probability density $\propto e^{-\mathrm{tr} A^4} dA$. See Deift (2000) for details on such an ensemble and Olver et al. (2015) for a method to sample such a matrix. Importantly, the entries of the matrix below the diagonal are correlated.

Here we have continuous and discrete, real and complex, and independent and dependent ensembles but nevertheless we see universality in Figure 2 where we take $N = 150$ and $\epsilon = 10^{-10}$.

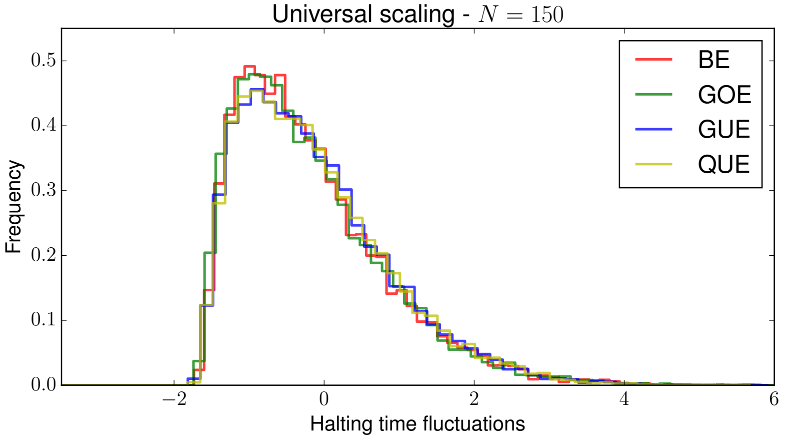


Figure 2: Empirical histograms for the halting time fluctuations $\tau_{\epsilon,N,\mathrm{QR},E}$ when $N = 150$, $\epsilon = 10^{-10}$ for various choices of ensembles E. This figure shows four normalized histograms, one each for $E =$ BE, GOE, GUE and QUE. It is clear that the fluctuations follow a universal law.

Remark 1.1. *The ensembles discussed above (GOE, GUE, BE and QUE) exhibit eigenvalue repulsion. That is, the probability that two eigenvalues are close*[1] *is much smaller than if the locations of the eigenvalues were just given by iid points on the line. It turns out that choosing a random matrix with iid eigenvalues breaks the universality that is observed in Figure 2. See Pfrang et al. (2014) for a more in-depth discussion of this.*

Remark 1.2. *To put the QR algorithm in the framework, let $B = UAU^*$ define $F(A, U)$ by*

$$\min\{j : \sqrt{N(N-k)}\|B(k+1:N, 1:k)\|_\infty < \epsilon \\ \textit{for some } 1 \leq k \leq N-1\}$$

We then use the QR algorithm to minimize F with respect to unitary matrices U using the initial condition $U = I$. If A is random then $F(A, U)$ represents a random field on the unitary group.

1.3 Core examples: Spin glass Hamiltonians and deep learning cost functions

A natural class of random fields is the class of Gaussian random functions on a high-dimensional sphere, known as p-spin spherical spin glass models in the physics literature (in the Gaussian process literature they are known as isotropic models). From the point of view of optimization, minimizing the spin glass model's Hamiltonian is fruitful because a lot is known about its critical points. This allows us to experiment with questions regarding whether the local minima and saddle points, due to the non-convex nature of landscapes, present an obstacle in the training of a system. Such observations on the Hamiltonian doesn't imply that it is a cost function or a simplified version of a cost function. Rather, the features that both systems have in common hint at a deeper underlying structure that needs to be discovered.

In recent years Dauphin et al. (2014) attacked the saddle point problem of non-convex optimization within deep learning. In contrast, Sagun et al. (2014) and the experimental second section of Choromanska et al. (2014) jointly argue that if the system is large enough, presence of saddle points is not an obstacle, and add that the local minimum practically gives a good enough solution within the limits of the model. However, Sagun et al. (2014) and Choromanska et al. (2014) hold different perspectives on what the qualitative similarities between optimization in spin glasses and deep learning might imply. The latter asserts a direct connection between the two systems based on these similarities. On the contrary, the former argues that these similarities hint at universal behaviors that are generically observed in vastly different systems rather than emphasizing a direct connection.

[1] By close, we mean that their distance is much less than $\mathcal{O}(1/N)$ where N is the size of the matrix.

In line with the asymptotic proof in Auffinger et al. (2013), the local minima are observed to lie roughly at the same energy level in spherical spin glasses. Auffinger et al. (2013) also gives asymptotic bounds on the value of the ground state and the exponential behavior of the average of the number of critical points below a given energy level. It turns out, when the dimension is large, the bulk of the local minima tend to have the same energy which is slightly above the global minimum. This level is called the *floor* level of the function. Simulations of the floor in spin glass can be found in Sagun et al. (2014). Sagun et al. (2014) also exhibits floor in a specially designed MNIST experiment: A student network is trained by the outputs of a pre-trained teacher network. Zero cost is achievable by the student, but the stochastic gradient descent cannot find zeros. It also does not have to because the floor level already gives a decent performance.

- Given data (i.e., from MNIST) and a measure $L(x^\ell, w)$ for determining the cost that is parametrized by $w \in \mathbb{R}^N$, the training procedure aims to find a point w^* that minimizes the empirical training cost while keeping the test cost low. Here $x^\ell \in Z$ for $\ell \in \{1, ..., S\}$, where Z is a random (ordered) sample of size S from the training examples. Total training cost is given by

$$F(Z, w) = \mathcal{L}_{\text{Train}}(w) = \frac{1}{S}\sum_{\ell=1}^{S} L(x^\ell, w). \tag{2}$$

- Given couplings $x_{(\cdot)} \sim$ Gaussian$(0, 1)$ that represent the strength of forces between triplets of spins. The state of the system is represented by $w \in S^{N-1}(\sqrt{N}) \subset \mathbb{R}^N$. The Hamiltonian (or energy) of the simplest complex[2] spherical spin glass model is given by:

$$F(x_{(\cdot)}, w) = H_N(w) = \frac{1}{N}\sum_{i,j,k}^{N} x_{ijk} w_i w_j w_k. \tag{3}$$

The two functions are indeed different in two major ways. First, the domain of the Hamiltonian is a compact space and the couplings are independent Gaussian random variables whereas the inputs for (2) are not independent and the cost function has a non-compact domain. Second, at a fixed point w, variance of the function $\mathcal{L}_{\text{Train}}(w)$ is inversely proportional to the number of samples, but the variance of $H_N(w)$ is N. As a result a randomly initialized Hamiltonian can take vastly different values, but a randomly initialized cost tend to have very similar values. The Hamiltonian has macroscopic extensive quantities: its minimum scales with a negative constant multiple of N. In contrast, the minimum of the cost function is bounded from below by zero. All of this indicates that landscapes with different geometries (glass-like, funnel-like, or another geometry) might still lead to similar phenomena such as existence of the floor level, and the universal behavior of the halting time.

1.4 Summary of results

We discuss the presence of universality in algorithms that are of a very different character. The conjugate gradient algorithm, discussed in Section 2.1, effectively solves a convex optimization problem. Gradient descent applied in the spin glass setting (discussed in Section 2.2) and stochastic gradient descent in the context of deep learning (MNIST, discussed in Section 2.3) are much more complicated non-convex optimization processes. Despite the fact that these algorithms share very little geometry in common, we demonstrate three things they share:

- A scaling region in which universality appears and performance is good.
- Regions where the computation is either ineffective or inefficient.
- A moment-based indicator for finding the universality class.

[2] 2-spin spherical spin glass, sum of $x_{ij} w_i w_j$ terms, has exactly $2N$ critical points. When $p \geq 3$, $p-$spin model has exponentially many critical points with respect to N. For the latter case, complexity is a measure on the number of critical points in an exponential scale. Deep learning problems are suspected to be complex in this sense.

2 Empirical observation of universality

2.1 The conjugate gradient algorithm

The conjugate gradient algorithm (Hestenes & Stiefel, 1952) for solving the $N \times N$ linear system $Ax = b$, when $A = A^*$ is positive definite, is an iterative procedure to find the minimum of the convex quadratic form:

$$F(A, y) = \frac{1}{2} y^* A y - y^* b,$$

where * denotes the conjugate-transpose operation. Given an initial guess x_0 (we use $x_0 = b$), compute $r_0 = b - Ax_0$ and set $p_0 = r_0$. For $k = 1, \ldots, N$,

1. Compute $r_k = r_{k-1} - a_{k-1} A p_{k-1}$ where[3] $a_{k-1} = \langle r_{k-1}, r_{k-1} \rangle / \langle p_{k-1}, A p_{k-1} \rangle$.
2. Compute $p_k = r_k + b_{k-1} p_{k-1}$ where $b_{k-1} = \langle r_k, r_k \rangle / \langle r_{k-1}, r_{k-1} \rangle$.
3. Compute $x_k = x_{k-1} + a_{k-1} p_{k-1}$.

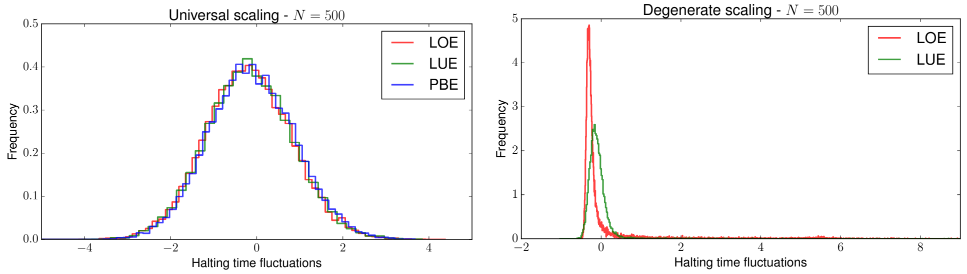
If A is strictly positive definite $x_k \to x = A^{-1} b$ as $k \to \infty$. Geometrically, the iterates x_k are the best approximations of x over larger and larger affine Krylov subspaces $\mathcal{K}_k$,

$$\begin{aligned} \|Ax_k - b\|_A &= \min_{x \in \mathcal{K}_k} \|Ax - b\|_A, \\ \mathcal{K}_k &= x_0 + \operatorname{span}\{r_0, Ar_0, \ldots, A^{k-1} r_0\}, \\ \|x\|_A^2 &= \langle x, A^{-1} x \rangle, \end{aligned}$$

as $k \uparrow N$. The quantity one monitors over the course of the conjugate gradient algorithm is the norm $\|r_k\|$:

$$T_{\epsilon, N, \mathrm{CG}, E}(A, b) := \min\{k : \|r_k\| < \epsilon\}.$$

In exact arithmetic, the method takes at most N steps: In calculations with finite-precision arithmetic the number of steps can be much larger than N and the behavior of the algorithm in finite-precision arithmetic has been the focus of much research (Greenbaum, 1989; Greenbaum & Strakos, 1992). What is important for us here is that it may happen that $\|r_k\| < \epsilon$ but the true residual $\hat{r}_k := b - Ax_k$ (which typically differs from r_k in finite-precision computations) satisfies $\|\hat{r}_k\| > \epsilon$.



(a)

(b)

Figure 3: Empirical histograms for the halting time fluctuations $\tau_{\epsilon, N, \mathrm{CG}, E}$ when $N = 500$, $\epsilon = 10^{-10}$ for various choices of ensembles E. (a) The scaling $M = N + 2\lfloor \sqrt{N} \rfloor$ demonstrating the presence of universality. This plot shows three histograms, one each for $E =$ LUE, LOE and PBE. (b) The scaling $M = N$ showing two histograms for $E =$ LUE and LOE and demonstrating the non-existence of universality.

Now, we discuss our choices for ensembles E of random data. In all computations, we take $b = (b_j)_{1 \leq j \leq N}$ where each b_j is iid uniform on $(-1, 1)$. We construct positive definite matrices A by $A = XX^*$ where $X = (X_{ij})_{1 \leq i \leq N,\, 1 \leq j \leq M}$ and each $X_{ij} \sim \mathcal{D}$ is iid for some distribution $\mathcal{D}$. We make the following three choices for $\mathcal{D}$, Positive definite Bernoulli ensemble (PDE), Laguerre orthogonal ensemble (LOE), Laguerre unitary ensemble (LUE):

[3] We use the notation $\|y\|^2 = \langle y, y \rangle = \sum_j |y_j|^2$ for $y = (y_1, y_2, \ldots, y_N) \in \mathbb{C}^N$.

- PBE $\mathcal{D}$ a Bernoulli ±1 random variable (equal probability).
- LOE $\mathcal{D}$ is a standard normal random variable.
- LUE $\mathcal{D}$ is a standard complex normal random variable.

The choice of the integer M, which is the inner dimension of the matrices in the product XX^*, is critical for the existence of universality. In Deift et al. (2014) and Deift et al. (2015) it is demonstrated that universality is present when $M = N + \lfloor c\sqrt{N} \rfloor$ and the ϵ-accuracy is small, but fixed. Universality is not present when $M = N$ and this can be explained by examining the distribution of the condition number of the matrix A in the LUE setting (Deift et al., 2015). We demonstrate this again in Figure 3(a). We also demonstrate that universality does indeed fail for $M = N$ in Figure 3(b).

2.2 SPIN GLASSES AND GRADIENT DESCENT

The gradient descent algorithm for the Hamiltonian of the p-spin spherical glass will find a local minimum of the non-convex function (3). Since variance of $H_N(w)$ is typically of order N, a local minimum has size N. More precisely, by Auffinger et al. (2013), the energy of the floor level where most of local minima are located is asymptotically at $-2\sqrt{2/3}N \approx -1.633N$ and the ground state is around $-1.657N$. The algorithm starts by picking a random element w of the sphere with radius $\sqrt{N}$, $S^{N-1}(\sqrt{N})$, as a starting point for each trial. We vary the environment for each trial and introduce ensembles by setting $x_{(\cdot)} \sim \mathcal{D}$ for a number of choices of distributions. For a fixed dimension N, accuracy ϵ that bounds the norm of the gradient, and an ensemble E: (1) Calculate the gradient steps: $w^{t+1} = w^t - \eta_t \nabla_w H(w^t)$, (2) Normalize the resulting vector to the sphere: $\sqrt{N}\frac{w^{t+1}}{||w^{t+1}||} \leftarrow w^{t+1}$, and (3) Stop when the norm of the gradient size is below ϵ and record $T_{\epsilon,N,\text{GD},E}$. This procedure is repeated 10,000 times for different ensembles (i.e. different choices for $\mathcal{D}$). Figure 4 exhibit the universal halting time presenting evidence that $\tau_{\epsilon,N,\text{GD},E}$ is independent of the ensemble.

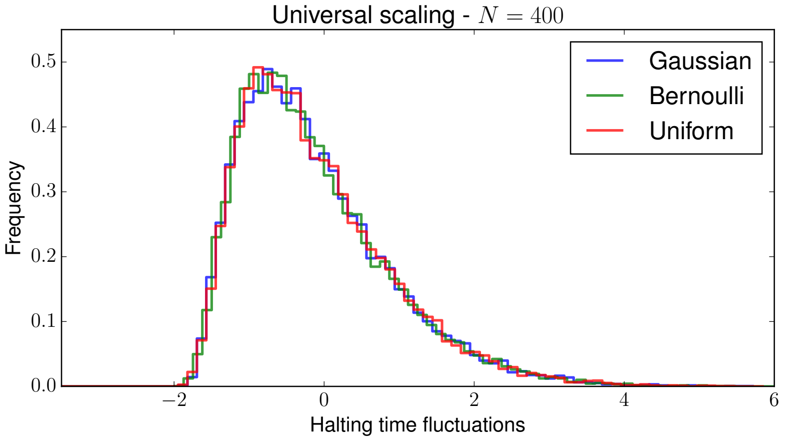


Figure 4: Universality across different distributions: We choose $\mathcal{D} \sim$ Gaussian$(0, 1)$, $\mathcal{D} \sim$ uniform on $(-(3/2)^{1/3}, (3/2)^{1/3})$ and $\mathcal{D} \sim$ Bernoulli $\pm 1/\sqrt{2}$ with equal probability.

2.3 DIGIT INPUTS VS. RANDOM INPUTS IN DEEP LEARNING

A deep learning cost function is trained on two drastically different ensembles. The first is the MNIST dataset, which consists of 60,000 samples of training examples and 10,000 samples of test examples. The model is a fully connected network with two hidden layers, that have 500 and 300 units respectively. Each hidden unit has rectified linear activation, and a cross entropy cost is attached at the end. To randomize the input data we sample 30K samples from the training set each time we set up the model and initialize the weights randomly. Then we train the model by the stochastic gradient descent method with a minibatch size of 100. This model gets us about 97%

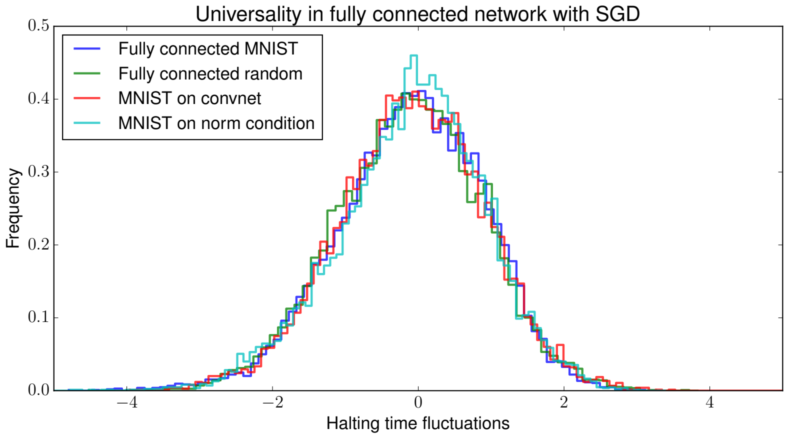


Figure 5: Universality in the halting time for deep learning cost functions. MNIST digit inputs and independent Gaussian noise inputs give rise to the same halting time fluctuations, as well as a convnet with a different stopping condition.

accuracy without any further tuning. The second ensemble uses the same model and outputs, but the input data is changed from characters to independent Gaussian noise. This model, as expected, gets us only about 10% accuracy: it randomly picks a number! The stopping condition is reached when the average of successive differences in cost values goes below a prescribed value. As a comparison we have also added a deep convolutional network (convnet), and we used the fully connected model with a different stopping condition: one that is tied to the norm of the gradient. Figure 5 demonstrates universal fluctuations in the halting time in all of the four cases.

3 Conclusions

What are the conditions on the ensembles and the model that lead to such universality? What constitutes a good set of hyperparameters for a given algorithm? How can we go beyond inspection when tuning a system? How can we infer if an algorithm is a good match to the system at hand? What is the connection between the universal regime and the structure of the landscape? This research attempts to exhibit cases where one can extract answers to these questions in a robust and quantitative way. The examples we have presented clearly exhibit universality. The normalized moment analysis, presented in the Appendix, gives a quantitative way to test for universality. And we further believe that an algorithm that exhibits universality is running in a scaling region of “high performance”: universality is a measure of insensitivity to initial data which is a beneficial property of a numerical method. Establishing this claim is a difficult task, beyond the scope of this primarily empirical work.

More specifically, the current work gives empirical evidence that within an appropriate scaling region, the halting time can often be approximated as

$$T_{\epsilon,N,\mathbb{A},E} \approx \mu + \sigma \tau^*_{\mathbb{A}},$$

where $\tau^*_{\mathbb{A}}$ is a mean-zero, variance one universal distribution. If this holds, a simple estimate of the mean $\mu = \mu_{\epsilon,N,\mathbb{A},E}$ and the standard deviation $\sigma = \sigma_{\epsilon,N,\mathbb{A},E}$ using a few samples will give a good *a priori* estimate of algorithm run time

$$\mathbb{P}\left(|T_{\epsilon,N,\mathbb{A},E} - \mu| \geq \sigma\ell\right) \approx \mathbb{P}\left(|\tau^*_{\mathbb{A}}| \geq \ell\right).$$

If $\tau^*_{\mathbb{A}}$ has (or is just conjectured to have) exponential tails, for example, this can be quite useful.

This work also validates the broad claims made in Deift et al. (2015) that universality is present in all or nearly all (sensible) computation. Future work will be along the lines of using these heuristics to identify when we have universality, to identify the different kinds of landscapes, and to guide both algorithm development and algorithm tuning. Furthermore, one would like theoretical estimates for the mean $\mu_{\epsilon,N,\mathbb{A},E}$ and the standard deviation $\sigma_{\epsilon,N,\mathbb{A},E}$.

Acknowledgments

We thank Percy Deift for valuable discussions and Gérard Ben Arous for his mentorship throughout the process of this research. The first author thanks very much to Uğur Güney for his availability for support and valuable contributions in countless implementation issues. This work was partially supported by the National Science Foundation under grant number DMS-1303018 (TT).

References

Robert J Adler and Jonathan E Taylor. *Random fields and geometry*. Springer Science & Business Media, 2009.

Antonio Auffinger, Gérard Ben Arous, and Jiří Černý. Random matrices and complexity of spin glasses. *Communications on Pure and Applied Mathematics*, 66(2):165–201, 2013.

Yuri Bakhtin and Joshua Correll. A neural computation model for decision-making times. *Journal of Mathematical Psychology*, 56(5):333–340, 2012.

Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surface of multilayer networks. *arXiv preprint arXiv:1412.0233*, 2014.

Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in Neural Information Processing Systems*, pp. 2933–2941, 2014.

Percy Deift. *Orthogonal polynomials and random matrices: a Riemann-Hilbert approach*, volume 3. American Mathematical Soc., 2000.

Percy Deift and Thomas Trogdon. Universality for the Toda algorithm to compute the eigenvalues of a random matrix. *arXiv Prepr. arXiv1604.07384*, apr 2016. URL http://arxiv.org/abs/1604.07384.

Percy Deift and Thomas Trogdon. Universality for eigenvalue algorithms on sample covariance matrices. *arXiv Preprint arXiv:1701.01896*, pp. 1–31, 2017.

Percy Deift, Govind Menon, Sheehan Olver, and Thomas Trogdon. Universality in numerical computations with random data. *Proceedings of the National Academy of Sciences*, 111(42):14973–14978, 2014.

Percy Deift, Govind Menon, and Thomas Trogdon. On the condition number of the critically-scaled laguerre unitary ensemble. *arXiv preprint arXiv:1507.00750*, 2015.

Anne Greenbaum. Behavior of slightly perturbed lanczos and conjugate-gradient recurrences. *Linear Algebra and its Applications*, 113:7–63, 1989.

Anne Greenbaum and Zdenek Strakos. Predicting the behavior of finite precision lanczos and conjugate gradient computations. *SIAM Journal on Matrix Analysis and Applications*, 13(1):121–137, 1992.

Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. *arXiv preprint arXiv:1509.01240*, 2015.

Magnus Rudolph Hestenes and Eduard Stiefel. Method of Conjugate Gradients for solving Linear Systems. *J. Res. Nat. Bur. Stand.*, 20:409–436, 1952.

Eric Kostlan. Complexity theory of numerical linear algebra. *Journal of Computational and Applied Mathematics*, 22(2):219–230, 1988.

Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent converges to minimizers. *University of California, Berkeley*, 1050:16, 2016.

Sheehan Olver, N Raj Rao, and Thomas Trogdon. Sampling unitary ensembles. *Random Matrices: Theory and Applications*, 4(01):1550002, 2015.

Christian W Pfrang, Percy Deift, and Govind Menon. How long does it take to compute the eigenvalues of a random symmetric matrix? *Random matrix theory, Interact. Part. Syst. Integr. Syst. MSRI Publ.*, 65:411–442, 2014.

Levent Sagun, V Uğur Güney, Gérard Ben Arous, and Yann LeCun. Explorations on high dimensional landscapes. *arXiv preprint arXiv:1412.6615*, 2014.

APPENDIX

FURTHER MOTIVATION FOR THE STUDY

Asymptotic error bounds for loss functions have been useful in the study of convergence properties of various models under various algorithms, for instance, at the heart of Hardt et al. (2015) and Lee et al. (2016) lies a bound that depends largely on the number of steps. Such bounds, even when they are tight, hold asymptotically. The finite time behaviour may be less pessimistic and it may prove to be useful in many practical concerns. For example, assuming the assumptions for a possible universality we can estimate the `Google`$^{\text{TM}}$ search time with high probability, whereas an upper bound of an asymptotic nature could give results that are a lot more pessimistic.

EFFECTS OF VARYING ACCURACY IN OPTIMIZATION

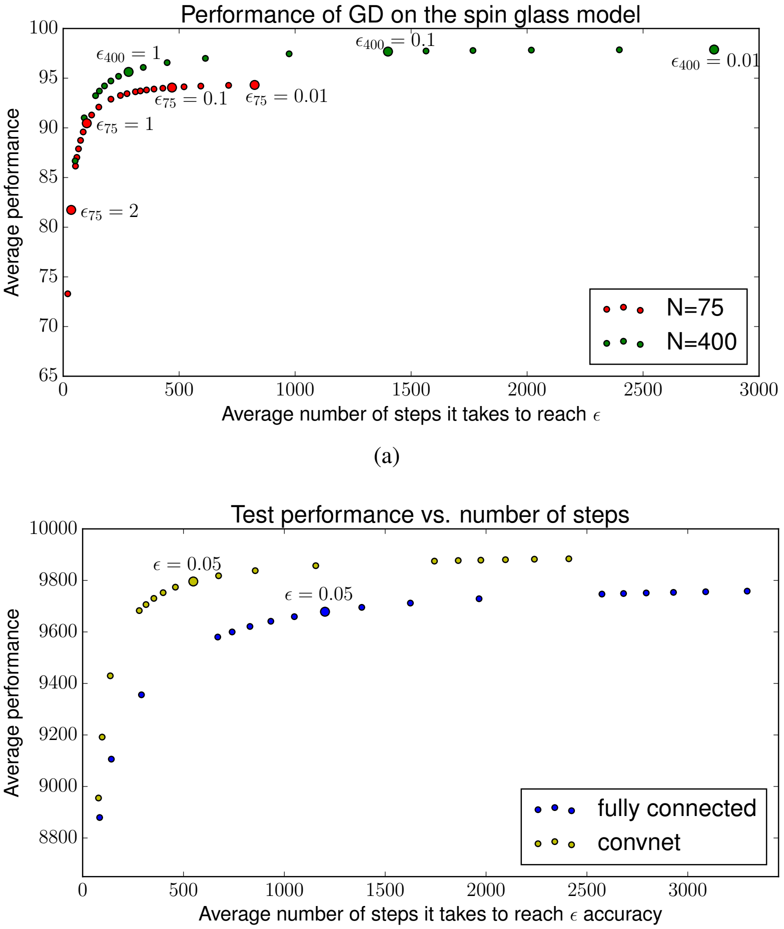
In Figure 6, we plot ensemble averages of efficiency versus accuracy for different ϵ's. A sharp plateau in the accuracy is seen, indicating that the extra computation for small values of ϵ is unnecessary. In MNIST and the spin glass example, the extra computation does not come with a gain in performance.

In the spin glass setting, the floor value gives a natural bound on the value that the Hamiltonian can practically reach. That value is above the ground state at an energy level where most local minima lie. This level presents a natural barrier for an algorithm like the gradient descent. Therefore a natural measure of performance at the point w^* is $H(w^*)/(\text{floor value})$. In MNIST, performance is the percentage of correct guesses in the test.

NORMALIZED-MOMENT ANALYSIS

We use the normalized third and fourth moments of the data, also referred to as the skewness and kurtosis, to identify which class the distributions belong to. Note that the first and second moments are zero and one since the date is normalized.

Intuitively, in gradient based methods, the halting time is effected by the curvature of the surface. And the curvature of the surface describes the landscape along the path of decay. The Gaussian-like behavior of halting time in MNIST might allow us to speculate that it has a funnel like non-convex landscape rather than a glassy landscape. This observation is consistent with Sagun et al. (2014) in its landscape exploration for spin glasses and deep learning.



(a)

(b)

Figure 6: (a) Norm of the gradient varies from 5 to 0.01 for the spin glass. (b) Averages of consecutive costs on MNIST that varies from 0.6 to 0.005.

Model	Ensemble	Mean	St.dev.	3rd	4th
CG: $M = N$	LOE	970	164	5.1	35.2
CG: $M = N$	LUE	921	46	15.7	288.5
CG: $M = N + 2\lfloor\sqrt{N}\rfloor$	LOE	366	13	0.08	3.1
CG: $M = N + 2\lfloor\sqrt{N}\rfloor$	LUE	367	9	0.07	3.0
CG: $M = N + 2\lfloor\sqrt{N}\rfloor$	PBE	365	13	0.08	3.0
Spin Glass	Gaussian	192	79.7	1.10	4.58
Spin Glass	Bernoulli	192	80.2	1.10	4.56
Spin Glass	Uniform	193	79.6	1.10	4.54
QR	BE	26	15	1.18	4.77
QR	GOE	24	14	1.17	4.78
QR	GUE	22	12	1.04	4.32
QR	QUE	22	12	1.02	4.16
Fully connected	MNIST	2929	106	-0.32	3.24
Fully connected	Random	4223	53	-0.08	2.98
Convnet	MNIST	2096	166	-0.11	3.18
Cond. on gradient	MNIST	3371	118	-0.34	3.31

Table 1: Skewness (3rd moment) and kurtosis (4th moment) for the experiments: (1) In the $M = N + 2\lfloor\sqrt{N}\rfloor$ it is clear that these normalized moments nearly coincide and they are quite distinct for $M = N$. (2) Gumbel like distribution in spin glasses and QR. (3) Gaussian-like distribution, with a flat left tail for deep learning.