SGD from a random start in high dimensions

Aukosh Jagannath (Waterloo)
Joint w/ G Ben Arous (NYU) R Gheissari (UC Berkeley)

Basic setting

Given:

- $(P_x)_{x \in \mathbb{R}^N}$ — Parametric family of distributions
- $(Y^\ell)_{\ell=1}^M$ — i.i.d. observations from P_{x_0}

Goal: Estimate x_0

Risk minimization

Approach: $L(x; Y)$ — Loss function

Population Loss: $\Phi(x) = \mathbb{E}L(x; Y)$

Ideal Estimator: $\hat{x} = \operatorname{argmin}_x \mathbb{E}L(x; Y)$

Issue: Don't have access to "true" distribution

Fix: Empirical Risk Minimization or Stochastic Approximation

Stochastic Approximations

- Sequentially optimize loss on new data points [Robbins-Monro '51]
- Each sample gives approximation to population:

$$L(x, Y^{\ell}) = \Phi(x) + \underbrace{H_{\ell}(x)}$$

Sample-wise error

- Proxy for gradient descent on population

Stochastic gradient descent (SGD)

Algorithm:

Input: $\underbrace{X_0}, L, (Y^\ell)_{\ell=1}^M, \underbrace{\delta}$

initial guess step-size

Update: $X_{t+1} = X_t + \delta \nabla L(X_t, \underbrace{Y^{t+1}})$

new sample

Output: X_M

Q: How many samples needed for convergence?

"Sample complexity"

Two phases of stochastic gradient descent

Heurstic picture: [Bottou '99, Mandt-Hoffman-Blei '17]

1. **Search phase**
   - Start in high entropy region
   - Fluctuations dominates
   - Walker wandering in complex landscape

2. **Descent phase**
   - outperforms a random guess
   - Descends to minimum
   - Trust region (or at least a "basin"?)

Limit theory (Fixed N)

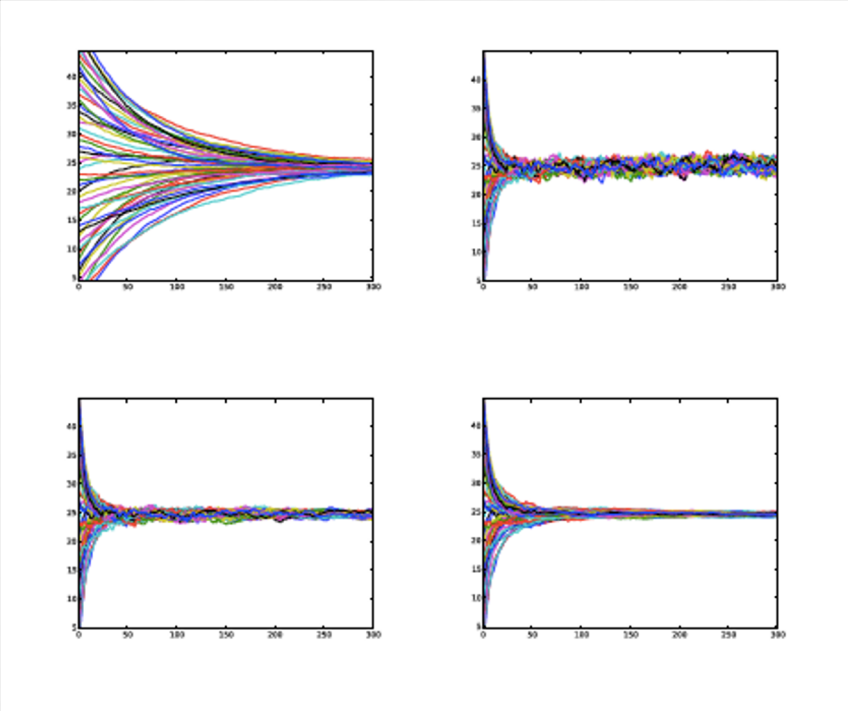
Stochastic approximation:

$$\underbrace{\nabla L}_{\text{loss}}(x, Y) = \underbrace{\nabla \Phi}_{\text{population loss}}(x) + \underbrace{\nabla H}_{\text{fluctuation}}(x)$$

Perturbation of gradient flow in infinite size $M \to \infty$ for fixed N

Limit theory [Robbins-Monro '51, McLeish '76, Ljung '77, Benaïm-Hirsch '96]

- Ignore "burn-in time"
- Convergence to GD for $\Phi(x)$
- Connects to dynamical systems



[credit: J. Le Ny '09]

Toward a high-dimensional theory

- One of go-to methods in modern data science
- Used to tackle extremely complex inference tasks
  - High-dimensional data
  - Complex models
- Performs well in very diverse domains
  - Computer vision/Image processing
  - Prediction
  - Healthcare

Today's talk:

- How many samples needed in high-dim? ($\mathrm{M} \asymp poly(N)$)

Recent progress in high-dimensions

1. Convexity, Quasi-convexity, … [Bottou '98, Bottou-Le Cun '04, Needel-Srebro-Ward '14, Harvey-Liaw-Plan-Randhawa '19, Dieuleveut-Durmus-Bach '19…]
   - Ignore search and focus on rates assuming shape of basins
2. Langevin dynamics or SDE approximations [Raginsky-Rakhlin-Telegarsky '17, Zhang-Liang-Charikar '17, Ma-Chen-Jin-Flammarion-Jordan '19, Cheng-Yin-Bartlett-Jordan '20…]
   - Study an SDE approximation to the dynamics
   - Polynomial mixing time bounds $O(poly(N)e^{LR^2})$
     - Empirical risk is L-Lipschitz, K-smoothness (gradient is K-lipschitz)
     - Fixed domain: B(0,R)
     - Ellipticity, reversibility …

High-dim statistical models don't fit

Issue: Standard tasks don't fit either setting

1. SGD still performs well with non-convex problems
   - Complex data (tensors, neural networks, …)
2. Dimension dependence of Lipschitz constants
   - With high probability in realization Lip ~ N^c
     - Linear regression, Phase retrieval, Spiked matrix models
   - Normalizing can render invariant measure uninformative

Concentration of measure:

"1-Lipschitz functions of many variables are nearly constant"

Recent progress in high-dimensions

1. Convexity, Quasi-convexity, … [Bottou '98, Bottou-Le Cun '04, Needel-Srebro-Ward '14, Harvey-Liaw-Plan-Randhawa '19, Dieuleveut-Durmus-Bach '19…]
   - Ignore search and focus on rates assuming shape of basins
2. Langevin dynamics or SDE approximations [Raginsky-Rakhlin-Telegarsky '17, Zhang-Liang-Charikar '17, Ma-Chen-Jin-Flammarion-Jordan '19, Cheng-Yin-Bartlett-Jordan '20…]
   - Study an SDE approximation to the dynamics
   - Polynomial mixing time bounds $O(poly(N)e^{LR^2})$
     - Empirical risk is L-Lipschitz, K-smoothness (gradient is K-lipschitz)
     - Fixed domain: B(0,R)
     - Ellipticity, reversibility …
3. **Scaling limits and bounding flows** [Cugliandolo-Kurchan '92, Saad-Solla '95, Ben Arous-Dembo-Guionnet '04, Tan-Vershynin '19+, Goldt-Mézard-Krzakala-Zdeborová 20, Ben Arous-Gheissari-J '20-21,…]

Today's talk

Focus for today's talk:

Regimes relevant to high-dimensional statistics
Uninformed initializations

1. How many samples do you need? (sample complexity)
2. How much time does it take to beat a random guess?
3. What are the fundamental properties of a problem that govern the answer to these questions?

Model and assumptions

A simple class of models

Assumptions:

- Population loss: $\Phi(x) = \mathbb{E}L(x; Y)$
- (Non-linear) function of distance to ground truth
- Bounded domain + fixed noise level → know norm

Parameter space: S^{N-1} unit sphere in R^N

Population loss: $\Phi(x) = \phi(m(x))$

$$m(x) = x \cdot x_0$$

x_0 – parameter to be inferred

Stochastic gradient descent

Algorithm:

Input: $\underbrace{X_0}_{\text{initial guess}}, L, (Y^\ell)_{\ell=1}^M, \underbrace{\delta}_{\text{step-size}}$

Update:
$$\begin{cases} \tilde{X}_{t+1} = X_t + \delta \nabla L(Y^{t+1}, X_t) \\ X_{t+1} = \sqrt{N} \frac{\tilde{X}_{t+1}}{\|X_{t+1}\|} \end{cases} \leftarrow \text{projection}$$

Output: X_M

Assumption A: Regularity

$$\nabla L(x, Y) = \nabla \Phi(x) + \underbrace{\nabla H(x)}$$

Sample-wise error

Naively: Worst case if error term is "completely random"
Idea: gradient of error no worse than *i.i.d.* vector with a few moments

Def: A sequence of data distributions and losses, (P_N, L_N) satisfies ***Assumption A*** if the sample-wise error satisfies the following norm bounds:

$$\sup_{x \in \mathbb{S}^{N-1}} \mathbb{E}[(\nabla H(x) \cdot x_0)^2] \leq C$$

$$\sup_{x \in \mathbb{S}^{N-1}} \mathbb{E}[\|\nabla H(x)\|^{4+\epsilon}] \leq C N^{\frac{4+\epsilon}{2}}$$

x_0 – parameter to be inferred

Assumption B: Fisher-type consistency

Fisher consistency: estimator correct given population.

- Gradient descent on ϕ consistent with random start
- ϕ even → can only determine up to a sign
- Random start is on upper half sphere with prob ½

Def: A population loss satisfies ***Assumption B*** if:

$$\phi'(t) < 0 \text{ for } t \in (0,1)$$

Sample complexity

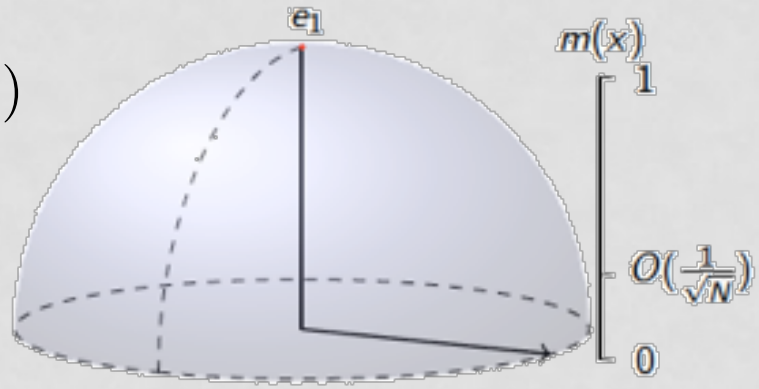
Information exponent

Def: A population loss Φ_N has **information exponent** k if $\phi \in C^{k+1}([-1,1])$ and

$$\begin{cases} \frac{d^\ell}{dm^\ell}\phi(0) = 0 & \ell \le k-1 \\ \frac{d^k}{dm^k}\phi(0) < -c \end{cases}$$

Recall: $\Phi_N(x) = \mathbb{E}L_N(Y;x) = \phi_N(m(x))$

Typical start: $x_1^k \simeq (1/\sqrt{N})^k$



Examples

k = 1:

- Linear regression with random covariates
- Generalized linear models with random covariates
- Asymmetric two component Gaussian mixture

k = 2:

- Symmetric Gaussian mixture
- Phase retrieval
- Online PCA
- Spiked Wigner models

k ≥ 3:

- Tensor PCA

Variable:

- Single-layer networks (exponent depends on activation)

Performance guarantee

Initialization: μ_N^+ uniform measure conditioned on upper-half sphere

Thm 1: Suppose Assumptions **A** and **B** hold.

For information exponent **k,** if $M = \alpha_N N$ has

1. **(k=1)** $\alpha_N >> 1 = \alpha_c(N,1)$
2. **(k=2)** $\alpha_N >> \log(N)^2 = \alpha_c(N,2)\log(N)$
3. **(k≥3)** $\alpha_N >> N^{k-1}\log(N)^2 = \alpha_c(N,k)\log(N)^2$,

then SGD started from μ_N^+ with step size $\delta_N \sim \alpha_N^{-1+\varepsilon}$ produces a consistent estimator:

$$m(X_M) \to 1 \text{ in probability.}$$

Critical sample complexity: $\alpha_c(N,k) = \begin{cases} 1 & k=1 \\ \log N & k=2 \\ N^{k-2} & k \geq 3 \end{cases}$

Refutation

Initialization: μ_N^+ uniform measure conditioned on upper-half sphere

Thm 2: Suppose Assumptions **A** and **B** hold.

For information exponent **k,** if $M = \alpha_N N$ has

1. **(k=1)** $\alpha_N << \alpha_c(N,1)$ and $\delta_N = O(1)$
2. **(k ≥ 2)** $\alpha_N << \alpha_c(N,2)$ and $\delta_N = O(\alpha_N^{-1/2+\varepsilon})$

then SGD started from μ_N^+ does not corellate:

$$m(X_M) \to 0 \text{ in probability.}$$

Critical sample complexity: $\alpha_c(N,k) = \begin{cases} 1 & k=1 \\ \log N & k=2 \\ N^{k-2} & k \geq 3 \end{cases}$

Rapid descent

- τ_ϵ- first hitting time for $\{m(x) = \epsilon\}$

Thm 3: Suppose Assumptions **A** and **B** hold.
For information exponent **k ≥ 2**, if M = α_NN as in Theorem 1, then for any $\epsilon > 0$, the first hitting time for latitude ϵ and $1 - \epsilon$ satisfy $|\tau_\epsilon - \tau_{1-\epsilon}| = O(N)$.
Furthermore, $m(\mathrm{X}_t) > 1 - 2\epsilon$ for t > $\tau_{1-\epsilon}$.

Summary

- For random initializations there are **three regimes**:
  1. Linear (k =1): needs linear in N samples
  2. Quasi-linear (k = 2): needs ≥ N log(N) and ≤ N log(N)^2
  3. Polynomial (k ≥ 3): needs ~N^{k-1} samples
- Critical sample complexity:

$$\alpha_c(N,k) = \begin{cases} 1 & k = 1 \\ \log N & k = 2 \\ N^{k-2} & k \geq 3 \end{cases}$$

- Once at latitude ϵ:
  - can get to $1 - \epsilon$ in linear time.
  - Law of large numbers (back to finite dim story)

Examples

Linear (k = 1):

- Linear regression with random covariates
- Generalized linear models with random covariates
- asymmetric two component Gaussian mixture

Quasilinear (k = 2):

- symmetric Gaussian mixture
- phase retrieval
- Online PCA
- spiked Wigner models

Polynomial (k ≥ 3):

- Tensor PCA

Variable:

- Single-layer networks (exponent depends on activation)

Some Insights

A motivating example

Task: supervised learning with one-layer networks

- Teacher-Student networks, single-index or non-linear factor model, perceptron, generalized phase retrieval (GLM)…

Given:

- Activation function: f
- (Random) feature vectors: (a^ℓ)
- M i.i.d. non-linear measurements of unknown unit N-vector

$$y^\ell = f(a^\ell \cdot x_0) + \varepsilon^\ell$$

Goal: Estimate optimal weight $x_0 \in R^N$

Approach: SGD on ℓ_2 loss from a **random** start

- **Spectral initializations:** Candès-Li-Soltanolkotabi '15, Li-Lu '20, Mondelli-Montanari '18, Maillard-Krzakala-Lu-Zdeborová '21

Supervised learning with Gaussian features

- i.i.d. Standard gaussian features (a^ℓ)
- i.i.d. centered errors (ε^ℓ) with finite 4^+ moment

Population loss: (let $u_j(f)$ = j[th] Hermite coefficient)

$$\Phi(x) = 2\sum_{j=1}^{\infty} u_j(f)^2(1 - m(x)^j) + c$$

Information Exponent:

index of first nonzero Hermite coefficient

Examples

Linear (k = 1):

- Adaline ($f(x) = x$) has exponent 1
- Sigmoid ($f(x) = (1+\exp(-x))^{-1}$) has exponent 1
- ReLu ($f(x) = \max(x,0)$) has exponent 1

Quasi-linear (k = 2):

- **Phase retrieval** $f(x) = x^2$ or $|x|$ has exponent 2
- Monomial $f(x) = x^k$ has exponent 1 or 2 depending on parity

Polynomial (k ≥ 3):

- Hermite polynomials: $f(x) = h_k(x)$ has exponent k
- Activations in subspace spanned by Hermite polynomials of degree at least 3.

How much data for search phase?

Search v.s. descent trade-off

Q: How much data do I spend in the "burn-in"

- Total amount of data for estimation
  - Exponent = 1 → N samples
  - Exponent ≥ 2 → needs at least N log N samples
- Amount of data used in "descent"
  - Once at latitude ϵ, can get to $1 - \epsilon$ in $O(N)$ time

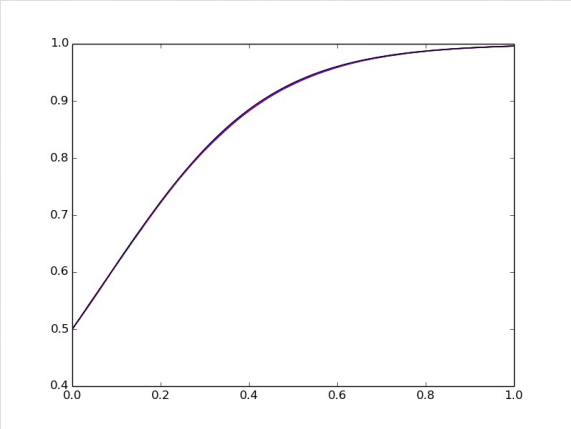
Most of time spent/data used is for search phase!

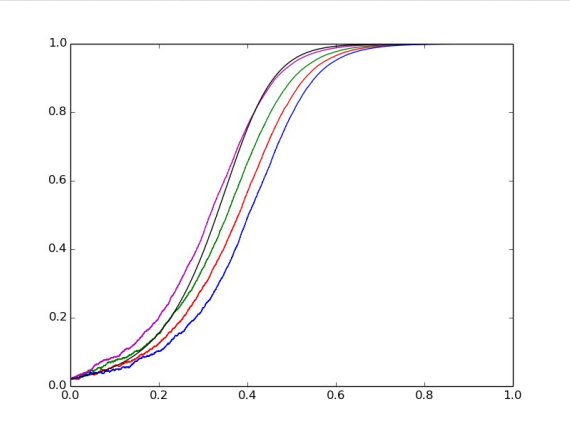
Most of data spent in search phase

Warm start

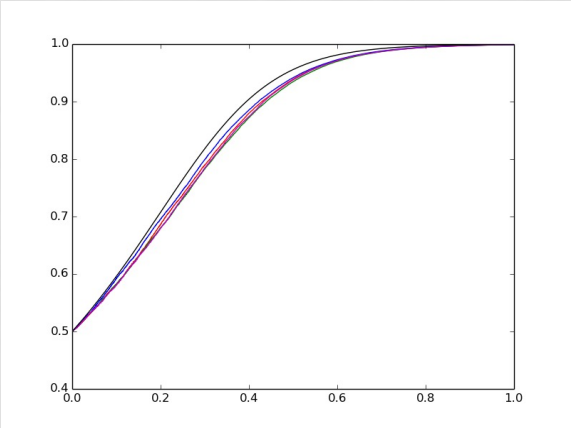
Random initialization

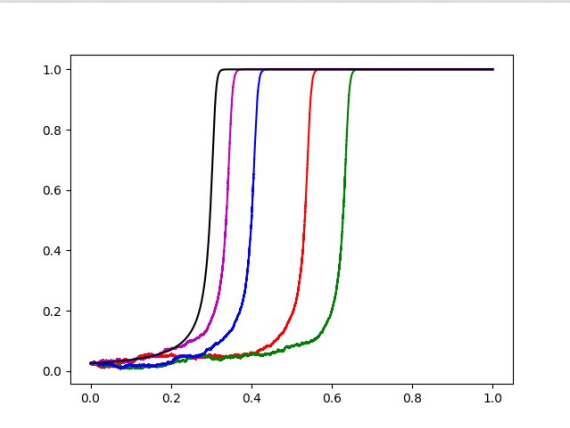
Phase Retrieval x^2





Hermite: x^3-3x



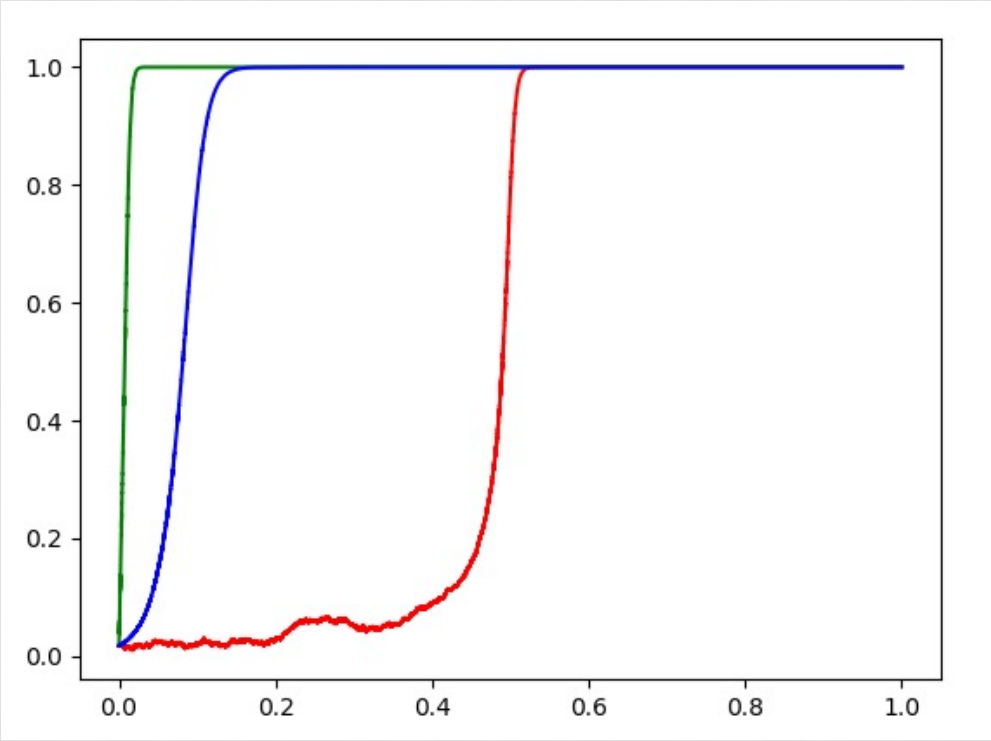


Once correlated all problems in "easy phase"
Pathwise LLN once warm

Impact of activation on run-time

Small changes to activation can dramatically change runtime!

N=3000, M = 90,000,000



cubic (x^3) vs quadratic (x^2) vs 3rd hermite (x^3-3x)

Summary

- Stochastic gradient descent for "rank 1" models
- Sample complexity determined by information exponent **k**
- For random initializations there are **three regimes:**
  1. Linear (k = 1): needs N samples
  2. Quasi-linear (k = 2): needs $\geq N \cdot \log(N)$ and $\leq N \cdot \log(N)^2$
  3. Polynomial (k ≥ 3): needs $\sim N^{k-1}$ samples

Take away's:

1. Many classical tasks have k ≤ 2
2. If k ≥ 2 → most of data used in search phase.
3. Performance depends dramatically on activation/loss (misspecification can cause major issues!)

Proof techniques

- Consider population dynamics $m_t \approx m_{t-1} + \frac{\delta}{N} c m^{k-1}$
- Direct analysis of difference equation if $m_0 \sim N^{-\zeta}$
  1. k < 2: needs time $\delta^{-1} N$
  2. k = 2: needs time $\delta^{-1} N \log N$
  3. k > 2: needs time $\delta^{-1} N^{1+\zeta(k-2)}$
- One idea: send N to infinity and step-size to zero first
- **Issue:** nontrivial fixed point at 0
- Most time spent on microscopic scales
- Instead use bounding flows approach of [Ben Arous-Gheissari-J '20]
- Due to martingale can avoid control of initialization

Thanks for listening!

References:

- G. Ben Arous, R. Gheissari, A.J., "Online stochastic gradient descent on non-convex losses from high-dimensional inference", *Jour. Mach. Learn. Res. 2021*
- G. Ben Arous, R. Gheissari, A.J., "Algorithmic thresholds for Tensor PCA", *Ann. Probab 2020*
- G. Ben Arous, R. Gheissari, A.J., "Bounding flows for spherical spin glass dynamics", *Commun. Math. Phys. 2020*

Cette recherche a été financée par CRSNG. This research was supported by NSERC.