Power Laws in isotropic linear least squares

aka, Spectral Origins of Power-Law Learning: A Triple Descent Phenomenon in Kaczmarz and SGD

Background

Suppose you train your neural network. Should you expect loss to decay

1. Exponentially with time
2. Power-law

Many optimization results are exponential https://chatgpt.com/c/68b1bd8d-84d8-8324-bf01-594c29c22afa

t=time
d=dimensions

1. t->infinity, d=fixed (classical optimization) *t>>d*
2. t=fixed, d->infinity (NTK) *d>>t*
3. t=d, t->infinity, d->infinity (**this work**) d≈t

Main take away

For uniform linear least squares

mean squared loss $\sqrt{\text{epochs}}$

RMS $(\text{epochs})^{1/4}$

Setup

Model

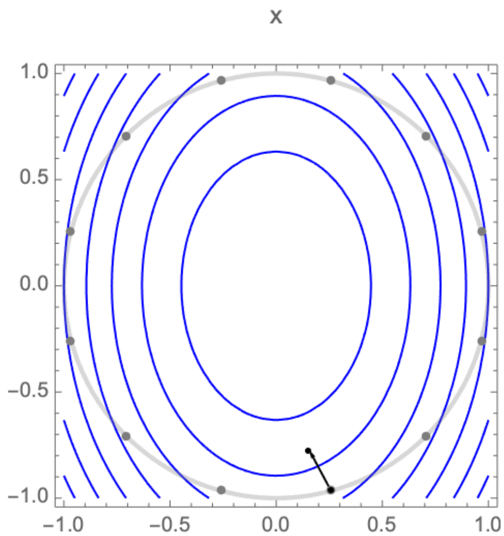
1. Realizable linear regression <xi, w> = bi forall xi for i in 1..N

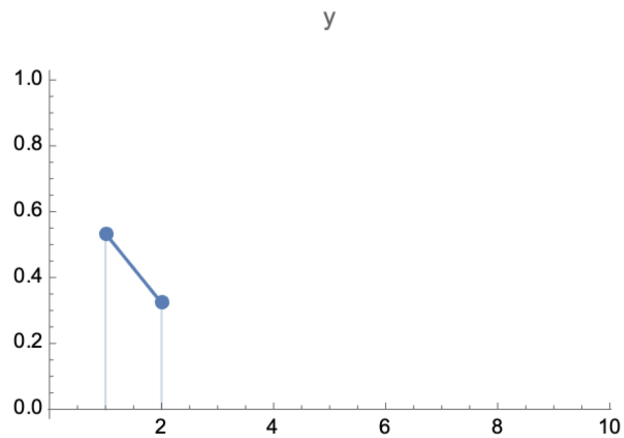
d dimensions

Minimize least squares loss, one example at a time.

Initialization

2. Isotropically initialized -- random w0 such that every w0-w* is equally likely





WLOG ||w0-w*||=1 if we care about relative drop
WLOG w*=0

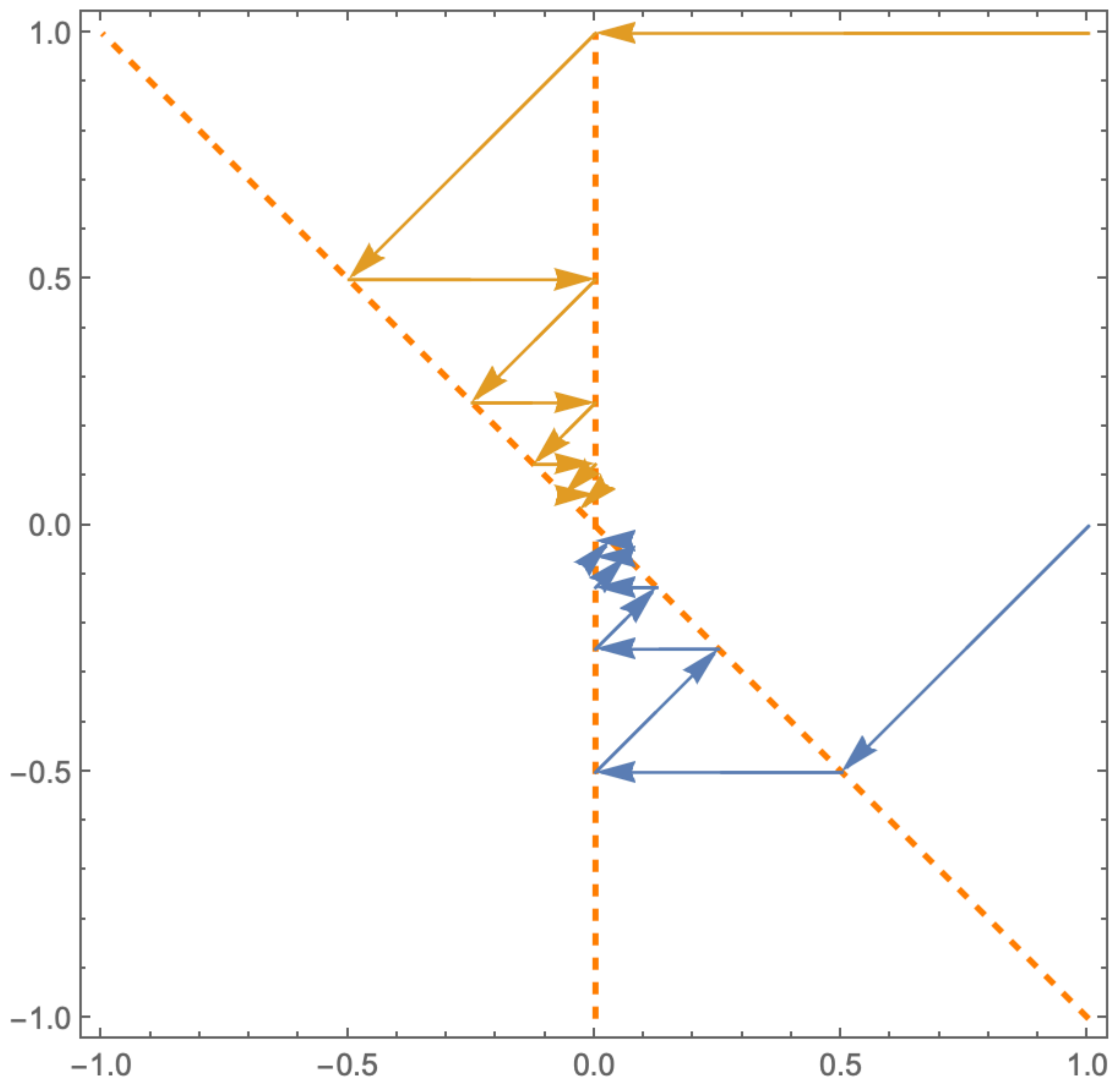
squared error at step_i = ||wi||^2

Dataset + step size

xi isotropic

3. Normalize data, such that ||xi||=1 for all i and use step-size=1

equivalent to Kaczmarz
equivalent to batch-size=1 SGD with greedy line search for each batch



Key question

How does mean squared loss behave as a function of time?

The math

Each step is a projection.

w1 = (I-xx')w0

After going through the whole epoch, it's the same as initial error vector by this matrix: https://mathoverflow.net/questions/475439/spectrum-of-prod-id-lefti-x-ix-it-right-for-isotropic-x-i

Suppose $x_i \in \mathrm{R}^d$ are IID isotropic random vectors with $\|x_i\| = 1$ and matrix A_d is defined as follows:

$$A_d = \prod_i^d \left(I - x_i x_i^T\right)$$

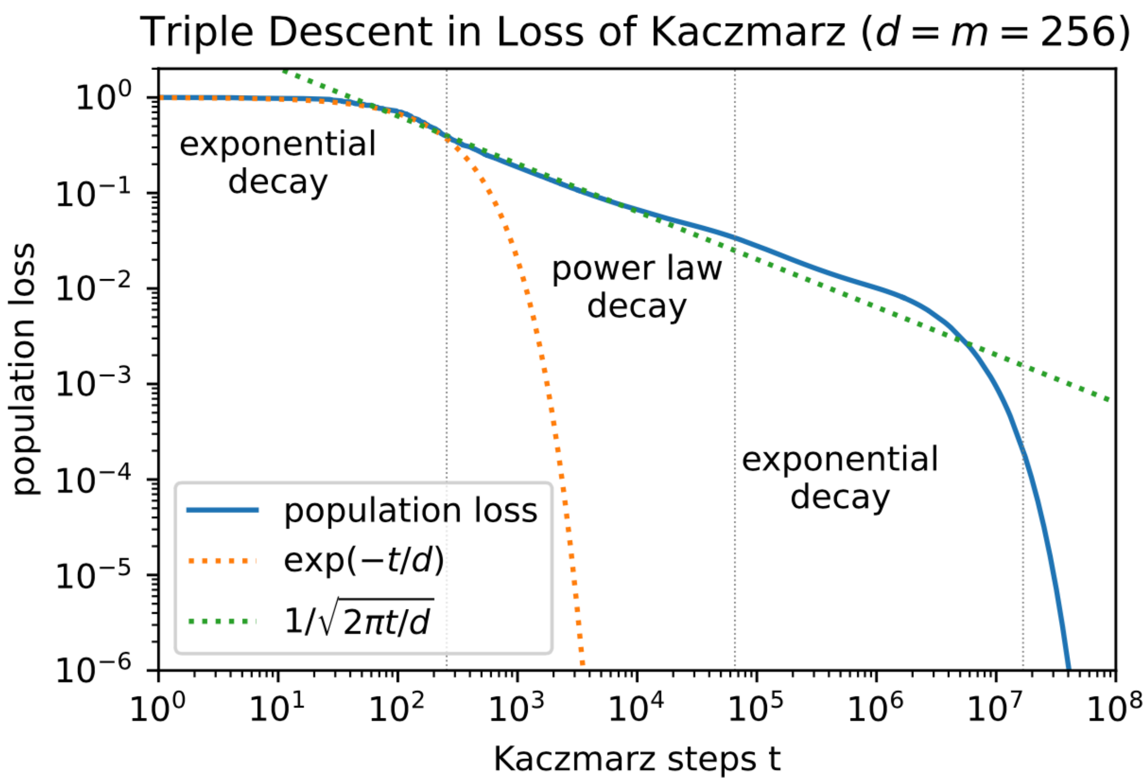
mean squared loss after s epochs =

$$\frac{s^s}{e^s s!}$$

Finite data

1. Exponential in the head

3. Exponential in the tail



History

1. Trying to get closed form,
https://mathoverflow.net/questions/475439/spectrum-of-prod-id-lefti-x-ix-it-right-for-isotropic-x-i

2. Chris Re/Chris DeSa
    - got closed form using free probability
    - results for other step-sizes

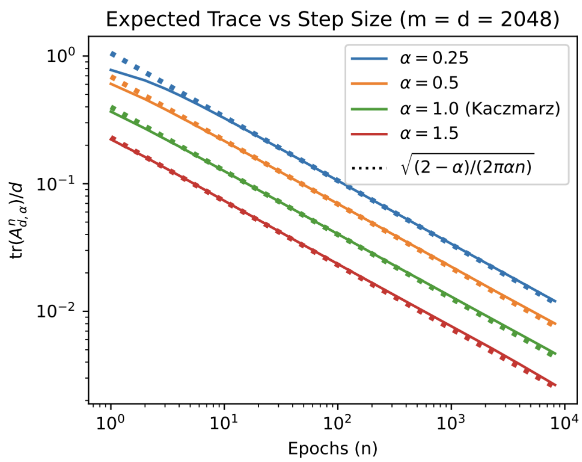


Figure 6: Empirical validation of Theorem 4.1 by running SGD for various α on a random linear regression problem.

Thomas Ahle:

- generalize beyond flip-flop
- shuffling strategies (shuffling hurts)

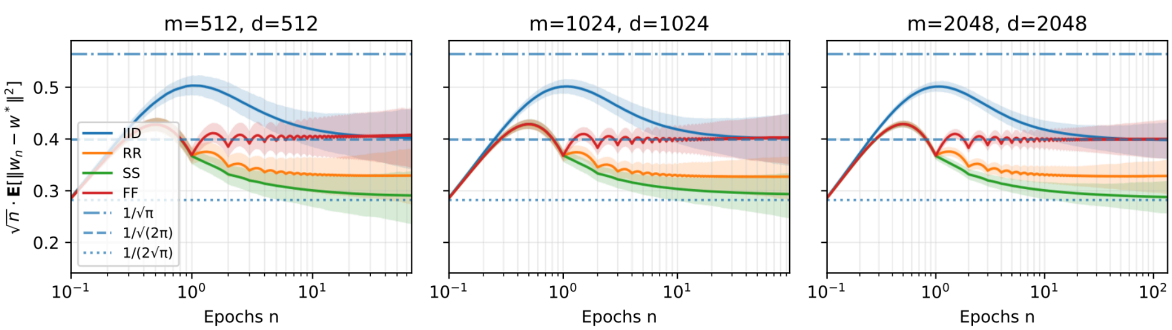


Figure 2: **Empirical Test Loss Normalized by $\sqrt{\text{epochs}}$.** We plot the mean and quartiles at 1000 independent runs of the four shuffling methods discussed. In particular Flip-Flop (red) and Single Shuffle (green) can be seen to closely follow the theorems above. At epoch $n = 1$ we have loss $1/e$ for both methods, while for larger n they follow the asymptotics of $1/\sqrt{2\pi n}$ and $1/(2\sqrt{\pi n})$.