QUARTERLY OF APPLIED MATHEMATICS
VOLUME LXXVI, NUMBER 2
JUNE 2018, PAGES 289–301
http://dx.doi.org/10.1090/qam/1483
Article electronically published on September 20, 2017

UNIVERSAL HALTING TIMES IN OPTIMIZATION AND MACHINE LEARNING

By

LEVENT SAGUN (*Department of Mathematics, New York University, New York, New York 10012*),

THOMAS TROGDON (*Department of Mathematics, University of California, Irvine, California 92697*),

AND

YANN LECUN (*Department of Computer Science, New York University, New York, New York 10012*)

Abstract. We present empirical evidence that the halting times for a class of optimization algorithms are universal. The algorithms we consider come from quadratic optimization, spin glasses and machine learning. A universality theorem is given in the case of the quadratic gradient descent flow. More precisely, 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 of the algorithm follow a distribution that, after centering and scaling, appears invariant under changes in the distribution on the landscape — universality is present.

1. Introduction. In this paper we discuss the presence of universality in optimization algorithms. More precisely, we analyze the required number of iterations of a given algorithm to optimize (or approximately optimize) an energy functional when the functional is random. We consider the following iterative routines: gradient descent and the conjugate gradient algorithm for solving a linear system [12], gradient descent for spin glasses, and stochastic gradient descent for deep learning [5].

A bounded, piecewise differentiable random field[1] where the randomness is nondegenerate, may yield a landscape with many saddle points and local minima. We refer to the value of the landscape at a given point as the *energy*. Consider a moving particle on the landscape that takes a sequence of steps, attempting to reach a (local) minimum. An essential quantity is the time (or number of steps) the particle requires to find this

Received June 16, 2017.
2010 *Mathematics Subject Classification.* Primary 68Q32, 78M50.
E-mail address: sagun@cims.nyu.edu
E-mail address: ttrogdon@math.uci.edu
E-mail address: yann@cs.nyu.edu

[1]See [1] for an account on the connection of random fields and geometry.

©2017 Brown University

minimum. We call this the *halting time*. Many useful bounds on the halting time are known for convex cases where the global minimum is necessarily found. In non-convex cases, however, the particle knows only information that can be calculated locally. And a locally computable stopping condition, such as the norm of the gradient at the present point, or the difference in energy between successive steps, 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, even though the global minimum may not be located.

A prototypical example of a random field is found in the class of polynomials with random coefficients. Spin glasses and deep learning cost functions are special cases of such fields that yield very different landscapes. We emphasize that polynomials with random coefficients are a broad class of functions but they are hard to study in any generality. Therefore, in order to capture essential features of such problems, we focus on subclasses of random polynomials 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 random input data — 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 (input) statistics on a microscopic scale. An example is the central limit theorem (CLT) which states that the sums of observations tend to follow the same distribution, independent of the distribution of the individual observations. This holds provided the contribution from each individual observation is reasonably small. The CLT may fail to hold if the microscopic statistics are not independent, do not have a finite second-moment or if we move beyond summation.

1.1. *Results.* We first present a universality theorem for the quadratic gradient descent flow. This shows that universality within optimization processes is a *bona fide* phenomenon. We use numerical experiments to show that universality persists when the flow is discretized. The focus then turns to attempts to put forward cases where we *see* universality (in addition to [6, 8, 9, 14]) with an emphasis on routines and landscapes from (or related to) machine learning. We present concrete evidence that the halting time in such optimization problems is universal (Sections 3.2 and 3.3). But, in the spirit of the potential failure of the CLT in degenerate cases, we show a degenerate case for the conjugate gradient algorithm (Section 3.1) where the halting time fails to follow a universal law.

Another example of halting time universality is in the work of Bakhtin and Correll [3]. In this experimental work, the time it takes a person to make a decision in the presence of a visual stimulus is shown to have universal fluctuations. The theoretically predicted distribution f_{BC} for this experiment is a Gumbel distribution. In a surprising connection, we sampled words uniformly at random from two different dictionaries (English and Turkish) and submitted search queries to the `Google`$^{\mathrm{TM}}$ search engine. The time it took `Google`$^{\mathrm{TM}}$ to present the results was recorded. After normalizing the times to mean zero

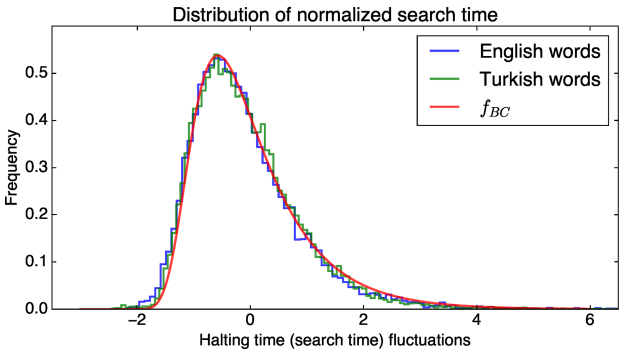


FIG. 1. Normalized Google^{TM} search times for randomly selected words from two ensembles is compared with the curve f_{BC} in [3] that is estimated from the decision times in an experiment conducted on humans. This plot indicates that the search time is universal for searches performed randomly in different languages.

and variance one, the normalized search times closely follow the same Gumbel curve; see Figure 1 — Google^{TM} search times are universal.

In the cases we observe below, we find two qualitative universality classes: (1) A non-symmetric Gumbel-like distribution that appears in Google^{TM} searches, human decision times and spin glasses, and (2) a symmetric Gaussian-like distribution that appears in the conjugate gradient algorithm and in deep learning.

1.2. *Definition of universality.*

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

To each algorithm we attach a precise ϵ-dependent halting criterion for the algorithm. The halting time, which is a random variable, is the time (i.e. the number of iterations) it takes to meet this criterion. 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) on $\mathbf{x}$. We use the empirical distribution of $T_{\epsilon,N,\mathbb{A},E}$ to provide heuristics for understanding the qualitative performance of the algorithms — we look for universality.

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 *fluctuation* 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. We note that universality in this sense can also be used as a measure of stability in an algorithm. Some remarks must be made:

- A statement like (1) is known to hold rigorously for a few algorithms (see [8, 9]) but in practice, it is verified experimentally. Theorem 2.1 presents another rigorous case. The experimental verification was first done in [14] and expanded in [6] 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 the limit $N \to \infty$ 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.

1.3. *Core empirical examples: Spin glass Hamiltonians and deep learning cost functions.* A natural class of random fields is that of Gaussian random functions on a high-dimensional sphere. These Hamiltonians are also known as p-spin spherical spin glass models in the physics literature.[2] From the point of view of optimization, minimizing the spin glass Hamiltonian is fruitful because much is known about its critical points. This allows one to experiment with questions regarding whether the local minima and saddle points present an obstacle in the training of a system without convexity.

Following the asymptotic proof in [2], the local minima of the spin glass Hamiltonian lie at roughly the same energy level. Moreover, the values of the ground states are known and it has been shown that there is an exponential growth in the average number of critical points below any given energy level. When the dimension (or the number of spins) is large, it turns out that the bulk of the local minima have the same energy — an energy level that is slightly above the global minimum. This level is called the *floor* level of the function. An optimization simulation for this model can only locate the values at the floor level, and not deeper. This same phenomenon is present in optimization for the MNIST[3] classification problem [16]. We emphasize that this striking similarity is only at the level of analogy, and the two systems are in fact vastly different. To the best of the authors' knowledge, there are no known theoretical arguments that connect spin glass Hamiltonians to deep learning. However, the feasibility of the observation of the floor level in the optimization of the two problems may give a hint at universal properties that can also be observed in other systems. In more detail, our two core random fields are:

- **Deep learning:** Given data (i.e., from MNIST) and a measure $L(x^\ell, w)$ for determining the cost, 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 then given by

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

[2]In the Gaussian process literature they are known as isotropic models.

[3]MNIST is a database of handwritten numerical digits that is commonly used as a means of benchmarking deep learning networks [13].

- **Spin glass Hamiltonians:** Let $x_{(\cdot)} \sim$ Gaussian$(0,1)$ be couplings that represent the strength of forces between triplets of spins and let the state (spins) of the system be represented by $w \in S^{N-1}(\sqrt{N}) \subset \mathbb{R}^N$. The Hamiltonian (or energy) of the simplest complex[4] spherical spin glass model is then 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 (2) and (3) are indeed different in two major ways. First, the domain of the Hamiltonian (3) is a compact space and the couplings are independent Gaussian random variables whereas the inputs for the cost function (2) are not independent and the cost function has a non-compact domain. Second, at a fixed point w, the variance of the function $\mathcal{L}_{\text{Train}}(w)$ is inversely proportional to the number of samples while the variance of $H_N(w)$ is N. A randomly initialized Hamiltonian can take vastly different values, but randomly initialized costs 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.

2. Universality for gradient descent. We begin by considering a simpler linear algebra problem. In this case we can present a universality theorem. Consider the problem of solving $Ax = b$ where A is (strictly) positive definite. This is turned directly into a quadratic convex optimization problem by setting

$$F((A,b),w) = \mathcal{L}(w) = \frac{1}{2} w^* A w - \operatorname{Re} w^* b.$$

Here * denotes the conjugate-transpose operation. Given an initial condition $w(0) = w_0$, the flow

$$\dot{w}(t) = -\nabla \mathcal{L}(w(t)), \quad \nabla \mathcal{L}(w(t)) = Aw(t) - b,$$

will converge to $x = A^{-1}b$ as $t \to \infty$. We let A and b be random, and ask how long it takes to converge to x. This time will be universal. We choose A to be a sample covariance matrix.

Definition 2.1 (Sample covariance matrix (SCM)). A sample covariance matrix (ensemble) is a real symmetric ($\beta = 1$) or complex Hermitian ($\beta = 2$) matrix $H = V^*V/M$, $V = (V_{ij})_{1 \le i \le M, 1 \le j \le N}$ such that V_{ij} are independent random variables for $1 \le i \le M$, $1 \le j \le N$ given by a probability measure ν_{ij} with

$$\mathbb{E} V_{ij} = 0, \quad \mathbb{E}|V_{ij}|^2 = 1.$$

[4] 2-spin spherical spin glass, sum of $x_{ij} w_i w_j$ terms, has exactly $2N$ critical points. When $p \ge 3$, $p-$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.

Next, assume there is a fixed constant ν (independent of N, i, j) such that

$$\mathbb{P}(|V_{ij}| > x) \leq \nu^{-1} \exp(-x^{\nu}), \quad x > 1. \tag{4}$$

For $\beta = 2$ (when V_{ij} is complex-valued) the condition

$$\mathbb{E}V_{ij}^2 = 0,$$

must also be satisfied.

We choose b to be an independent (of A) random unit vector.

DEFINITION 2.2. A random unit vector is given by

$$v = Y/\|Y\|_2, \quad Y = (Y_1, Y_2, \ldots, Y_N)^T,$$

for the given (independent) random variables Y_j.

Assume $A = U\Lambda U^*$ where $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$ and $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N$. We halt the gradient descent flow when the norm of the gradient is small

$$T_{\epsilon,N,\mathrm{qGD},E} := \min\{t : \|\nabla \mathcal{L}(w(t))\|_2 \leq \epsilon\}.$$

An explicit expression for $\|\nabla \mathcal{L}(w(t))\|_2$ is available. Indeed, it follows that

$$w(t) = e^{-At}(w(0) - A^{-1}b) + A^{-1}b,$$

and therefore

$$\nabla \mathcal{L}(w(t)) = Aw(t) - b = e^{-At}(Aw(0) - b).$$

Using the spectral decomposition for A, assuming, for simplicity that $w(0) = 0$, we have

$$\|\nabla \mathcal{L}(w(t))\|_2^2 = E(t) := \sum_{n=1}^{N} e^{-2\lambda_n t} \lambda_n^2 q_n^2, \quad q = (q_n)_{1 \leq n \leq N} = U^* b.$$

We write

$$E(t) = q_1^2 e^{-2\lambda_1 t} \left(\lambda_1^2 + \sum_{n=2}^{N} \lambda_n^2 e^{-2(\lambda_n - \lambda_1)t} \frac{q_n^2}{q_1^2} \right).$$

We assume $M \sim N/d$ for $0 < d < 1$. The following lemmas are a consequence of [4,15] (see also [9]).

LEMMA 2.1. For $s > 0$ fixed, let M_C, $C > 1$, denote the set of matrices which satisfy the estimates

$$N^{-2/3-s} \leq |\lambda_1 - \lambda_2| \leq N^{-2/3+s}, \quad N^{-1/2-s/2} \leq q_1 \leq N^{-1/2+s/2},$$

and

$$C^{-1} \leq \lambda_j \leq C \quad \text{for all } j.$$

For SCMs, there exists $C > 1$ such that $\mathbb{P}(M_C) \to 1$ as $N \to \infty$ for every fixed $s > 0$.

LEMMA 2.2. There exists a distribution function $F_\beta(t)$ such that

$$\lim_{N \to \infty} \mathbb{P}(2^{-7/6} \lambda_-^{-2/3} N^{2/3} (\lambda_- - \lambda_1) \leq t) = F_\beta(t), \quad \lambda_- = (1 - d^{1/2})^2.$$

The distribution function $F_\beta(t)$ is called the Tracy–Widom distribution [17]. We preform calculations only for matrices in M_C for an appropriate choice of C. Since $E(t)$ is strictly monotonic, it follows that $T_{\epsilon,N,\mathrm{qGD},E}$ is the unique time at which $E(T_{\epsilon,N,\mathrm{qGD},E}) = \epsilon^2$. We construct an approximation to $T_{\epsilon,N,\mathrm{qGD},E}$ by setting

$$\lambda_1^2 q_1^2 e^{-2\lambda_1 \hat{T}} = \epsilon^2, \quad \hat{T} = \frac{\log \epsilon^{-1} - \log q_1 - \log \lambda_1}{\lambda_1}.$$

From the Mean-Value Theorem

$$|\hat{T} - T_{\epsilon,N,\mathrm{qGD},E}| = |E(\hat{T}) - E(T_{\epsilon,N,\mathrm{qGD},E})| \frac{1}{|E'(\zeta)|},$$

where ζ is between $\hat{T}$ and $T_{\epsilon,N,\mathrm{qGD},E}$. We assume $\epsilon = N^{-\sigma}$ for $\sigma > 1/2$. It also follows that $T_{\epsilon,N,\mathrm{qGD},E} \geq \hat{T}$ so that

$$\begin{aligned} |\hat{T} - T_{\epsilon,N,\mathrm{qGD},E}| &\leq \frac{|E(\hat{T}) - E(T_{\epsilon,N,\mathrm{qGD},E})|}{|E'(T_{\epsilon,N,\mathrm{qGD},E})|} \leq \frac{\epsilon^{-2}}{2\lambda_1} \sum_{n=2}^{N} e^{-2\lambda_n \hat{T}} \lambda_n^2 q_n^2 \\ &= \frac{\epsilon^{-2}}{2\lambda_1} \sum_{n=2}^{N} \left(\frac{\epsilon}{q_1}\right)^{2\frac{\lambda_n}{\lambda_1}} \lambda_n^2 q_n^2 \leq C^3 \frac{\epsilon^{-2}}{2} \left(\frac{\epsilon}{q_1 \lambda_1}\right)^{2\frac{\lambda_2}{\lambda_1}}. \end{aligned} \tag{5}$$

This last inequality follows because $\sum_n \lambda_n^2 q_n^2 \leq C^2$, $\lambda_1^{-1} \leq C$ and $\epsilon < \lambda_1 q_1$. This estimate is not without issue: $\lambda_2 - \lambda_1 \to 0$ as $N \to \infty$. But it will turn out to be sufficient that (5) is bounded by $C^3/2$. In other words, we require

$$\begin{aligned} \log \epsilon^{-2} + 2\frac{\lambda_2}{\lambda_1} \log \epsilon - 2\frac{\lambda_2}{\lambda_1} \log \lambda_1 q_1 &\leq 0 \\ \sigma \left(2 - 2\frac{\lambda_2}{\lambda_1}\right) &\geq 2\frac{\lambda_2 \log \lambda_1^{-1} q_1^{-1}}{\lambda_1 \log N}. \end{aligned}$$

By Lemma 2.1, for any $s > 0$, the probability is that

$$2\frac{\lambda_2 \log \lambda_1^{-1} q_1^{-1}}{\lambda_1 \log N} \leq 1 + s$$

tends to one as $N \to \infty$. Thus, it suffices to choose

$$\sigma \geq (1+s) \left(2 - 2\frac{\lambda_2}{\lambda_1}\right)^{-1}$$

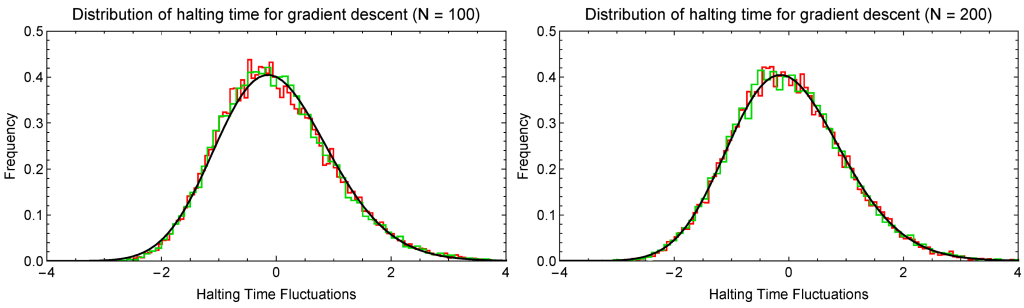
or, more simply, $\sigma \geq N^{2/3+\gamma}$ for any $\gamma > 0$. For such a choice of σ, on the set M_C we have

$$\frac{|\hat{T} - T_{\epsilon,N,\mathrm{qGD},E}|}{\log \epsilon^{-1}} \leq N^{-2/3-\gamma}.$$

Because $\mathbb{P}(M_C) \to 1$ as $N \to \infty$, it follows that

$$N^{2/3} \frac{|\hat{T} - T_{\epsilon,N,\mathrm{qGD},E}|}{\log \epsilon^{-1}} \to 0$$

in probability. It follows from Lemma 2.2 that $2^{-7/6} \lambda_-^{-2/3} N^{2/3} (\hat{T}/\log \epsilon^{-1} - \lambda_-)$ converges in distribution to the Tracy–Widom distribution, and therefore we have the following theorem.



(a) (b)

FIG. 2. A demonstration that Theorem 2.1 (approximately) holds outside the scaling region $\log \epsilon < -N^{2/3+\gamma} \log N$. The histograms in each panel are for the distributions given by $M = 2N$ and (V_{ij}) are iid standard normal or mean zero Bernoulli random variables (defined as PBE and LOE in the following section). (a) Universality across different distributions for gradient descent for $N = 100$, $\Delta t = 0.01$ and $\epsilon = 0.001$. Here $\tau_{\epsilon,N,\mathrm{qGD},A}$ is compared against the (mean zero, variance one) Tracy–Widom distribution. (b) Universality across different distributions for gradient descent for $N = 200$, $\Delta t = 0.01$ and $\epsilon = 0.0001$. Here $\tau_{\epsilon,N,\mathrm{qGD},A}$ is compared against the (mean zero, variance one) Tracy–Widom distribution.

THEOREM 2.1. For $\gamma > 0$, $\log \epsilon < -N^{2/3+\gamma} \log N$, if E is an SCM and $t \in \mathbb{R}$

$$\lim_{N\to\infty} \mathbb{P}\left(2^{-7/6}\lambda_-^{-2/3} N^{2/3}\left(\frac{T_{\epsilon,N,\mathrm{qGD},E}}{\log \epsilon^{-1}} - \frac{1}{\lambda_-}\right) \le t\right) = F_\beta(t).$$

The restriction on ϵ is unrealistic for what would be used in practice — it is exponential in N. It is our conjecture that this is necessary for a limit theorem — with F_β being the limit — to hold for gradient descent with this choice of ensembles. Nonetheless, numerical observations, which are presented in Figure 2, demonstrate that for an approximate integration of the flow $\dot{w}(t) = -\nabla \mathcal{L}(w(t))$, Theorem 2.1 still gives a good approximation of the normalized halting time, even when ϵ is not small. More precisely, for $\Delta t > 0$ define the discretization by Euler's method

$$w_n = w_{n-1} - \Delta t \nabla \mathcal{L}(w_{n-1}), \quad n \ge 0, \quad w_0 = 0.$$

The discrete halting time is then given by

$$\bar{T}_{\epsilon,N,\mathrm{dGD},E} := \min\{n\Delta t : \|\nabla \mathcal{L}(w_n)\|_2 \le \epsilon\}.$$

For $\Delta t = 0.01$, we plot histograms for the normalized halting time (fluctuations) $\tau_{\epsilon,N,\mathrm{dGD},E}$ in Figure 2. Particularly in Figure 2(b), the histograms significantly overlap with density $F_1'(t)$, after it is normalized to mean zero and variance one.

3. Empirical observation of universality. We discuss the presence of universality in algorithms that are of a very different character, building of the ideas in the previous section. The conjugate gradient algorithm, an improvement on quadratic gradient descent, discussed in Section 3.1, effectively solves a convex optimization problem. Gradient

descent applied in the spin glass setting (discussed in Section 3.2) and stochastic gradient descent in the context of deep learning (MNIST, discussed in Section 3.3) are much more complicated non-convex optimization processes. Despite the fact that these algorithms share very little geometry in common, they all (empirically) exhibit universality in their halting times (Table 1). Indeed, if the normalized third and fourth moments are close, across different ensembles, there is a strong indication that fluctuations are universal.

TABLE 1. Skewness (normalized third moment) and kurtosis (normalized fourth moment) for the halting times in the experiments performed below: (Rows 1-5, Section 3.1) In the $M = N + 2\lfloor\sqrt{N}\rfloor$ it is clear that these normalized moments nearly coincide and they are quite distinct for $M = N$. (Rows 6-8, Section 3.2) The Gumbel-like distribution in spin glasses. (Rows 9-12, Section 3.3) Gaussian-like distribution, with a flat left tail for deep learning.

MODEL	ENSEMBLE	MEAN	ST.DEV.	NORM. 3RD	NORM. 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
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

3.1. *The conjugate gradient algorithm.* The conjugate gradient algorithm [12] 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

$$F(A, w) = \frac{1}{2} w^* A w - \operatorname{Re} w^* b.$$

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 $a_{k-1} = \langle r_{k-1}, r_{k-1}\rangle / \langle p_{k-1}, Ap_{k-1}\rangle$.
(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 $x_k = x_{k-1} + a_{k-1} p_{k-1}$.

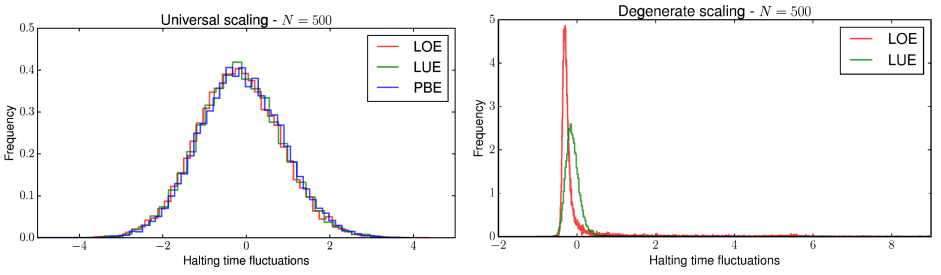
If A is strictly positive definite $x_k \to x = A^{-1} b$ as $k \to \infty$. Geometrically, the iterates x_k are the best approximations of x over larger and larger affine Krylov subspaces $\mathcal{K}_k$,

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

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 [10, 11].



(a) (b)

FIG. 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 $b = (b_j)_{1\le j\le N}$ where each b_j is iid uniform on $(-1, 1)$. We construct SCMs A by $A = VV^*$ where $V = (V_{ij})_{1\le i\le N,\ 1\le j\le M}$ and each $V_{ij} \sim \mathcal{D}$ is iid with distribution $\mathcal{D}$. We make the following three choices for $\mathcal{D}$:

PBE Positive definite Bernoulli ensemble: $\mathcal{D}$ a Bernoulli ± 1 random variable (equal probability).

LOE Laguerre orthogonal ensemble: $\mathcal{D}$ is a standard normal random variable.

LUE Laguerre unitary ensemble: $\mathcal{D}$ is a standard complex normal random variable.

The choice of the integer M, which is the inner dimension of the matrices in the product VV^*, is critical for the existence of universality. In [6] and [7] it is demonstrated that universality is present when $M = N + \lfloor c\sqrt{N}\rfloor$ and ϵ is small, but fixed. Universality is not present when $M = N$ and this can be explained heuristically by examining the distribution of the condition number of the matrix A in the LUE setting [7]. We demonstrate this again in Figure 3(a). We also demonstrate that universality does indeed fail[5] for $M = N$ in Figure 3(b). We refer to Table 1(Rows 1-5) for a quantitative indicator of universality.

3.2. *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 the $H_N(w)$ is typically of order N, a local minimum scale like $-N$. More precisely, from [2], 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 approximately

[5]For those familiar with random matrix theory, this might not be surprising as real and complex matrices typically lie in different universality classes. From this point of view, it is yet more striking that Figure 3(a) gives a universal curve.

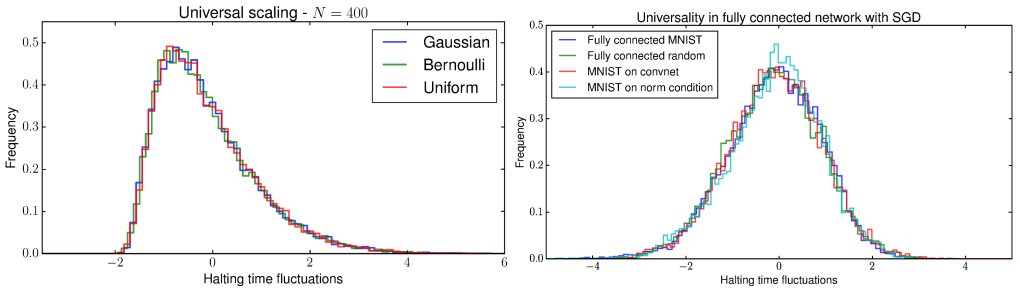


FIG. 4. (a) Spin glass landscape: Universality across different distributions: We choose $\mathcal{D} \sim$ Gaussian$(0,1)$, $\mathcal{D} \sim$ uniform on $(-(3/2)^{1/3}, (3/2)^{1/3})$ and $\mathcal{D} \sim$ Bernoulli $\pm 1/\sqrt{2}$ with equal probability. (b) Deep learning landscape: Universality 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. The same is true of a convnet with a different stopping condition.

$-1.657N$. The algorithm starts by picking a random element w of the sphere with radius $\sqrt{N}$ as a starting point for each trial. For a fixed dimension N, accuracy ϵ and ensemble E: (1) Calculate the gradient step: $w^{t+1} = w^t - \eta_t \nabla_w H(w^t)$, (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 $T_{\epsilon,N,\mathrm{GD},E}$. This procedure is repeated 10,000 times for different ensembles. We vary the environment for each trial and introduce ensembles by setting $x_{(\cdot)} \sim \mathcal{D}$ for a number of choices of distributions $\mathcal{D}$. Figure 4(b) exhibits the universal halting time which presents evidence that $\tau_{\epsilon,N,\mathrm{GD},E}$ is independent of the ensemble. We refer to Table 1(Rows 6-8) for a clear quantitative verification of universality.

3.3. *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 30,000 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 gives approximately 97% 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, gives about 10% accuracy: It randomly picks a number! For these two methods, the stopping condition is calculated through the running average of the stochastic costs. To be more precise, let $F(Z,w) = \mathcal{L}_{\mathrm{Train}}(w)$ be the cost function associated with a given model. The cost is formed by the average of the costs per example: $\mathcal{L}_{\mathrm{Train}}(w) = \frac{1}{S}\sum_{\ell=1}^{S} L(x^\ell, w)$ (see equation (2)). Then the stochastic gradients are the following $\mathcal{L}'_{\mathrm{Train}}(w) = \frac{1}{S'}\sum_{\ell=1}^{S'} L(x^\ell, w)$ where $S' \subset S$.

Then the stopping criterion is reached when the running average of the last 100 stochastic costs is below 10^{-2}.

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. In this one, we calculate the norm of the gradient on the cost itself, so the stopping criteria is reached when $||\nabla_w \mathcal{L}_{\text{Train}}(w)||_2 < \epsilon$. Figure 4(a) demonstrates universal fluctuations in the halting time in all four cases. Again, we refer to Table 1(Rows 9-12) for a quantitative verification of universality — despite the large amount of noise in the dataset the skewness and kurtosis remain very close across different ensembles.

Acknowledgments. We thank Percy Deift and Andrew Stuart for valuable discussions and Gérard Ben Arous for his mentorship throughout the process of this research. The first author thanks Uğur Güney very much for his availability, 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

[1] Robert J. Adler and Jonathan E. Taylor, *Random fields and geometry*, Springer Monographs in Mathematics, Springer, New York, 2007. MR2319516

[2] Antonio Auffinger, Gérard Ben Arous, and Jiří Černý, *Random matrices and complexity of spin glasses*, Comm. Pure Appl. Math. **66** (2013), no. 2, 165–201, DOI 10.1002/cpa.21422. MR2999295

[3] Yuri Bakhtin and Joshua Correll, *A neural computation model for decision-making times*, J. Math. Psych. **56** (2012), no. 5, 333–340, DOI 10.1016/j.jmp.2012.05.005. MR2983392

[4] Alex Bloemendal, Antti Knowles, Horng-Tzer Yau, and Jun Yin, *On the principal components of sample covariance matrices*, Probab. Theory Related Fields **164** (2016), no. 1-2, 459–552, DOI 10.1007/s00440-015-0616-x. MR3449395

[5] Léon Bottou, *Large-scale machine learning with stochastic gradient descent*, Proceedings of COMPSTAT'2010, Physica-Verlag/Springer, Heidelberg, 2010, pp. 177–186. MR3362066

[6] Percy A. Deift, Govind Menon, Sheehan Olver, and Thomas Trogdon, *Universality in numerical computations with random data*, Proc. Natl. Acad. Sci. USA **111** (2014), no. 42, 14973–14978, DOI 10.1073/pnas.1413446111. MR3276499

[7] Percy A. Deift, Thomas Trogdon, and Govind Menon, *On the condition number of the critically-scaled Laguerre unitary ensemble*, Discrete Contin. Dyn. Syst. **36** (2016), no. 8, 4287–4347, DOI 10.3934/dcds.2016.36.4287. MR3479516

[8] Percy Deift and Thomas Trogdon, *Universality for the Toda algorithm to compute the largest eigenvalue of a random matrix*, Commun. Pure Appl. Math. (2017), 1–27. doi:10.1002/cpa.21715.

[9] Percy Deift and Thomas Trogdon, *Universality for eigenvalue algorithms on sample covariance matrices*, SIAM J. Num. Anal. (2017), 1–31.

[10] A. Greenbaum, *Behavior of slightly perturbed Lanczos and conjugate-gradient recurrences*, Linear Algebra Appl. **113** (1989), 7–63, DOI 10.1016/0024-3795(89)90285-1. MR978581

[11] A. Greenbaum and Z. Strakoš, *Predicting the behavior of finite precision Lanczos and conjugate gradient computations*, SIAM J. Matrix Anal. Appl. **13** (1992), no. 1, 121–137, DOI 10.1137/0613011. MR1146656

[12] Magnus R. Hestenes and Eduard Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Research Nat. Bur. Standards **49** (1952), 409–436 (1953). MR0060307

[13] Yann LeCun, Corinna Cortes, and Burges Christopher J. C. Mnist database. `http://yann.lecun.com/exdb/mnist/`, accessed 2017.

[14] 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, interacting particle systems, and integrable systems, Math. Sci. Res. Inst. Publ., vol. 65, Cambridge Univ. Press, New York, 2014, pp. 411–442. MR3380694

[15] Natesh S. Pillai and Jun Yin, *Universality of covariance matrices*, Ann. Appl. Probab. **24** (2014), no. 3, 935–1001, DOI 10.1214/13-AAP939. MR3199978

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

[17] Craig A. Tracy and Harold Widom, *Level-spacing distributions and the Airy kernel*, Comm. Math. Phys. **159** (1994), no. 1, 151–174. MR1257246