Spin Glass and High Dimensional Energy Landscape

Binan Gu

Department of Mathematical Sciences, New Jersey Institute of Technology

New Jersey Institute of Technology
Spring 2021 Machine Learning Talk II

NJIT
New Jersey's Science & Technology University

Main scientists

Gérard Ben Arous (Mathematician)

- Director of Courant (2011-2016)

Main scientists

Gérard Ben Arous (Mathematician)

- Director of Courant (2011-2016)
- A probabilist working on Spin Glass, large deviations.

Main scientists

Gérard Ben Arous (Mathematician)

- Director of Courant (2011-2016)
- A probabilist working on Spin Glass, large deviations.
- For today's work: theoretical basis for minima search of random functions in high dimensions.

Yann LeCun (Computer Scientist)

Main scientists

Gérard Ben Arous (Mathematician)

- Director of Courant (2011-2016)
- A probabilist working on Spin Glass, large deviations.
- For today's work: theoretical basis for minima search of random functions in high dimensions.

Yann LeCun (Computer Scientist)

- Chief AI scientist at Facebook, Turing award.

Main scientists

Gérard Ben Arous (Mathematician)

- Director of Courant (2011-2016)
- A probabilist working on Spin Glass, large deviations.
- For today's work: theoretical basis for minima search of random functions in high dimensions.

Yann LeCun (Computer Scientist)

- Chief AI scientist at Facebook, Turing award.
- Professor at NYU CS, Data Science, Neural Science, EE and CE.

Main scientists

Gérard Ben Arous (Mathematician)

- Director of Courant (2011-2016)
- A probabilist working on Spin Glass, large deviations.
- For today's work: theoretical basis for minima search of random functions in high dimensions.

Yann LeCun (Computer Scientist)

- Chief AI scientist at Facebook, Turing award.
- Professor at NYU CS, Data Science, Neural Science, EE and CE.
- A computer scientist working on AI, machine learning, computer vision and computational neuroscience.

Main scientists

Gérard Ben Arous (Mathematician)

- Director of Courant (2011-2016)
- A probabilist working on Spin Glass, large deviations.
- For today's work: theoretical basis for minima search of random functions in high dimensions.

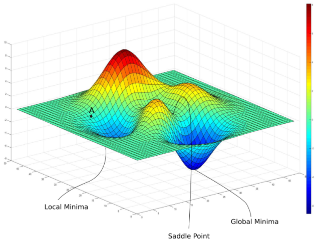
Yann LeCun (Computer Scientist)

- Chief AI scientist at Facebook, Turing award.
- Professor at NYU CS, Data Science, Neural Science, EE and CE.
- A computer scientist working on AI, machine learning, computer vision and computational neuroscience.
- For today's work: gradient-based machine learning algorithmic setup.

An Example of ML

Binary Classification Problems with many layers

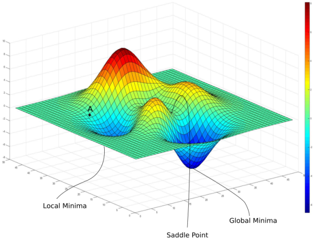
- Provided feature data and their labels (say, cats and dogs).



An Example of ML

Binary Classification Problems with many layers

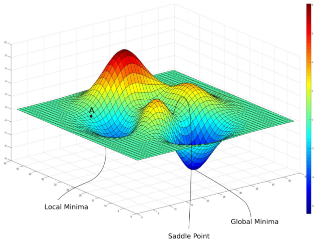
- Provided feature data and their labels (say, cats and dogs).
- Learn a function between features and labels (through several filters, say, eye shape, fur density, etc.). Error minimisation takes place here.



An Example of ML

Binary Classification Problems with many layers

- Provided feature data and their labels (say, cats and dogs).
- Learn a function between features and labels (through several filters, say, eye shape, fur density, etc.). Error minimisation takes place here.
- Test this function/quality of minimisation on new data.



ML Algorithmic Setup

- A probability measure μ that represents given labelled data.

ML Algorithmic Setup

- A probability measure μ that represents given labelled data.
- An unknown labelling function $\mathcal{G}$ (true relation).

ML Algorithmic Setup

- A probability measure μ that represents given labelled data.
- An unknown labelling function $\mathcal{G}$ (true relation).
- A metric *d* on appropriate function spaces.

ML Algorithmic Setup

- A probability measure μ that represents given labelled data.
- An unknown labelling function $\mathcal{G}$ (true relation).
- A metric d on appropriate function spaces.

Choose your favourite loss (error) L and minimise with stochastic gradient descent (SGD) on

$$\min_{\boldsymbol{w}} \left\{ \frac{1}{N} \sum_{n=1}^{N} L\left(h_{\boldsymbol{w}}\left(\boldsymbol{x}_n\right), y_n \right) \right\} = f\left(\boldsymbol{w}\right)$$

ML Algorithmic Setup

- A probability measure μ that represents given labelled data.
- An unknown labelling function $\mathcal{G}$ (true relation).
- A metric *d* on appropriate function spaces.

Choose your favourite loss (error) L and minimise with stochastic gradient descent (SGD) on

$$\min_{\boldsymbol{w}} \left\{ \frac{1}{N} \sum_{n=1}^{N} L\left(h_{\boldsymbol{w}}\left(\boldsymbol{x}_n\right), y_n\right) \right\} = f\left(\boldsymbol{w}\right)$$

Write

$$\nabla f\left(\boldsymbol{w}\right) = \frac{1}{N} \sum_{n=1}^{N} \nabla f_n\left(\boldsymbol{w}\right) \approx \frac{1}{|B|} \sum_{n \in B} \nabla f_n\left(\boldsymbol{w}\right)$$

ML Algorithmic Setup

- A probability measure μ that represents given labelled data.
- An unknown labelling function $\mathcal{G}$ (true relation).
- A metric d on appropriate function spaces.

Choose your favourite loss (error) L and minimise with stochastic gradient descent (SGD) on

$$\min_{\boldsymbol{w}} \left\{ \frac{1}{N} \sum_{n=1}^{N} L\left(h_{\boldsymbol{w}}\left(\boldsymbol{x}_n\right), y_n\right) \right\} = f\left(\boldsymbol{w}\right)$$

Write

$$\nabla f\left(\boldsymbol{w}\right) = \frac{1}{N} \sum_{n=1}^{N} \nabla f_n\left(\boldsymbol{w}\right) \approx \frac{1}{|B|} \sum_{n \in B} \nabla f_n\left(\boldsymbol{w}\right)$$

and the update adaptively,

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \frac{\mu^t}{|B|} \sum_{n \in B} \nabla f_n\left(\boldsymbol{w}\right), \text{with } \mu^t \to 0 \text{ appropriately.}$$

Convergence

Law of large numbers yields

$$\frac{1}{N}\sum_{n=1}^{N} L\left(h_{\mathbf{w}}\left(\boldsymbol{x}_n\right), y_n\right) \overset{a.s.}{\to} \mathbb{E}_{\mu}\left[L\left(h_{\mathbf{w}}\left(\boldsymbol{x}\right), y\right)\right]$$

while CLT yields

$$\sqrt{N}\left(\frac{1}{N}\sum_{n=1}^{N} L\left(h_{\mathbf{w}}\left(\boldsymbol{x}_n\right), y_n\right) - \mathbb{E}_{\mu}\left[L\left(h_{\mathbf{w}}\left(\boldsymbol{x}\right), y\right)\right]\right) \overset{law}{\to} \mathcal{N}\left(0, \sigma^2\left(\boldsymbol{w}\right)\right)$$

A Practical Problem in Loss Minimisation

Loss function characteristics

- Non-convex.

A Practical Problem in Loss Minimisation

Loss function characteristics

- Non-convex.
- High-dimensional domain Ω (lots of parameters!).

A Practical Problem in Loss Minimisation

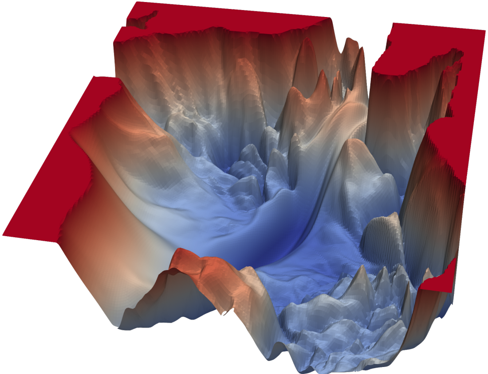
Loss function characteristics

- Non-convex.
- High-dimensional domain Ω (lots of parameters!).
- Exponentially many critical points of various indices, i.e. some negative eigenvalues of the Hessian.

A Practical Problem in Loss Minimisation

Loss function characteristics

- Non-convex.
- High-dimensional domain Ω (lots of parameters!).
- Exponentially many critical points of various indices, i.e. some negative eigenvalues of the Hessian.



A Practical Problem in Loss Minimisation

Loss function characteristics

- Non-convex.
- High-dimensional domain Ω (lots of parameters!).
- Exponentially many critical points of various indices, i.e. some negative eigenvalues of the Hessian.

Problems with GD/SGD

A Practical Problem in Loss Minimisation

Loss function characteristics

- Non-convex.
- High-dimensional domain Ω (lots of parameters!).
- Exponentially many critical points of various indices, i.e. some negative eigenvalues of the Hessian.

Problems with GD/SGD

- Searcher gets stuck in a critical point or flat region, and thus long search time.

A Practical Problem in Loss Minimisation

Loss function characteristics

- Non-convex.
- High-dimensional domain Ω (lots of parameters!).
- Exponentially many critical points of various indices, i.e. some negative eigenvalues of the Hessian.

Problems with GD/SGD

- Searcher gets stuck in a critical point or flat region, and thus long search time.
- Perturbing the gradient in this case (thereby changing energy landscape) is of insignificant improvement.

Remedies

- Shrink search space by specifying a “floor” (existence proved), a level set of the loss in which bulk of the low index critical points lie in the absence of an external field.

Remedies

- Shrink search space by specifying a “floor” (existence proved), a level set of the loss in which bulk of the low index critical points lie in the absence of an external field.

Advantage

Floor has energy low enough that global and local minima are about the same.

Remedies

- Shrink search space by specifying a "floor" (existence proved), a level set of the loss in which bulk of the low index critical points lie in the absence of an external field.

Advantage

Floor has energy low enough that global and local minima are about the same.

- Add a tunable random external field and reduce strength as SGD progresses (AnnealedSGD [3]).

$$L(\mathbf{x}, \mathbf{w}) = \sum_n x_{i_1,\ldots,i_n} w_{i_1} \ldots w_{i_n} + \sum_n r_n w_{i_n}$$

where $r_n \sim \mathcal{N}\left(0, v^2\right)$, v tunable.

Remedies

- Shrink search space by specifying a “floor” (existence proved), a level set of the loss in which bulk of the low index critical points lie in the absence of an external field.

Advantage

Floor has energy low enough that global and local minima are about the same.

- Add a tunable random external field and reduce strength as SGD progresses (AnnealedSGD [3]).

$$L(\mathbf{x}, \mathbf{w}) = \sum_n x_{i_1,\ldots,i_n} w_{i_1} \ldots w_{i_n} + \sum_n r_n w_{i_n}$$

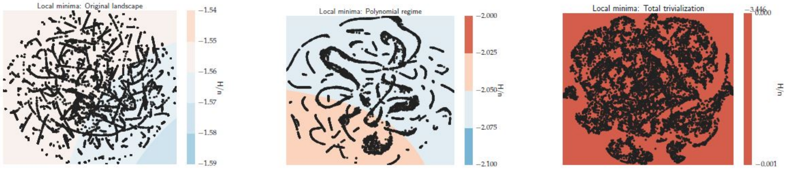
where $r_n \sim \mathcal{N}\left(0, v^2\right)$, v tunable.

Advantage

From complex terrain to degeneracy without affecting the locations of local minima of the original problem (proved in [3]).

Non-sharp phase transition

There exists a threshold of noise such that a non-sharp phase transition occurs.



(A) Exponential regime: Local minima are seen here as isolated dots surrounded by high energy barriers while saddle points are seen as narrow connected regions, viz., regions where the gradient is very small in all but a few directions.

(B) Polynomial regime: The number of isolated clusters, i.e., local minima, is significantly smaller as compared to Fig. 2a. As the discussion in Sec. 4.3 predicts, the energy landscape seems to be full of saddle points in the polynomial regime.

(C) Trivial regime: Gradient descent always converges to the same location. The average cosine distance (on $S^{n-1}(\sqrt{n})$) here is 0.02 as compared to 1.16 for Fig. 2a which suggests that this is indeed a unique local minimum.

Ferromagnetism model

Spin Glass overview

Ferromagnetism model

- Particles with magnetic spins ± 1 on $\mathbb{Z}^d$.

Spin Glass overview

Ferromagnetism model

- Particles with magnetic spins ± 1 on $\mathbb{Z}^d$.
- Random interaction, short (sparse) or long (dense) ranged.

Spin Glass overview

Ferromagnetism model

- Particles with magnetic spins ± 1 on $\mathbb{Z}^d$.
- Random interaction, short (sparse) or long (dense) ranged.
- Minimising total "energy" finds the equilibrium state.

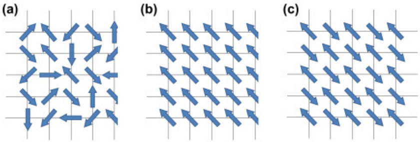
Spin Glass overview

Ferromagnetism model

- Particles with magnetic spins ± 1 on $\mathbb{Z}^d$.
- Random interaction, short (sparse) or long (dense) ranged.
- Minimising total "energy" finds the equilibrium state.
- Perturb to check stability.

Ferromagnetism model

- Particles with magnetic spins ± 1 on $\mathbb{Z}^d$.
- Random interaction, short (sparse) or long (dense) ranged.
- Minimising total "energy" finds the equilibrium state.
- Perturb to check stability.



Nearest neighbor constant:	Ising
Nearest neighbor random:	Edward-Anderson
Long range random:	Sherrington-Kirkpatrick

Spin Glass Setup

Consider $w = (w_1, \ldots, w_n)$ an array of ± 1's sitting on some domain Ω, e.g. $\mathbb{Z}^d$. Let x_{ij} be centered correlated Gaussian variables. Then, the *Hamiltonian* of w, a 2-spin system, is given by

$$-H(w) = \sum_{i,j} x_{ij} w_i w_j + \sum_j h_j w_j$$

where h is some external field. The above example is the Sherrington-Kirkpatrick model under magnetic field h.

Properties of Spin Glass

- If x_{ij} is constant, then under h the energy minimiser is when all spins align.

Spin Glass Setup

Consider $w = (w_1, \ldots, w_n)$ an array of ± 1's sitting on some domain Ω, e.g. $\mathbb{Z}^d$. Let x_{ij} be centered correlated Gaussian variables. Then, the *Hamiltonian* of w, a 2-spin system, is given by

$$-H(w) = \sum_{i,j} x_{ij} w_i w_j + \sum_j h_j w_j$$

where h is some external field. The above example is the Sherrington-Kirkpatrick model under magnetic field h.

Properties of Spin Glass

- If x_{ij} is constant, then under h the energy minimiser is when all spins align.
- If x_{ij} is random, then we obtain a glassy state, where energy landscape becomes rugged. Local minima are hard to find or even exist.

Spin Glass Setup

Consider $w = (w_1, \ldots, w_n)$ an array of ± 1's sitting on some domain Ω, e.g. $\mathbb{Z}^d$. Let x_{ij} be centered correlated Gaussian variables. Then, the *Hamiltonian* of w, a 2-spin system, is given by

$$-H(w) = \sum_{i,j} x_{ij} w_i w_j + \sum_j h_j w_j$$

where h is some external field. The above example is the Sherrington-Kirkpatrick model under magnetic field h.

Properties of Spin Glass

- If x_{ij} is constant, then under h the energy minimiser is when all spins align.
- If x_{ij} is random, then we obtain a glassy state, where energy landscape becomes rugged. Local minima are hard to find or even exist.
- Extend $w \in \mathcal{S}^{n-1}(\sqrt{n})$ for continuous interpretation.

A Theoretical Result with Zero-one loss

Feature vector ξ (data) and p hidden layers.

A Theoretical Result with Zero-one loss

Feature vector ξ (data) and p hidden layers. Target labels $Y^t \sim Ber(q) \in \{0, 1\}$ modeled as (with denoising autoencoders)

$$Y = g\left(W^{p+1} g\left(W^p \dots g\left(W^1 \xi - \frac{d}{3}\mathbf{1}_n\right) - \frac{d}{3}\mathbf{1}_n\right) \dots - \frac{d}{3}\mathbf{1}_n\right)$$

where d is expected degree of nodes and g is some thresholding function.

A Theoretical Result with Zero-one loss

Feature vector ξ (data) and p hidden layers. Target labels $Y^t \sim Ber(q) \in \{0, 1\}$ modeled as (with denoising autoencoders)

$$Y = g\left(W^{p+1} g\left(W^p \dots g\left(W^1 \xi - \frac{d}{3}\mathbf{1}_n\right) - \frac{d}{3}\mathbf{1}_n\right) \dots - \frac{d}{3}\mathbf{1}_n\right)$$

where d is expected degree of nodes and g is some thresholding function. Suppose $Y^t \sim Ber(q)$.

A Theoretical Result with Zero-one loss

Feature vector ξ (data) and p hidden layers. Target labels $Y^t \sim Ber(q) \in \{0,1\}$ modeled as (with denoising autoencoders)

$$Y = g\left(W^{p+1} g\left(W^p \dots g\left(W^1 \xi - \frac{d}{3}\mathbf{1}_n\right) - \frac{d}{3}\mathbf{1}_n\right) \dots - \frac{d}{3}\mathbf{1}_n\right)$$

where d is expected degree of nodes and g is some thresholding function. Suppose $Y^t \sim Ber(q)$. Then up to a constant,

$$\mathbb{E}_{Y^t}\left[\hat{Y} - Y^t\right] \overset{law}{=} -H_{n,p}(w)$$

where the Hamiltonian

$$-H_{n,p}(w) = \frac{J}{n^{(p-1)/2}} \sum_{i_1,\dots,i_p=1}^{n} J_{i_1,\dots,i_p} w_{i_1,\dots,i_p}$$

with $J_{i_1,\dots,i_p}$ standard Gaussian and $w \in \mathcal{S}^{n-1}(\sqrt{n})$.

Take-aways

- ML algorithms and spin glass systems have some similarities but not entirely analogous – weights and graph connectivity are still quite different notions.

Take-aways

- ML algorithms and spin glass systems have some similarities but not entirely analogous – weights and graph connectivity are still quite different notions.
- [2] conjectures that a more universal yet undiscovered phenomenon exists, and ML algorithms and spin glasses are mere special cases of it.

Take-aways

- ML algorithms and spin glass systems have some similarities but not entirely analogous – weights and graph connectivity are still quite different notions.
- [2] conjectures that a more universal yet undiscovered phenomenon exists, and ML algorithms and spin glasses are mere special cases of it.
- Statistical mechanics is a subject worth studying to facilitate (mean-field) analysis of algorithms with random features on high-dimensional problems.

References

- Auffinger, A., Arous, G.B., Complexity of Random Smooth Functions on the High-Dimensiontal Sphere. *The Annals of Probability*. 2013.
- Sagun, L, Güney, V., Arous, G.B., LeCun, Y., Explorations on High Dimensional Landscapes. International Conference on Learning Representations. 2015.
- Chaudhari, P., Soatto, S., On the Energy Landscape of Deep Networks. 2017. `https://arxiv.org/pdf/1511.06485.pdf`.