Statistical physics of stochastic gradient descent

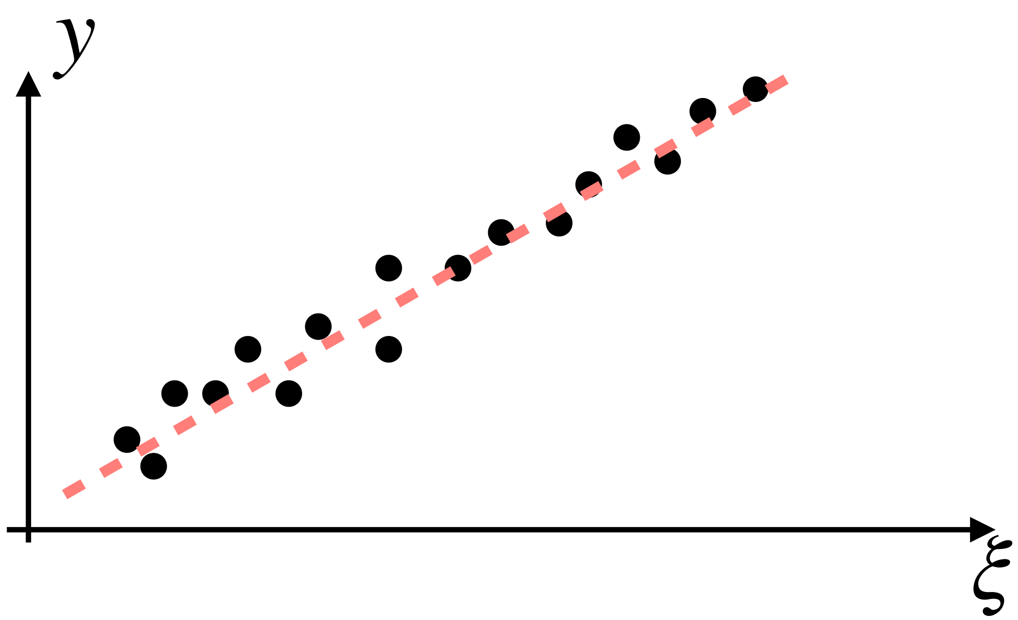
Statistical physics of learning at the IPhT

Pierfrancesco Urbani

**Université Paris-Saclay, CNRS, CEA,
Institut de Physique Théorique**

IPhT, November 2023

Linear Regression



Dataset $(y_\mu, \xi_\mu)_{\mu=1,\ldots,P}$

The rule: $y = m\xi + q$

Find $m,\ q$ that fit at best your dataset

<u>Empirical Risk Minimization</u>

Empirical Risk $$H[m, q] = \frac{1}{2}\sum_{\mu}^{P}(y_\mu - m\xi_\mu - q)^2$$

Minimize H via ***gradient descent*** $$\dot{m} = -\frac{\partial H}{\partial m} \qquad \dot{q} = -\frac{\partial H}{\partial q}$$

This is a low dimensional (***underparametrized***) problem (many data, few parameters).

High dimension: Deep Learning

$f($  $) =$ Cat / Dog

Dataset $(y_\mu, \underline{\xi}_\mu)_{\mu=1,\ldots,P}$ $\underline{\xi}_\mu \in \mathbb{R}^d$

The rule: $y_\mu = f(\underline{\xi}_\mu, \underline{w})$

Empirical Risk Minimization

$$H[\underline{w}] = \frac{1}{2}\sum_{\mu}^{P}(y_\mu - f(\underline{\xi}_\mu, \underline{w}))^2$$

Gradient Descent

$$\dot{\underline{w}} = -\frac{\partial H}{\partial \underline{w}}$$

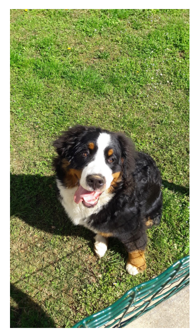
High dimension

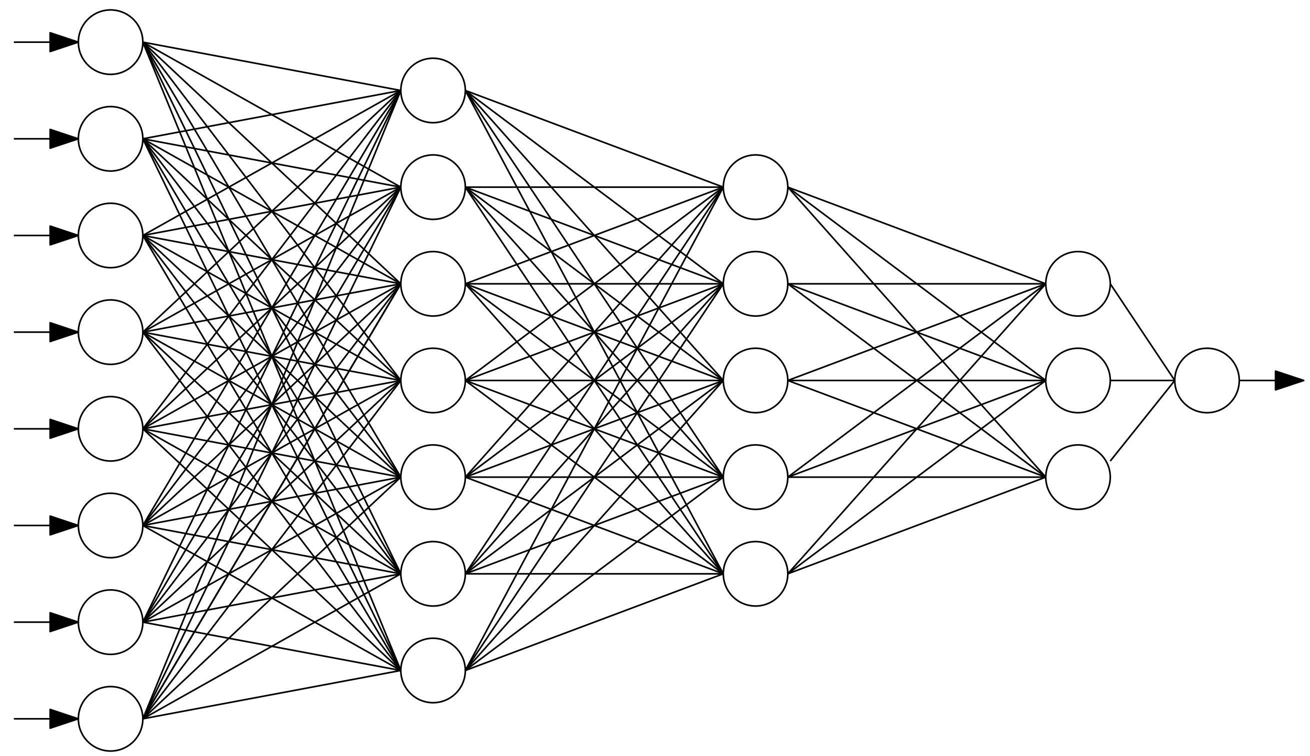
$d \gg 1$ Complex data structures

$P \gg 1$ Big datasets

$\dim(\underline{w}) = N \gg 1$ Huge number of fitting parameters

A computational problem





$$H[\underline{w}] = \frac{1}{2}\sum_{\mu}^{P}(y_\mu - f(\underline{\xi}_\mu, \underline{w}))^2 \qquad H[\underline{w}] = \sum_{\mu}^{P} v_\mu(\underline{w})$$

$$\dot{\underline{w}} = -\frac{\partial H}{\partial \underline{w}} = -\sum_{\mu=1}^{P}\frac{\partial v_\mu(\underline{w})}{\partial \underline{w}}$$

1. Each term of the sum is costly to compute: inevitable
2. **<u>We need to perform a huge sum over the dataset</u>**

Stochastic gradient descent

$$\dot{\underline{w}} = -\frac{\partial H}{\partial \underline{w}} = -\sum_{\mu=1}^{P} \frac{\partial v_\mu(\underline{w})}{\partial \underline{w}}$$

Partition of the dataset in minibatches

$$\sum_{\mu=1}^{P} \frac{\partial v_\mu(\underline{w})}{\partial \underline{w}} \to \sum_{\mu \in \mathscr{B}(t)} \frac{\partial v_\mu(\underline{w})}{\partial \underline{w}}$$

Minibatches are shuffled at random and proposed during training at random.

SGD is a noisy algorithm.
There is "information flow" during the dynamics

In deep learning, architectures and tasks change.

However: all of them are trained with stochastic gradient descent & **it works!!!**
Unexpectedly…

Questions: why SGD works? is SGD noise helpful for optimization?

Understanding SGD

Understanding SGD is a crucial part of the program aiming at understanding Deep Learning

1. *How does SGD explore the loss landscape (=Empirical Risk)?*
2. *Is the SGD noise useful for optimization? To what extent?*
3. *How much SGD is similar to Langevin/Gradient descent?*
4. *…*

This talk: focus on the algorithm.

✓ Develop DMFT to study the performances of SGD in a *prototypical* **hard** high-d optimization problem:

✗ Missing: interplay with the architecture/task/data structure

Statistical Physics of Learning

The space of interactions in neural network models 1987

E Gardner

Department of Physics, Edinburgh University, Mayfield Road, Edinburgh EH9 3JK, UK

Optimal storage properties of neural network models 1988

E Gardner† and B Derrida‡

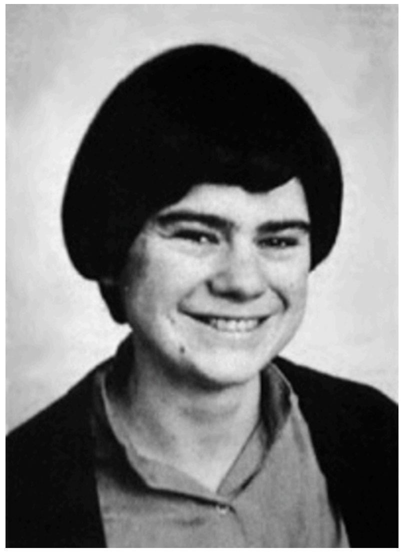
† Department of Physics, Edinburgh University, Mayfield Road, Edinburgh, EH9 3JZ, UK
‡ Service de Physique Theorique, CEN Saclay, F 91191 Gif sur Yvette, France

EG thanks the Service de Physique Theorique for their hospitality whilst in Saclay.

Three unfinished works on the optimal storage capacity of networks 1989

E Gardner and B Derrida

The Institute for Advanced Studies, The Hebrew University of Jerusalem, Jerusalem, Israel
and Service de Physique Théorique de Saclay†, F-91191 Gif-sur-Yvette Cedex, France



A teacher-student model

Still following Gardner and Derrida…

$\underline{x}^* = \{x_1^*, \ldots, x_N^*\}$	$\lvert \underline{x}^* \rvert^2 = N$	Signal/ground truth
$J_{ij}^\mu \sim \mathcal{N}(0,1)$	$\mu = 1,\ldots,\alpha N$ $i,j = 1,\ldots,N$	Randomness
$y_\mu = \frac{1}{N} \sum_{i<j} J_{ij}^\mu x_i^* x_j^*$	$\mu = 1,\ldots,\alpha N$	Labels: *The Rule*
$\{y_\mu, J^\mu\}_{\mu=1,\ldots,\alpha N}$		The dataset

Can we recover $\underline{x}^*$ given the dataset ***and*** knowing *the structure* of the rule?

We want to study the high-dimensional (= thermodynamic) limit

$$N \to \infty$$

Empirical Loss (The Hamiltonian)

$$H = \frac{1}{2}\sum_{\mu}\left(y_\mu - \frac{1}{N}\sum_{i<j} J^{\mu}_{ij} x_i x_j\right)^2$$

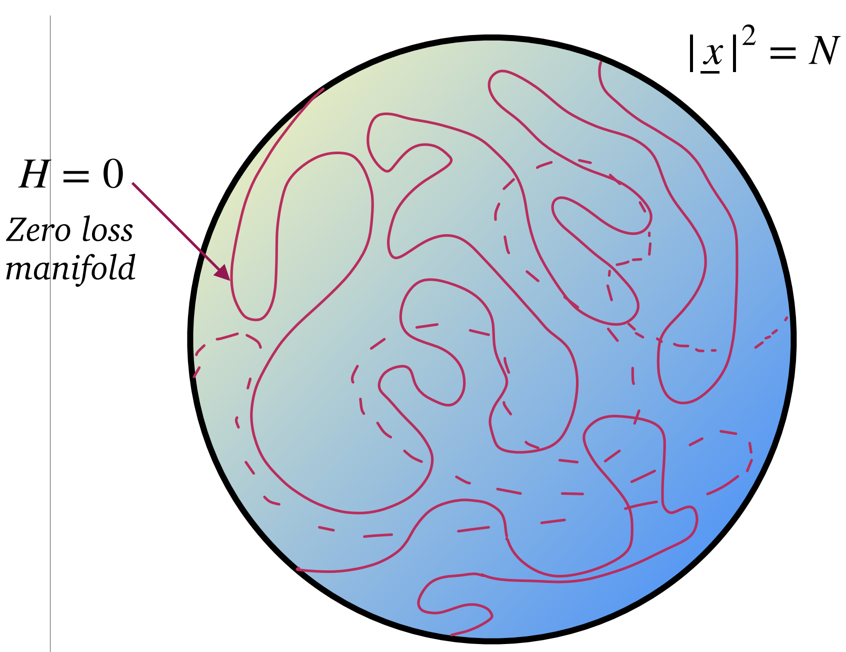
Empirical Risk = Empirical Loss
= the Hamiltonian

This is an (i) high-dimensional, (ii) non-convex, loss function

There are two regimes

The *overparametrized* regime $\alpha < 1$

Canyon landscape

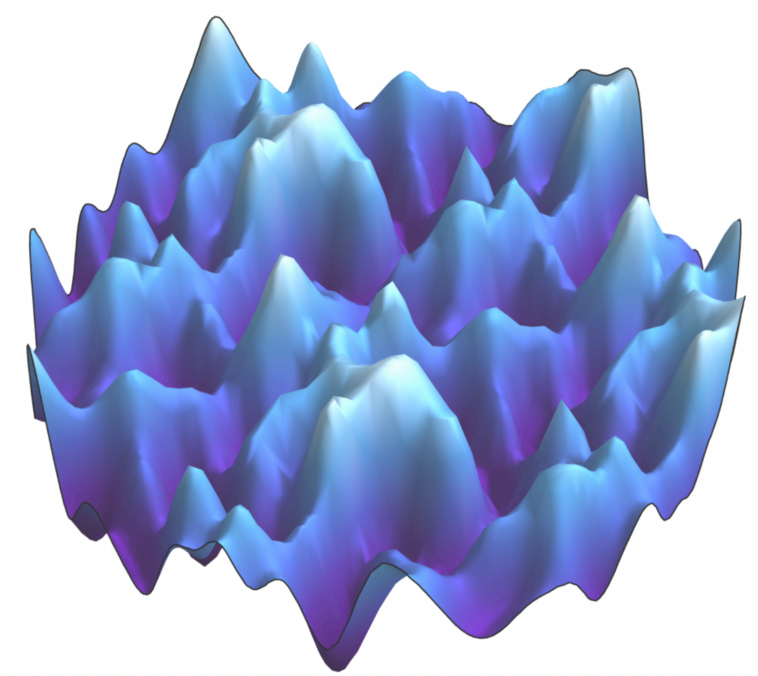


Relevant for Deep Learning

Interpolation

The *underparametrized* regime $\alpha > 1$

Rough landscape



Provides a prototypical hard high-d optimization problem where to benchmark algorithms

Empirical Risk

$$H = \frac{1}{2}\sum_{\mu}\left(y_\mu - \frac{1}{N}\sum_{i<j} J_{ij}^{\mu} x_i x_j\right)^2$$

We now focus on the ***underparametrized phase***: $\alpha > 1$

Two settings:

1. Thermodynamics

   Find the ground state of the loss (***Fyodorov 2018***).
   - the loss has only two global minima at zero energy.
   - The two minima are "Replica Symmetric"

2. Dynamics: ***minimize the loss via SGD.*** We expect
   - Hard high-d optimization problem.
   - Generated by a glassy landscape.
   - Only two good minima at zero energy => perfect generalization.

The landscape structure is an open problem

SGD minimization

Kamali, Urbani, arXiv:2306.06420
Kamali, Urbani, arXiv:2309.04788

$$H = \sum_\mu v_\mu(\underline{x}) \qquad v_\mu(\underline{x}) = \frac{1}{2}\left(y_\mu - \frac{1}{N}\sum_{i<j} J_{ij}^\mu x_i x_j\right)^2$$

1. Gradient Descent

$$x_i(t+1) = x_i(t) - \eta \frac{\partial H}{\partial x_i} = x_i(t) - \eta \sum_\mu \frac{\partial v_\mu}{\partial x_i}$$

2. Stochastic Gradient Descent

$$x_i(t+1) = x_i(t) - \frac{\eta}{b} \sum_\mu s_\mu(t) \frac{\partial v_\mu}{\partial x_i}$$

$$s_\mu(t) = \begin{cases} 0 & \text{with prob. } 1-b \\ 1 & \text{with prob. } b \end{cases}$$

Selection variables

Batch size $= b\alpha N$

This is a discrete algorithm and does not have a continuous time limit

Dynamical mean field theory

To study dynamics one can use *path integrals.*

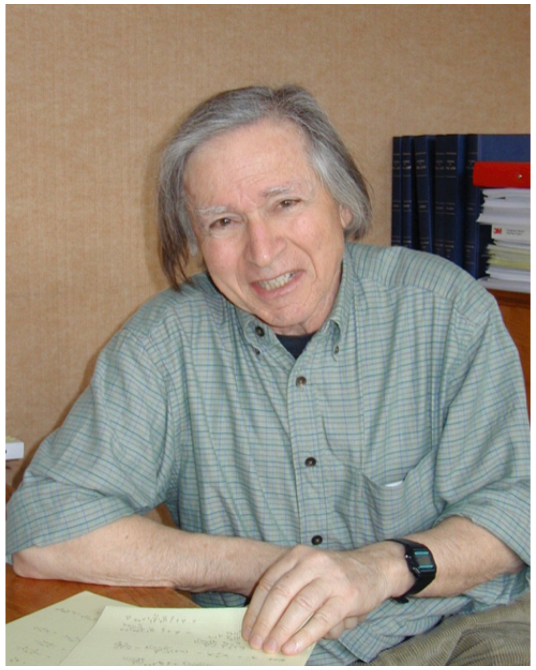
This technique takes the name of the
Martin-Siggia-Rose-Jannsen-De Dominicis formalism

TECHNIQUES DE RENORMALISATION
DE LA THÉORIE DES CHAMPS
ET DYNAMIQUE DES PHÉNOMÈNES CRITIQUES

C. DE DOMINICIS
Service de Physique Théorique, CEN, Saclay, BP nº 2, 91190 Gif-sur-Yvette, France

Résumé. — La dynamique des phénomènes critiques telle qu'elle est décrite par les équations stochastiques de type Ginzburg-Landau dépendant du temps, avec ou sans loi de conservation, est étudiée par les techniques de renormalisation de la théorie des champs.
Le cas des systèmes comportant un couplage mode-mode est brièvement abordé.

Abstract. — The dynamics of critical phenomena as is described by stochastic equations of the Landau-Ginzburg type with or without conservation law, is studied by the technique of field renormalization.
The case of mode coupling systems is briefly touched upon.



PHYSICAL REVIEW B **VOLUME 18, NUMBER 9** **1 NOVEMBER 1978**

Dynamics as a substitute for replicas in systems with quenched random impurities

C. De Dominicis*
Lyman Laboratory of Physics, Harvard University, Cambridge, Massachusetts 02138
*Permanent address: Service de Physique Theorique, CEA, CEN Saclay, BP2, Gif Sur Yvette, France.

Dynamical order parameters

$$C(t,t') = \frac{1}{N}\sum_i x_i(t)x_i(t')$$

Correlation function

$$R(t,t') = \frac{1}{N}\sum_i \frac{\delta x_i(t)}{\delta H_i(t')}$$

Response function

$$m(t) = \frac{1}{N}\sum_i x_i(t)x_i^*$$

Magnetization

$$\Delta(t) = \frac{1}{N}\sum_i (x_i(t) - x_i{}^*)^2 = 1 - 2m(t) + C(t,t)$$

The mean square displacement is a measure
of the distance from the true signal

Dynamical mean field theory

$$m(t+1)=m(t)-\eta^2\alpha\left(\sum_{s=0}^{t}(\Lambda_R(t,s)C(t,s)+\Lambda_C(t,s)R(t,s))m(s)-m(t)\sum_{s=0}^{t}\Lambda_R(t,s)\right)$$

$$C(t+1,t')=C(t,t')+\eta\Omega_1(t,t') \qquad \forall t'\leq t$$

$$R(t+1,t')=\delta_{t,t'}-\eta^2\alpha\sum_{s=t'+1}^{t}(\Lambda_R(t,s)C(t,s)+\Lambda_C(t,s)R(t,s))R(s,t')$$

$$C(t+1,t+1)=C(t,t)+2\eta\Omega_1(t,t)+\eta^2\Omega_2(t)$$

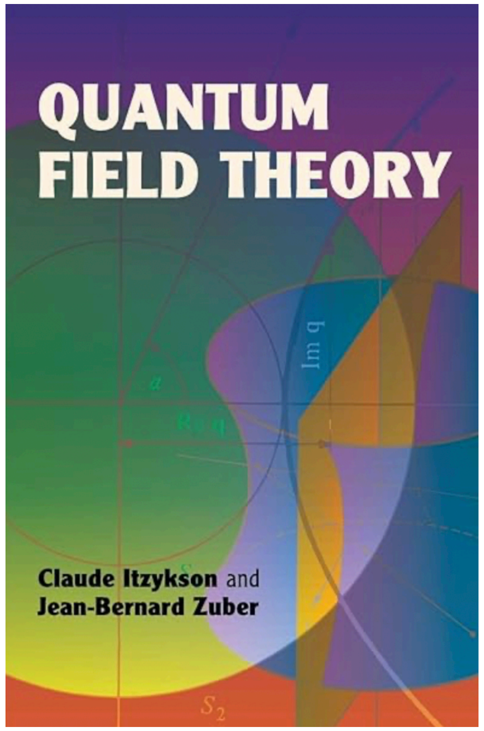
$$\Omega_1(t,t')=\alpha\eta\left[m(t)m(t')\sum_{s=0}^{t}\Lambda_R(t,s)-\sum_{s=0}^{t'}\Lambda_C(t,s)C(t,s)R(t',s)-\sum_{s=0}^{t}(\Lambda_R(t,s)C(t,s)+\Lambda_C(t,s)R(t,s))C(t',s)\right]$$

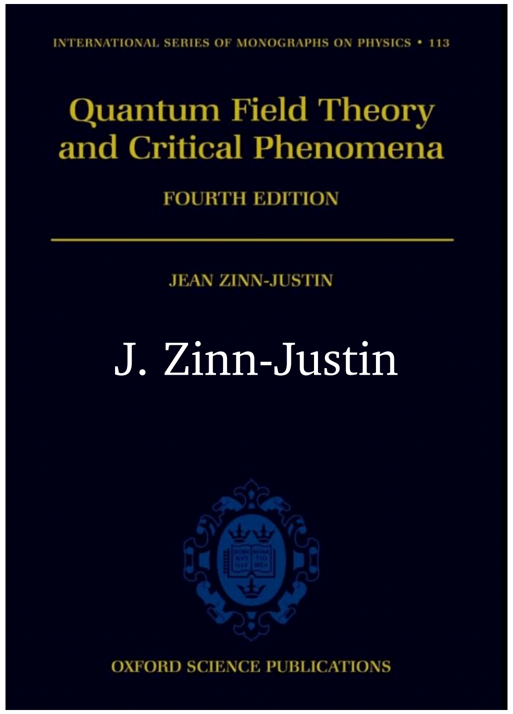
$$\begin{aligned}\Omega_2(t)=&\alpha^2\eta^2\sum_{s,s'=0}^{t}(\Lambda_R(t,s)C(t,s)+\Lambda_C(t,s)R(t,s))C(s,s')(\Lambda_R(t,s)C(t,s')+\Lambda_C(t,s)R(t,s'))\\&-2\alpha^2\eta^2m(t)\left(\sum_{s=0}^{t}\Lambda_R(t,s)\right)\left(\sum_{s=0}^{t}(\Lambda_R(t,s)C(t,s)+\Lambda_C(t,s)R(t,s))m(s)\right)\\&+2\alpha^2\eta^2\sum_{s=0}^{t}\sum_{s'=0}^{s}(\Lambda_R(t,s)C(t,s)+\Lambda_C(t,s)R(t,s))\Lambda_C(t,s')C(t,s')R(s,s')\\&-\alpha\Lambda_C(t,t)C(t,t)+\left(\alpha\eta\sum_{s=0}^{t}\Lambda_R(t,s)\right)^2\end{aligned}$$

Seem complicated but actually can be integrated very efficiently

Dynamical mean field theory

SUSY formalism





Superfields

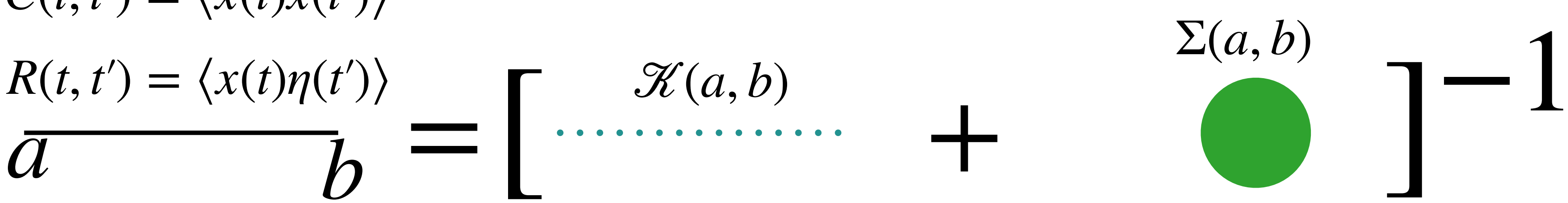
$$Q(a,b) = \langle x(a)x(b)\rangle$$
$$= C(t_a,t_b) + \theta_a R(t_b,t_a) + \theta_b R(t_b,t_a)$$

$$\color{red}{Q^{-1}(a,b) = \mathscr{K}(a,b) + \Sigma(a,b)}$$

<u>Dyson Equation</u>

$$C(t,t') = \langle x(t)x(t')\rangle$$

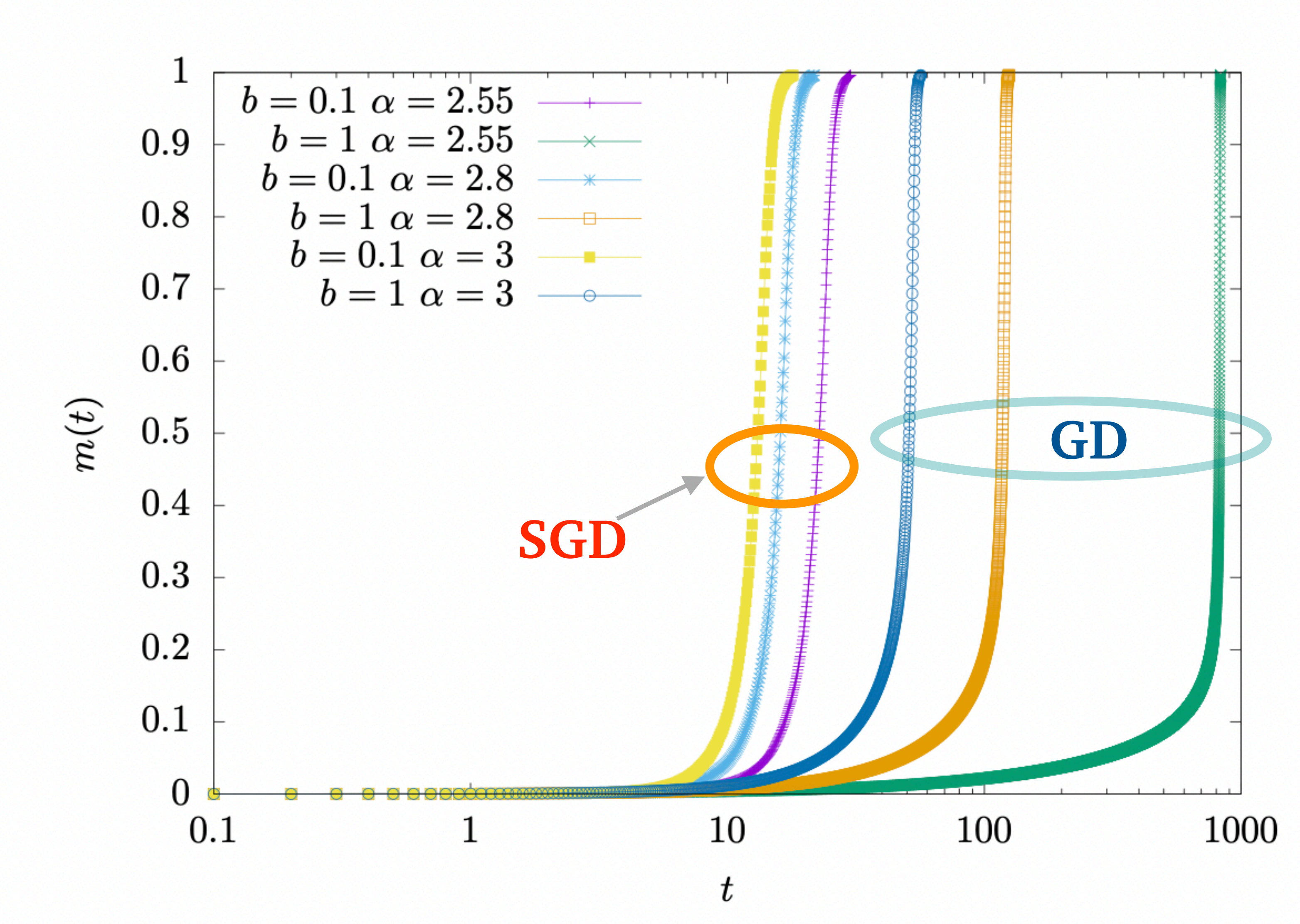
$$R(t,t') = \langle x(t)\eta(t')\rangle$$



The dynamical version has a *causal structure*
(unless one computes large deviations = instantons)

The coupling constant of the theory is the *sample complexity*.

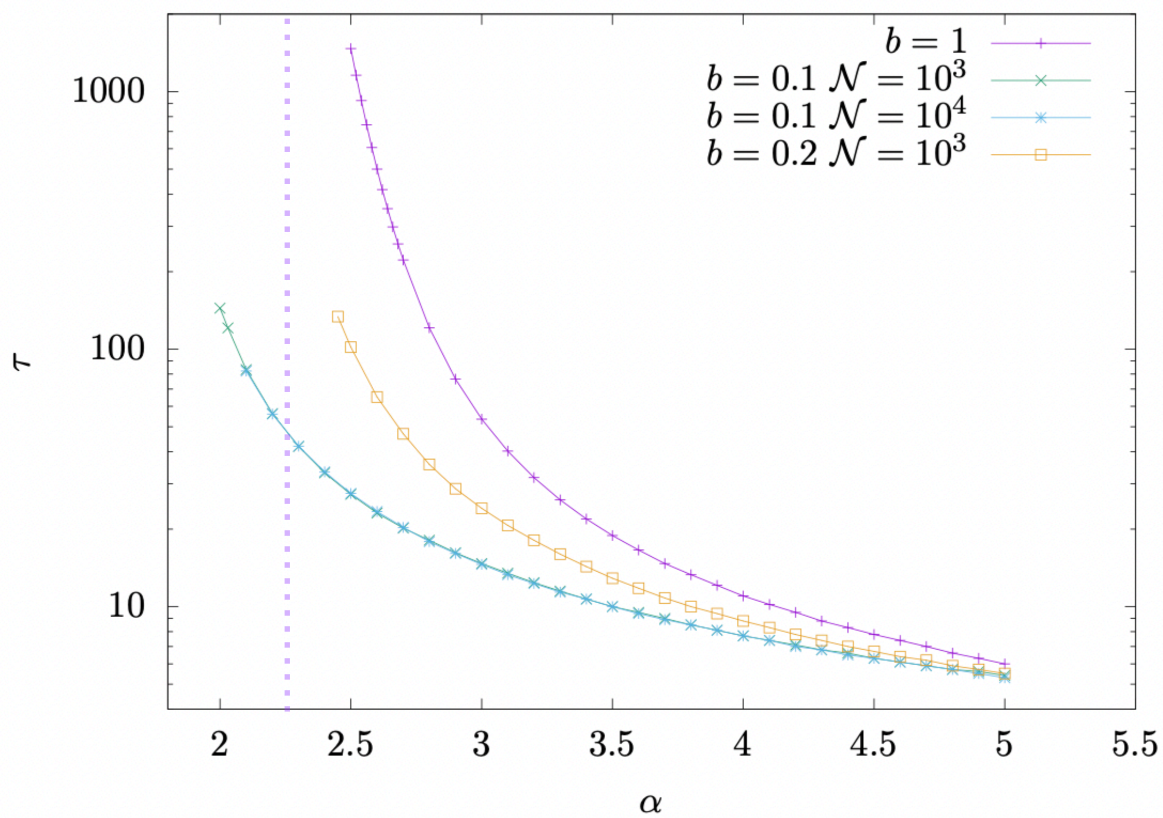
Results

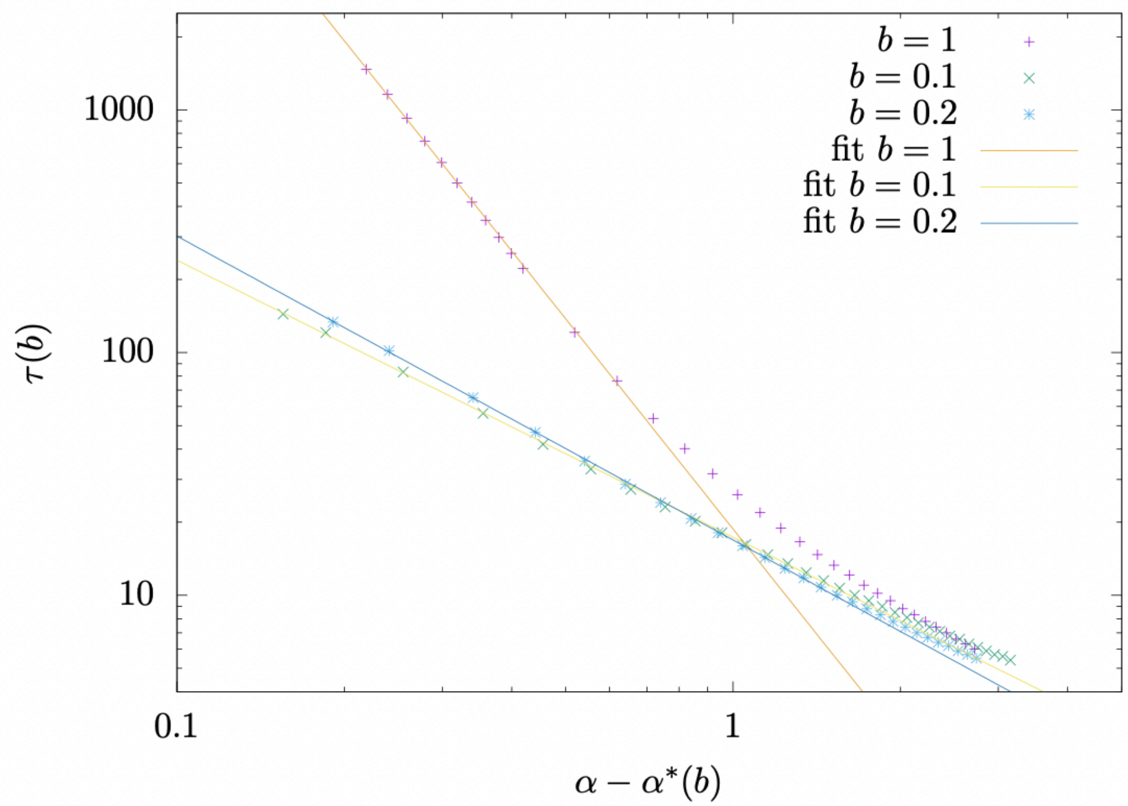


SGD is *faster* than GD.

Is it actually ***better***? = Does SGD recover the signal at smaller sample complexity than GD?

Results





Fit the relaxation time via power law $\tau(\alpha) \simeq \tau_0 |\alpha - \alpha^*(b)|^{-\nu}$

For GD $\alpha^*(1) \simeq 2.27$

SGD is has a different and ***smaller*** <u>recovery threshold</u> than GD.

Conclusions

- A theory of SGD can be developed and we just started (Plenty of questions still unanswered. We constructed mainly the tools).

- **We can establish that SGD is significantly better than GD.**

- A theory for the recovery threshold is possible: we need to have better understanding of the statistics of asymptotic configurations visited by SGD.

- SGD is a **non-equilibrium** algorithm. It *drives* the system preventing fully relaxational dynamics.

  => the asymptotic behavior is a *non-equilibrium stationary* state with interrupted aging.
  => One needs to develop a *macroscopic fluctuation theory* around the typical trajectory.

The statistical physics approach to learning and optimization is seeing a revival interest and it is shaping modern high-dimensional statistics and the theory of deep learning.