Hidden Thermodynamics in Deep Networks

Dmitry Vetrov
Professor of Computer Science in Constructor University, Bremen
Head of BayesGroup

CONSTRUCTOR UNIVERSITY

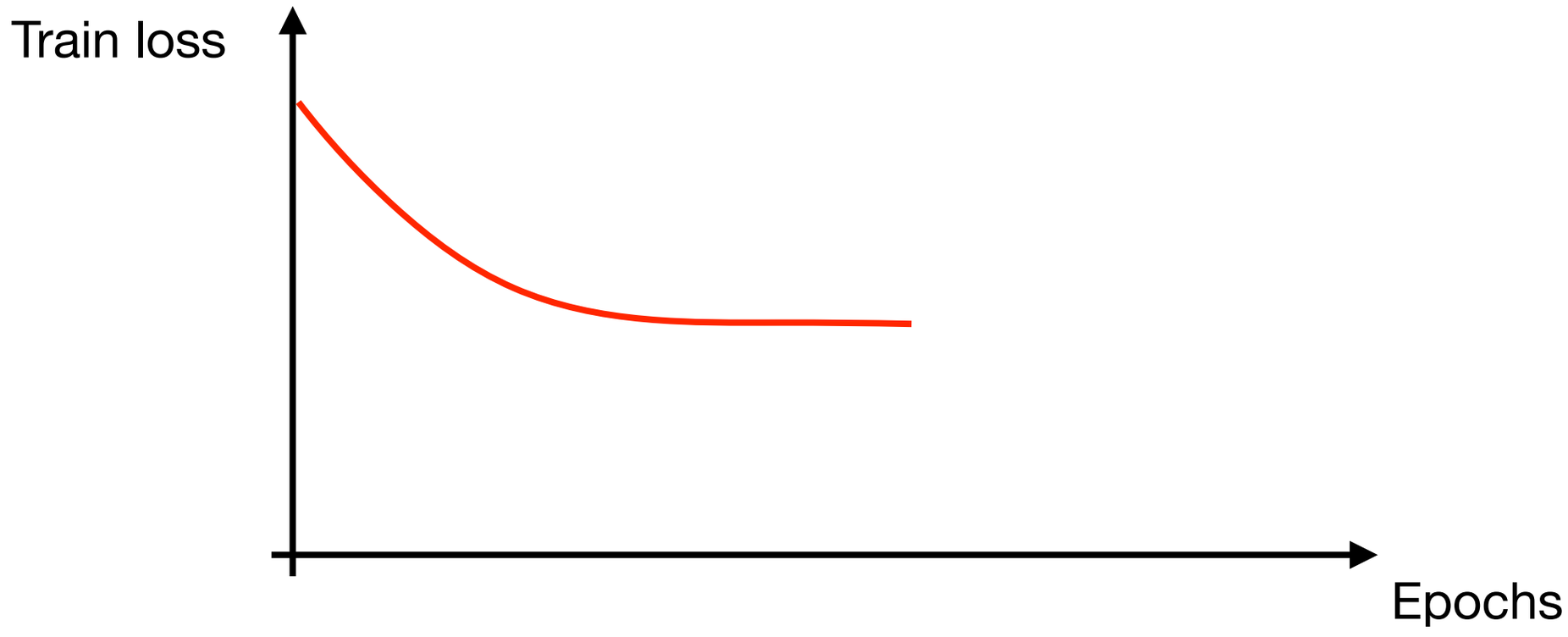
p(B|A)yesgroup

Motivation

- Lack of interpretability in modern DNN
- Bridges to other better-studied disciplines is needed
- **Main result:** training process in DNN has much in common with thermodynamical processes

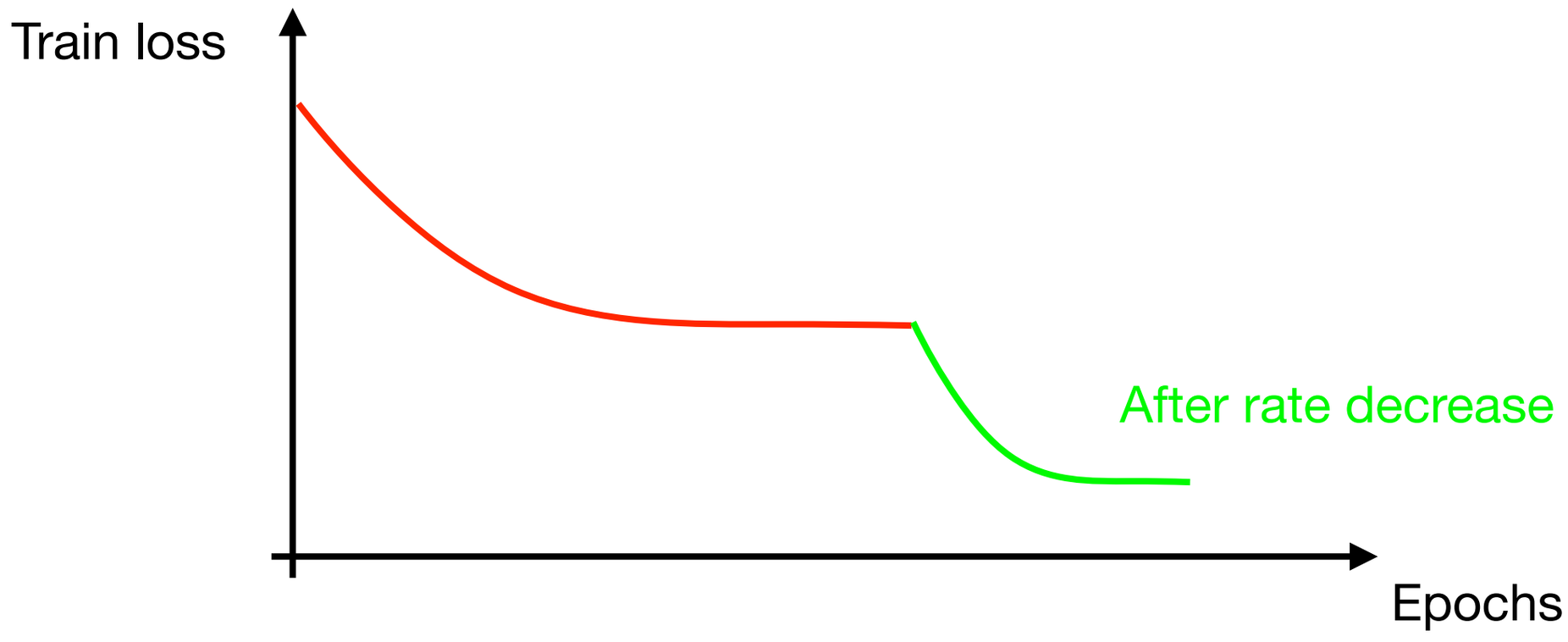
Motivation

- When we train DNN with SGD with constant learning rate our loss stabilizes but not in its minimum



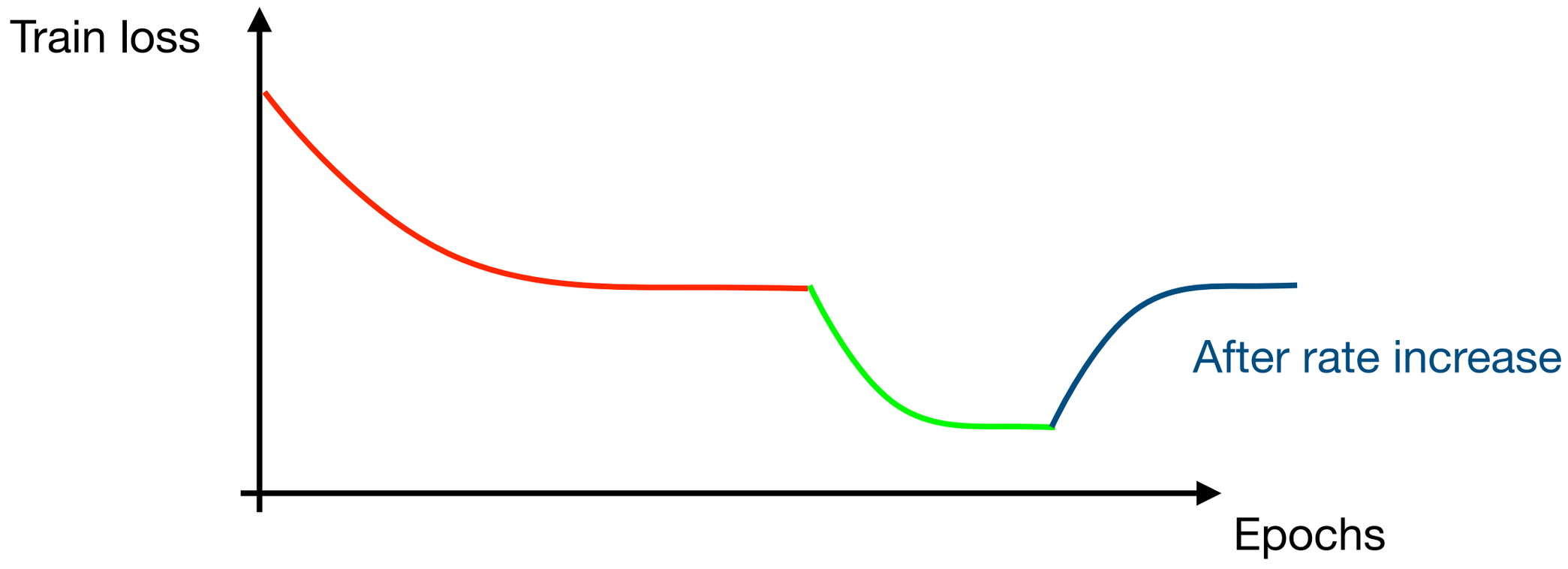
Motivation

- When we train DNN with SGD with constant learning rate our loss stabilizes but not in its minimum



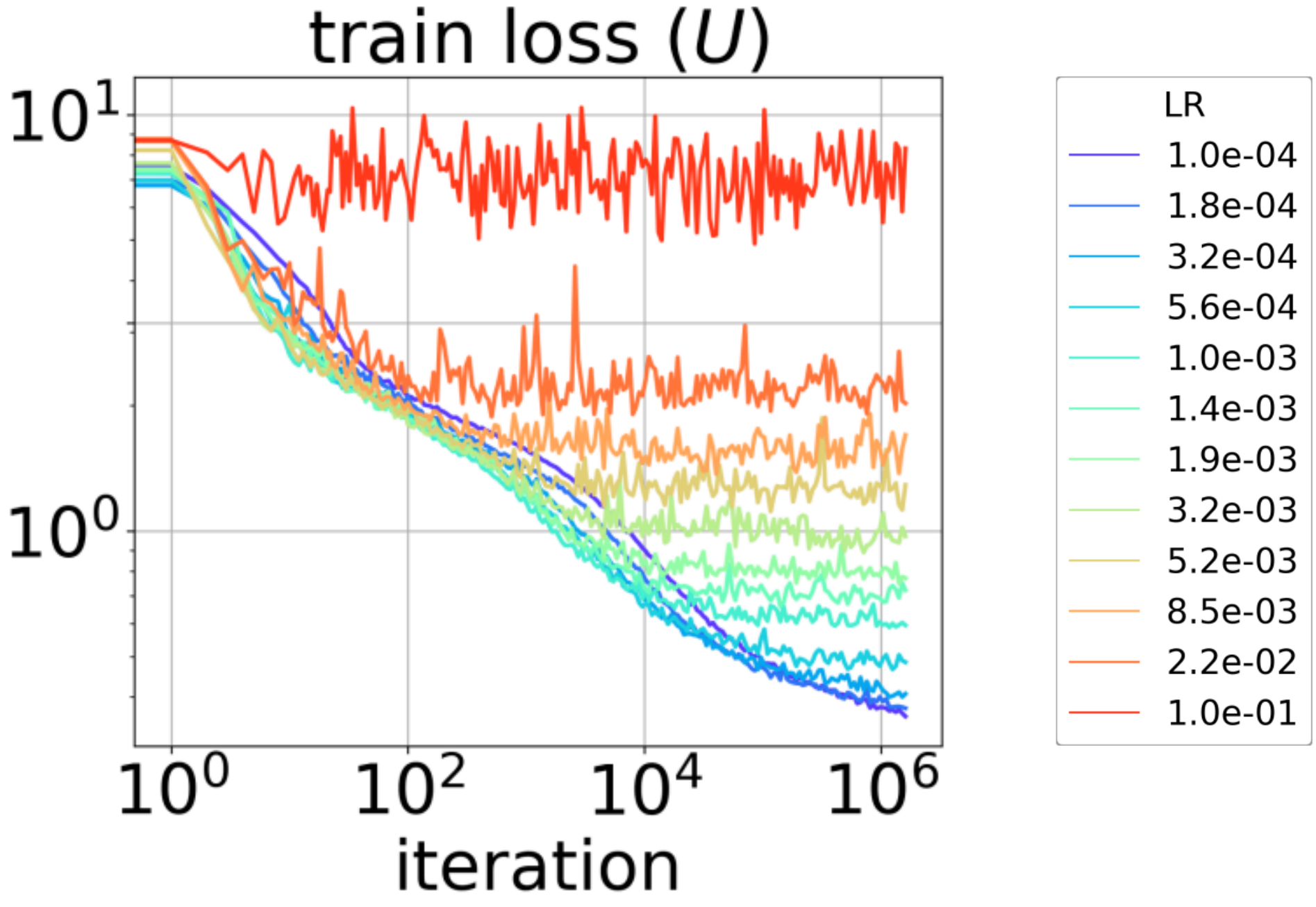
Motivation

- When we train DNN with SGD with constant learning rate our loss stabilizes but not in its minimum



- Hypothesis: maybe SGD with fixed rate minimizes something different rather than train loss?

Experiment


train loss (U)
10^1
10^0
10^0
10^2
10^4
10^6
iteration
LR
1.0e-04
1.8e-04
3.2e-04
5.6e-04
1.0e-03
1.4e-03
1.9e-03
3.2e-03
5.2e-03
8.5e-03
2.2e-02
1.0e-01

Outline

- Recap of thermodynamics
- Stochastic optimization of scale-invariant functions
- Different equilibriums in DNNs

Thermodynamics

- Thermodynamics studies the behaviour of the system of particles under thermal fluctuations

- Particles are described by (microsopic) variables z that define the total energy $U(z)$

- Microscopic variables include position and velocity of each particle

- Particles [in gas] are chaotically moving (fluctuating) given non-zero temperature

- In equilibrium state the distribution $\rho(z)$ does not change with time

Macroscopic variables in TD

- Parameters of individual molecules z are hard to measure
- Due to their chaotic oscillations we usually deal with distribution $\rho(z)$
- Thermodynamics deals with measurable (macroscopic) variables instead
- Such variables are:
  - Mean Internal energy $\bar{U}_\rho = \mathbb{E}_\rho U(z)$
  - Temperature T
  - Volume V
  - Pressure p
  - Entropy $S = -\mathbb{E}_\rho \log \rho(z)$

Ideal gas law

- For ideal gas the following law holds

$$\frac{pV}{T} = 1$$

- It was discovered empirically for real gases under normal conditions and later justified for model of ideal gas

First law of thermodynamics

- The internal energy (of gas) is the total energy (kinetic plus potential) of all of its molecules

- When temperature is non-zero the molecules are moving chaotically

- The internal energy can be changed either by heating or by doing work

$$d\bar{U}_\rho = dQ - dA$$

Heat **Work**

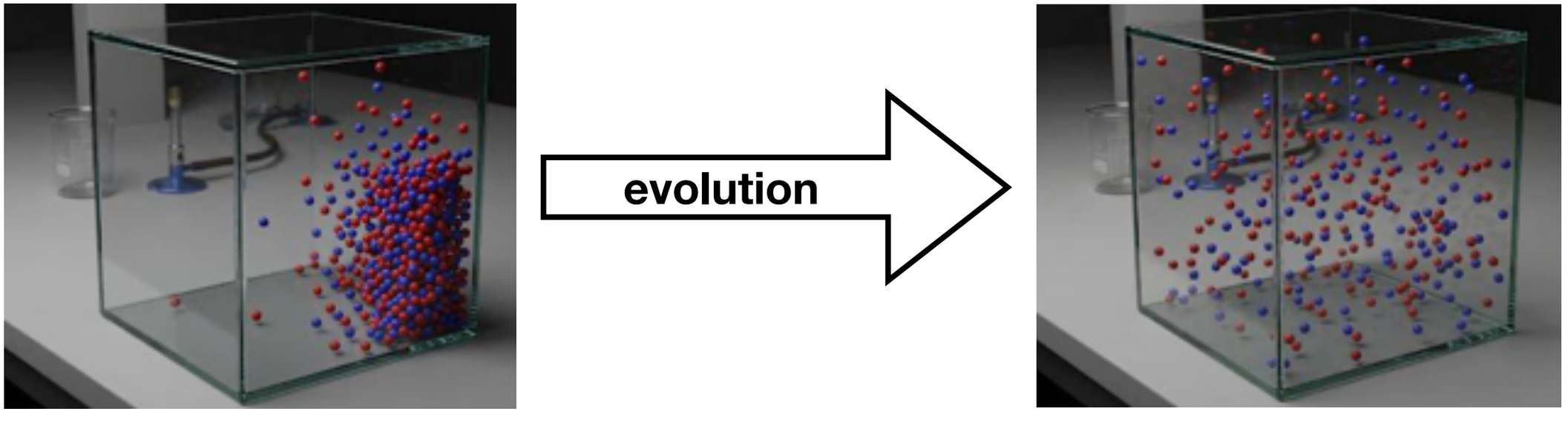
- The work is done by changing the volume of the system under given pressure

$$dA = pdV$$

Evolution of TD system

- Given some initial $\rho_0(z)$ system evolves until it reaches an stationary (equilibrium) state

- In equilibrium state the distribution $\rho(z)$ does not change in time



Non-stationary state

Stationary state

Second law of thermodynamics

- Second law of thermodynamics states that

$$dS \geq \frac{dQ}{T}$$

- In particular if there is no heat exchange the entropy cannot decrease $dS \geq 0$

- If the system is in equilibrium $dQ = TdS$

- By combining it with first law we get

$$d\bar{U}_{\rho} \leq TdS - pdV$$

Internal energy revisited

- From the first two laws of thermodynamics we have

$$d\bar{U}_\rho \leq TdS - pdV$$

- Now suppose that we consider the evolution of the system under constant S and V, i.e. $dS=0$, $dV=0$

- Then among all distributions with fixed volume and entropy the system evolves to the minimum internal energy

$$d\bar{U}_\rho = 0$$

Free energy

- What happens if we fix other macroscopic variables?
- Let us fix T and V
- Consider Helmholtz free energy $F = \bar{U}_\rho - TS$

Free energy

- What happens if we fix other macroscopic variables?
- Let us fix T and V
- Consider Helmholtz free energy $F = \bar{U}_\rho - TS$
- Its differential is

$$dF = d\bar{U}_\rho - d(TS) = dU - TdS - SdT$$

Free energy

- What happens if we fix other macroscopic variables?
- Let us fix T and V
- Consider Helmholtz free energy $F = \bar{U}_\rho - TS$
- Its differential is

$$dF = d\bar{U}_\rho - d(TS) = dU - TdS - SdT \leq$$

$$\leq TdS - pdV - TdS - SdT = -SdT - pdV$$

- This means that when $dT=0$ and $dV=0$ the free energy should be at extremal value (minimum) w.r.t. $\rho(z)$

Free energy

- When temperature T and volume V are fixed the system evolves to equilibrium

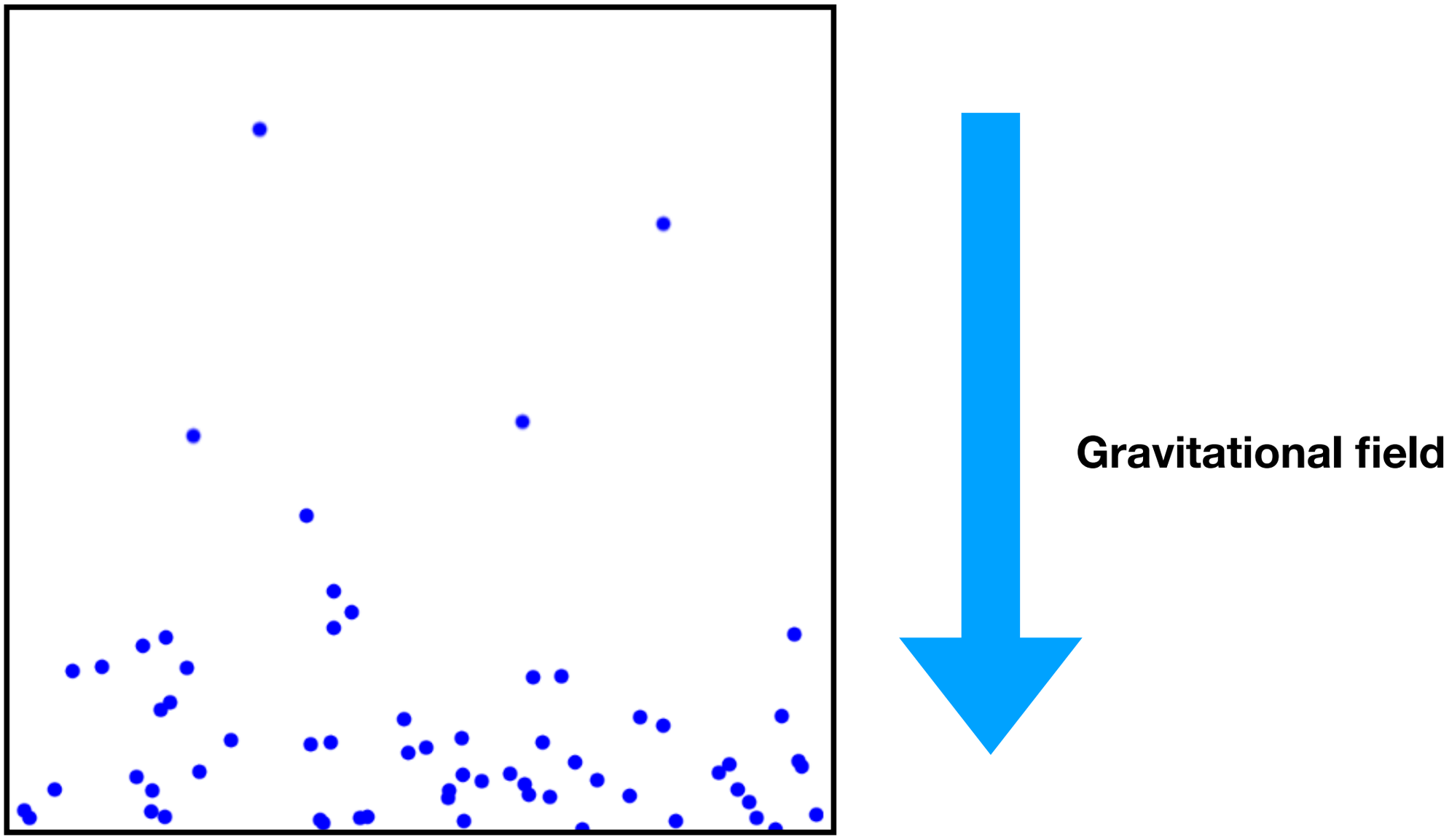
$$F = \bar{U}_\rho - TS \to \min_\rho$$

- It is not difficult to derive the expression for stationary distribution

$$\rho(z) = \frac{1}{C} \exp\left(-\frac{U(z)}{T}\right)$$

- This is Gibbs distribution

Gibbs distribution



Gibbs potential

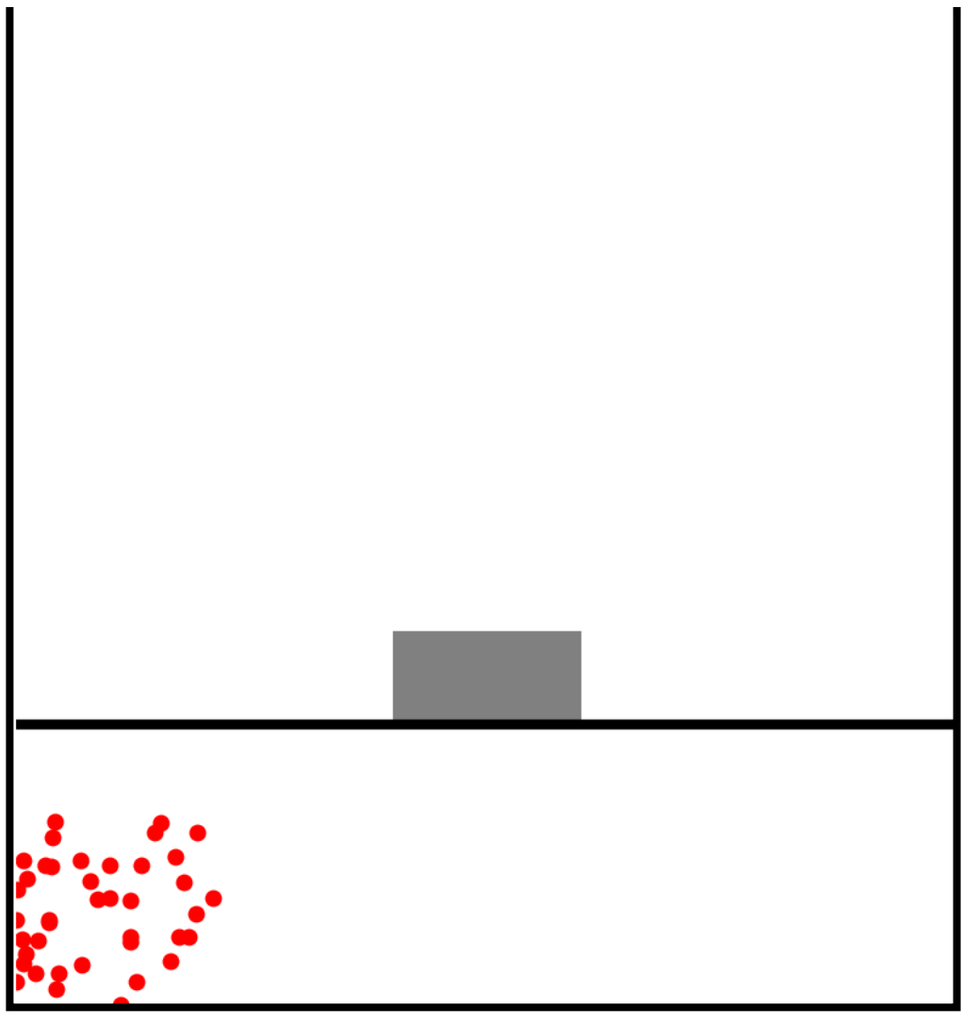
- Similarly we can consider the case when we fix T and p
- Define Gibbs potential as follows $G = \bar{U}_\rho - TS + pV$
- Then its differential is given by

$$dG = d\bar{U}_\rho - d(TS) + d(PV) \leq$$

$$\leq TdS - pdV - SdT - TdS + Vdp - pdV = -SdT + Vdp$$

- When $dT=0$ and $dp=0$ (constant temperature and pressure) system evolves to equilibrium state by minimizing G w.r.t. $\rho(z)$

Gibbs potential



Maxwell relations

- Consider, e.g. Gibbs potential as a function of two independent variables p and T

$$dG = -SdT + Vdp$$

- Then the derivatives are given by $\frac{dG}{dT} = -S, \quad \frac{dG}{dp} = V$

Maxwell relations

- Consider, e.g. Gibbs potential as a function of two independent variables p and T

$$dG = -SdT + Vdp$$

- Then the derivatives are given by $\frac{dG}{dT} = -S, \quad \frac{dG}{dp} = V$
- Since the second mixed derivatives must be equal

$$\frac{d^2G}{dTdp} = \frac{d^2G}{dpdT}$$

Maxwell relations

- Consider, e.g. Gibbs potential as a function of two independent variables p and T

$$dG = -SdT + Vdp$$

- Then the derivatives are given by $\frac{dG}{dT} = -S, \quad \frac{dG}{dp} = V$
- Since the second mixed derivatives must be equal

$$\frac{d^2G}{dTdp} = \frac{d^2G}{dpdT}$$

- We get

$$\left(\frac{\partial V}{\partial T}\right)_p = -\left(\frac{\partial S}{\partial p}\right)_T$$

Thermodynamics potentials

- By fixing different controllable variables we get different potentials that are being minimized w.r.t. $\rho(z)$ when system evolves to equilibrium state

- Mean Internal energy $\bar{U}_\rho(S, V)$

- Free energy $F(T, V) = \bar{U}_\rho - TS$

- Gibbs potential $G(T, p) = \bar{U}_\rho - TS + pV$

- Enthalpy $H(S, p) = \bar{U}_\rho + pV$

- Numerous Maxwell relations can be derived

Stochastic optimization in DNN

- Modern neural networks are trained via stochastic optimization with learning rate η

- In simplest form the update of the weights is given by

$$\theta_{t+1} = \theta_t - \eta(\nabla L(\theta_t) + \lambda\theta_t) + \xi, \quad \mathbb{E}\xi = 0, \quad \mathbb{D}\xi \sim \eta^2\sigma^2$$

Stochastic optimization in DNN

- Modern neural networks are trained via stochastic optimization with learning rate η

- In simplest form the update of the weights is given by

$$\theta_{t+1} = \theta_t - \eta(\nabla L(\theta_t) + \lambda\theta_t) + \xi, \quad \mathbb{E}\xi = 0, \quad \mathbb{D}\xi \sim \eta^2\sigma^2$$

- It can be shown that it corresponds to the discretization of the following stochastic differential equation (given η, σ fixed)

$$d\theta = -(\nabla L(\theta) + \lambda\theta)dt + \sqrt{\eta}\sigma dB_t$$

- Different η will correspond to different dynamics!

Scale-invariant DNN

- Modern DNNs contain normalization layers that make neural networks scale-invariant

$$L(\theta) = L(C\theta), \quad \forall C > 0$$

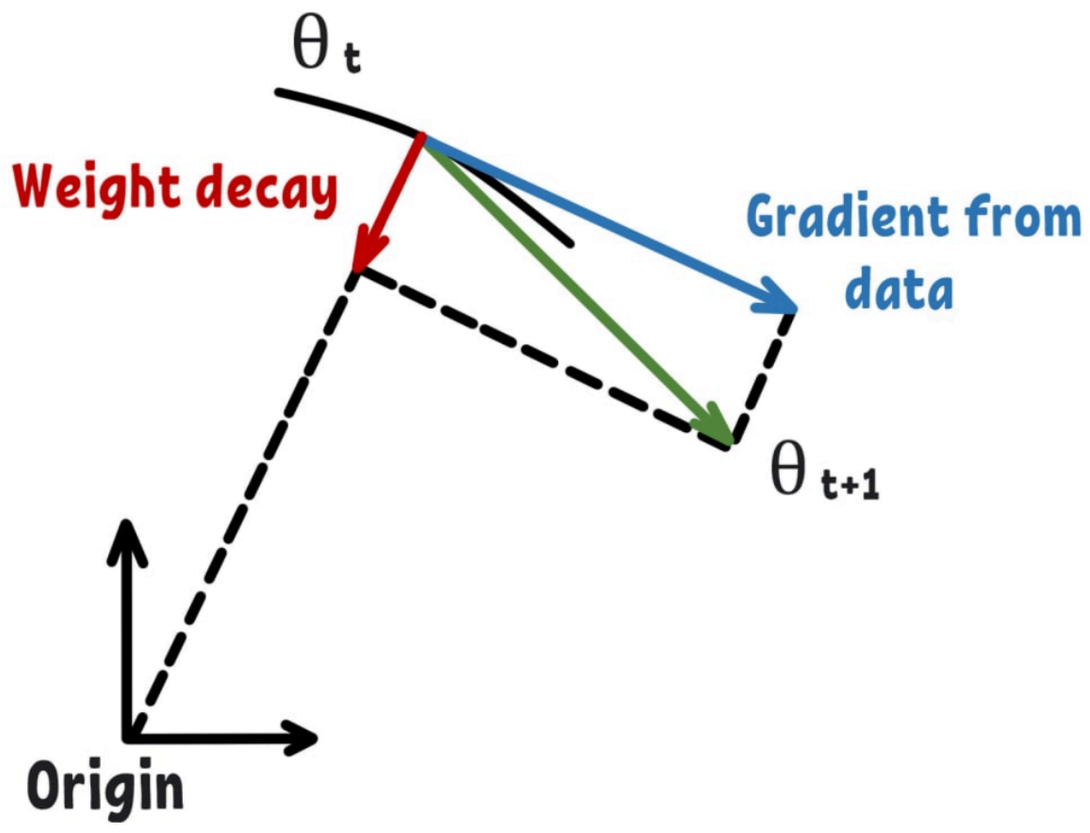
- In this case any point θ in the weight space has equivalent point on the unit sphere

$$u = \frac{\theta}{\|\theta\|}$$

Scale-invariant DNN

- Scale-invariant networks are constant along any beam emitted from the origin

- Hence the loss (stochastic) gradient is always orthogonal to the beam



Scale-invariant DNN

- We may switch to spherical coordinates

$$\theta = ru, \quad r = \|\theta\| \in \mathbb{R}, \quad u = \frac{\theta}{\|\theta\|} \in \mathbb{S}^d$$

- And write down the training dynamics for new variables

u – **Stochastic DE**

r – **Ordinary DE**

- Surprizingly the dynamics on r becomes deterministic!

- Under additional assumptions we may solve both equations

Effective LR

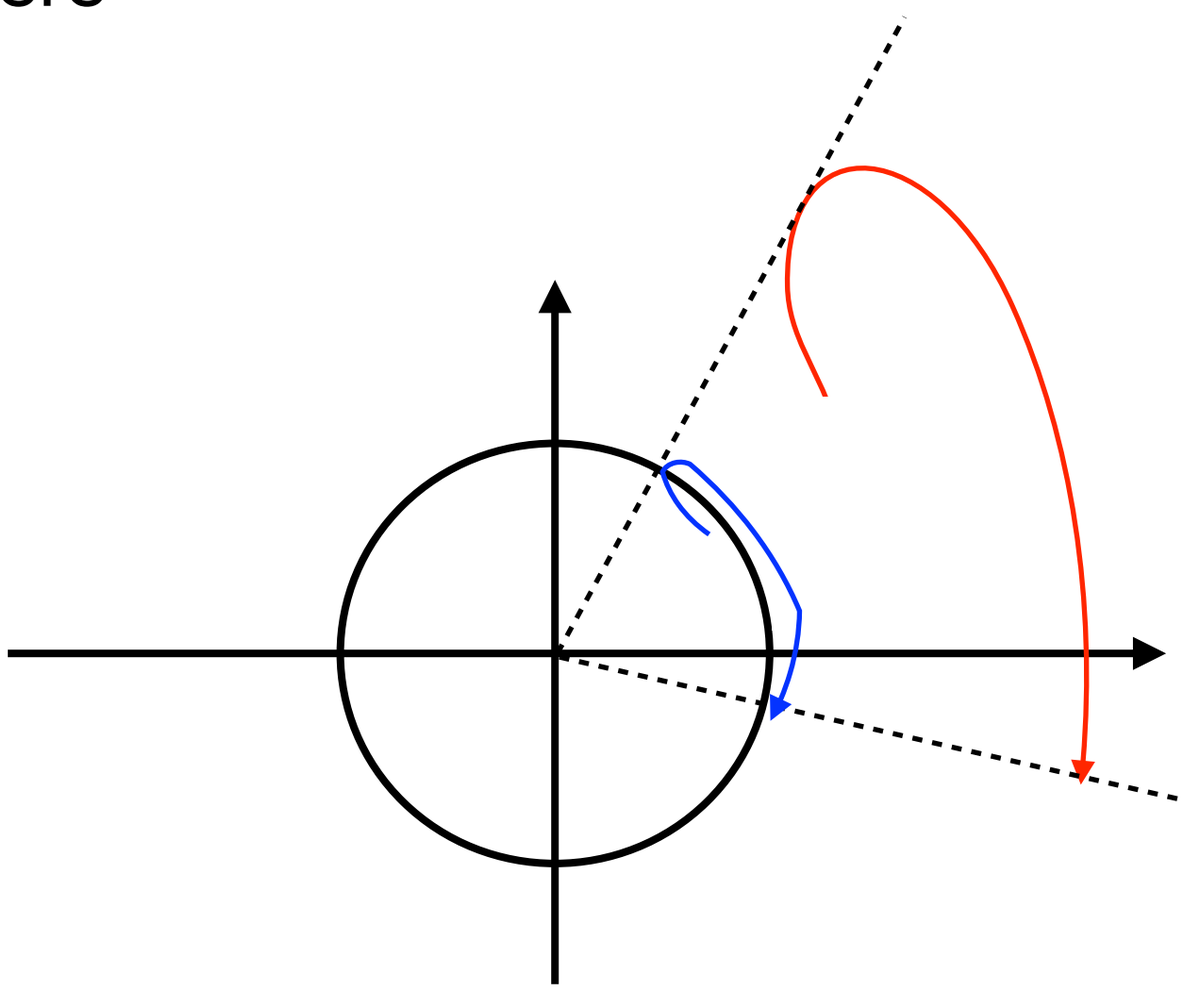
- Consider a simplistic training protocol first
- Here we fix so-called effective learning rate

$$\eta_{eff} = \frac{\eta}{\|\theta\|^2}$$

- That corresponds to fixing the resolution of our optimizer on a unit sphere

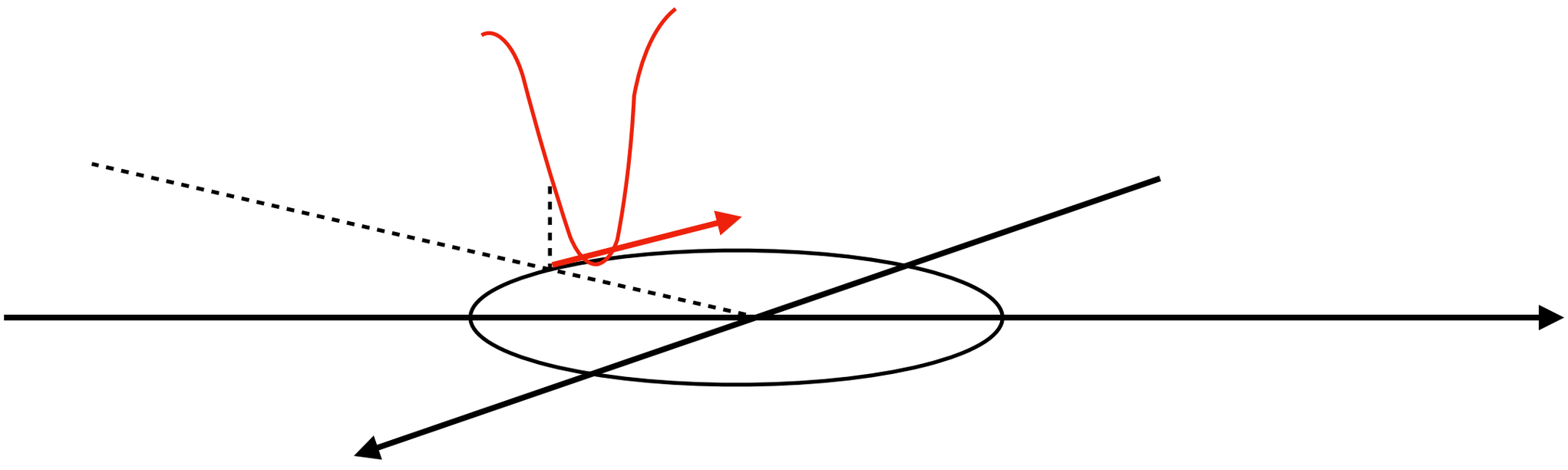
Effective LR

- Because of scale-invariance any optimization path in Euclidean space has an equivalent optimization path on a unit sphere



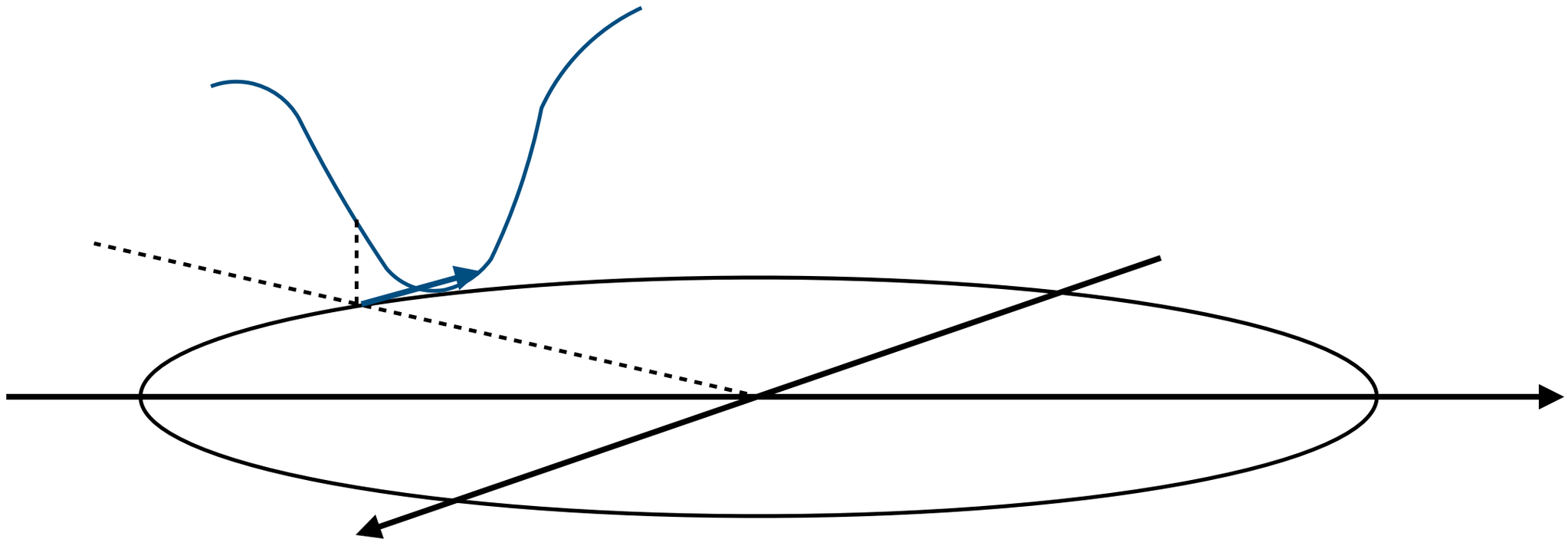
Effective LR

- The loss gradients become smaller with increase of weight norm



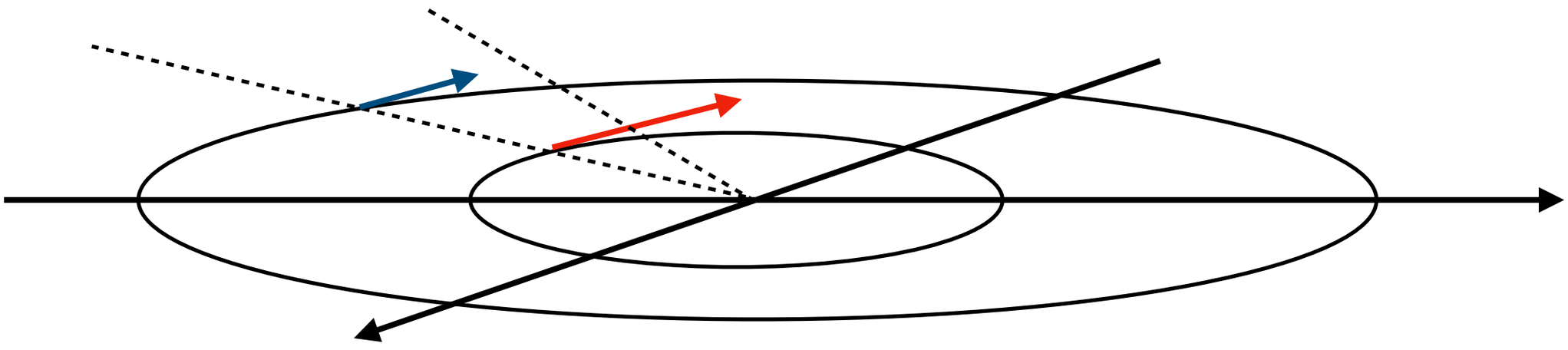
Effective LR

- The loss gradients become smaller with increase of weight norm



Effective LR

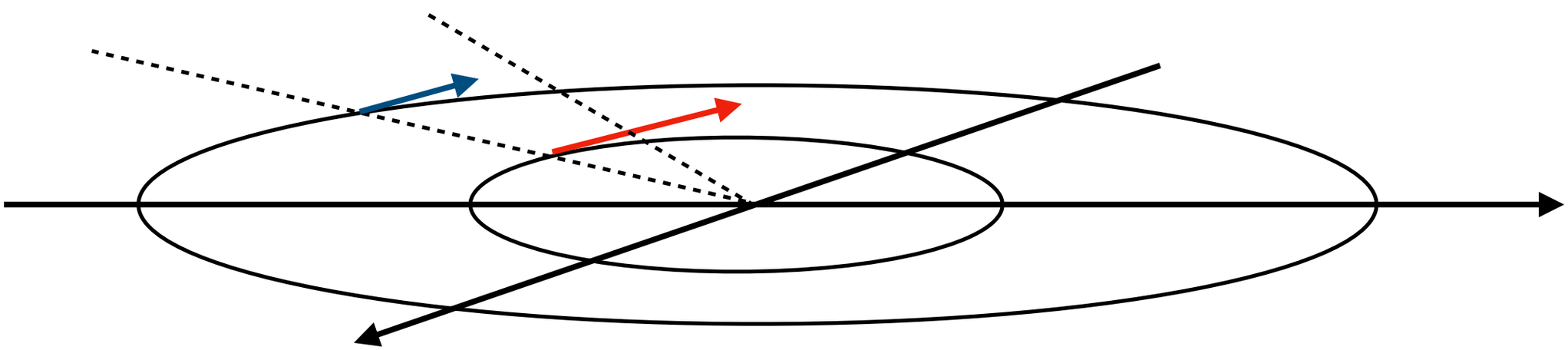
- But we need to pass longer Euclidean distance to move to the same angle, i.e. to make same update of DNN



Effective LR

- But we need to pass longer Euclidean distance to move to the same angle, i.e. to make same update of DNN

- To compensate this we correct learning rate by $\eta = \eta_{eff}\|\theta\|^2$



Solution for u

$u-$**Stochastic DE**

$r-$**Ordinary DE**

- Solution (stationary distribution) for u

$$\rho(u) = \frac{1}{C} \exp\left(-\frac{L(u)}{\eta_{eff}\sigma^2} \right)$$

Solution for u

$u-$**Stochastic DE**

$r-$**Ordinary DE**

- Solution (stationary distribution) for u

$$\rho(u) = \frac{1}{C} \exp\left(-\frac{L(u)}{\eta_{eff}\sigma^2} \right)$$

- But this is exactly Gibbs distribution where

$$L(u) = U(z)$$

$$T = \eta_{eff}\sigma^2$$

Solution for r

$u-$Stochastic DE

$r-$Ordinary DE

- Solution for r
- It converges to

$$r = \sqrt{\frac{\eta_{eff}\sigma^2}{\lambda}}$$

Ideal gas law for DNN

- We have $r = \sqrt{\frac{\eta_{eff}\sigma^2}{\lambda}}$ and we know that $T = \eta_{eff}\sigma^2$

- Weight decay coefficient λ defines the force that pressures the weights to zero just like pressure p

- If we define the volume as a squared norm of the weights $V = r^2$ we get

$$\frac{\lambda r^2}{\eta_{eff}\sigma^2} = 1 \iff \frac{pV}{T} = 1$$

Isothermal training

- Since $\eta_{eff}\sigma^2 = T$ training with constant effective learning rate corresponds to isothermal process

- Since weight decay $\lambda = p$ is kept constant it also isobaric

Isothermal training

- Since $\eta_{eff}\sigma^2 = T$ training with constant effective learning rate corresponds to isothermal process

- Since weight decay $\lambda = p$ is kept constant it also isobaric

- But we know from thermodynamics that when T and p are fixed our system minimizes Gibbs potential

$$G = \bar{U}_\rho + pV - TS$$

Isothermal training

- Since $\eta_{eff}\sigma^2 = T$ training with constant effective learning rate corresponds to isothermal process

- Since weight decay $\lambda = p$ is kept constant it also isobaric

- But we know from thermodynamics that when T and p are fixed our system minimizes Gibbs potential

$$G = \bar{U}_\rho + pV - TS$$

- Or in terms of DNN this corresponds to

$$G = \mathbb{E}_\rho L(\theta) + \lambda \|\theta\|^2 - \eta_{eff}\sigma^2 S$$

Isothermal training

- Since $\eta_{eff}\sigma^2 = T$ training with constant effective learning rate corresponds to isothermal process

- Since weight decay $\lambda = p$ is kept constant it also isobaric

- But we know from thermodynamics that when T and p are fixed our system minimizes Gibbs potential

$$G = \bar{U}_\rho + pV - TS$$

- Or in terms of DNN this corresponds to

$$G = \mathbb{E}_\rho L(\theta) + \lambda \|\theta\|^2 - \eta_{eff}\sigma^2 S$$

Regularized Loss function **Caused by noise**

Correspondence

Value in DL	DNN	Thermodynamics	Value in TD
Distribution of weights	$\rho(\theta)$	$\rho(z)$	Distribution of microscopic variables
Average loss function	$\mathbb{E}_\rho L(\theta)$	$\bar{U}_\rho = \mathbb{E}_\rho U(z)$	Average internal energy
Effective learning rate	$\eta_{eff}\sigma^2$	T	Temperature
Weight decay coefficient	λ	p	Pressure
Squared weights norm	$\|\|\theta\|\|^2$	V	Volume
Entropy	$\mathbb{E}_\rho \log \rho(\theta)$	$\mathbb{E}_\rho \log \rho(z)$	Entropy

Interesting fact

- In ideal gas it is assumed that total energy is given by the sum of energies of each molecule

$$U(z) = \sum_i U(z_i)$$

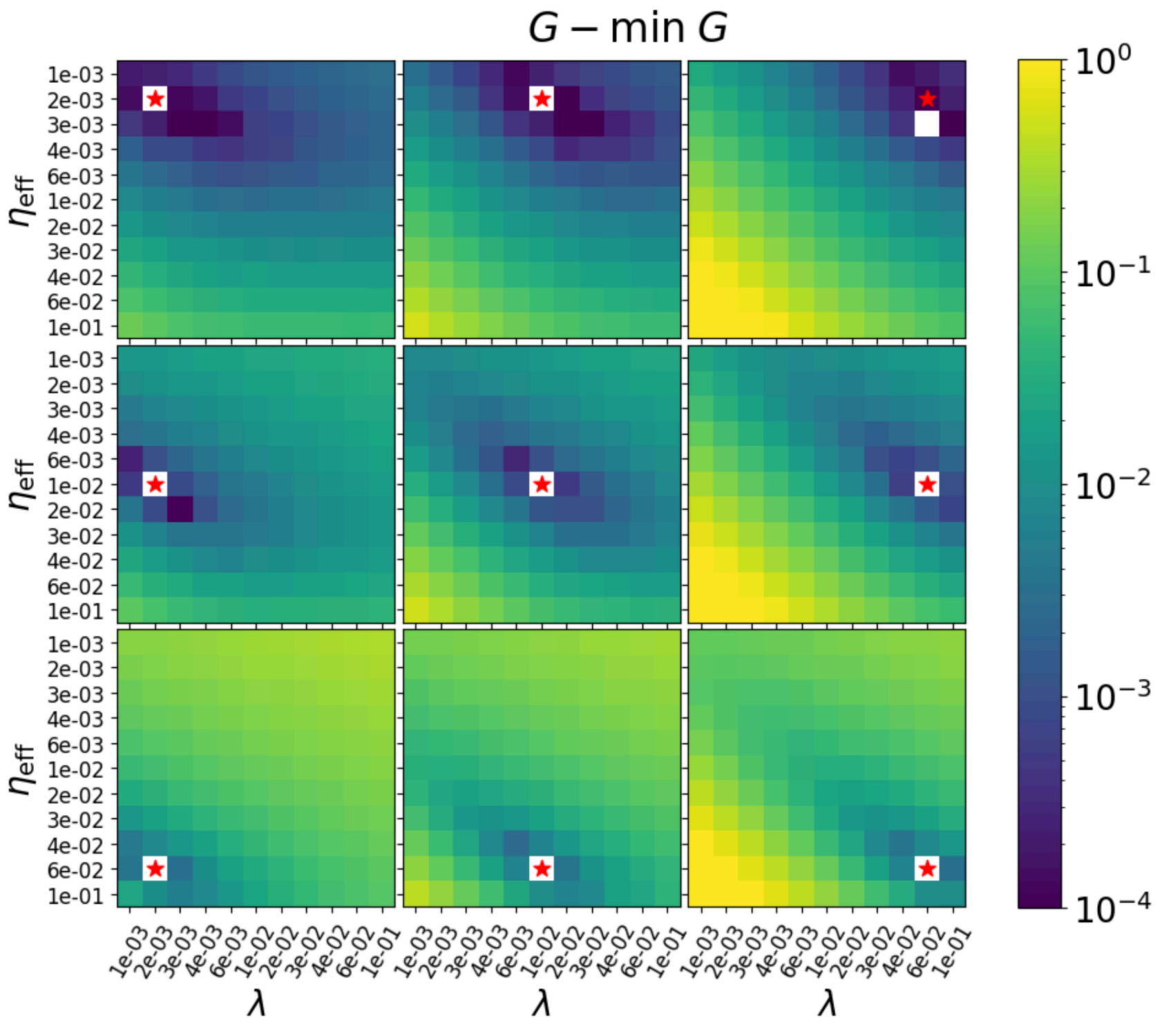
- Ideal gas law is derived based on this assumption

- In DNN the energy does not decompose

$$L(\theta) \neq \sum_i L(\theta_i)$$

- However ideal gas law still holds

Gibbs potential



Adiabatic process

- First law: $dQ = dU + pdV$
- In adiabatic process there's no heat exchange: $dQ = 0$

Adiabatic process

- First law: $dQ = dU + pdV$
- In adiabatic process there's no heat exchange: $dQ = 0$
- Consider heat capacity $c = \frac{dQ}{dT}$ under either constant V or p

Adiabatic process

- First law: $dQ = dU + pdV$
- In adiabatic process there's no heat exchange: $dQ = 0$
- Consider heat capacity $c = \frac{dQ}{dT}$ under either constant V or p

$$c_V = \frac{dQ}{dT} = \frac{dU}{dT} \qquad c_p = \frac{dQ}{dT} = \frac{dU + pdV}{dT} = c_V + 1$$

Adiabatic process

- First law: $dQ = dU + pdV$
- In adiabatic process there's no heat exchange: $dQ = 0$
- Consider heat capacity $c = \dfrac{dQ}{dT}$ under either constant V or p

$$c_V = \frac{dQ}{dT} = \frac{dU}{dT} \qquad c_p = \frac{dQ}{dT} = \frac{dU + pdV}{dT} = c_V + 1$$

$$\gamma = \frac{c_p}{c_V}$$

- Adiabatic law: $pV^\gamma = \text{Const}$

Adiabatic process

- First law: $dQ = dU + pdV$

- In adiabatic process there's no heat exchange: $dQ = 0$

- Consider heat capacity $c = \dfrac{dQ}{dT}$ under either constant V or p

$$c_V = \frac{dQ}{dT} = \frac{dU}{dT}$$

Can estimate empirically

$$c_p = \frac{dQ}{dT} = \frac{dU + pdV}{dT} = c_V + 1$$

$$\gamma = \frac{c_p}{c_V}$$

- Adiabatic law: $pV^\gamma = \text{Const}$

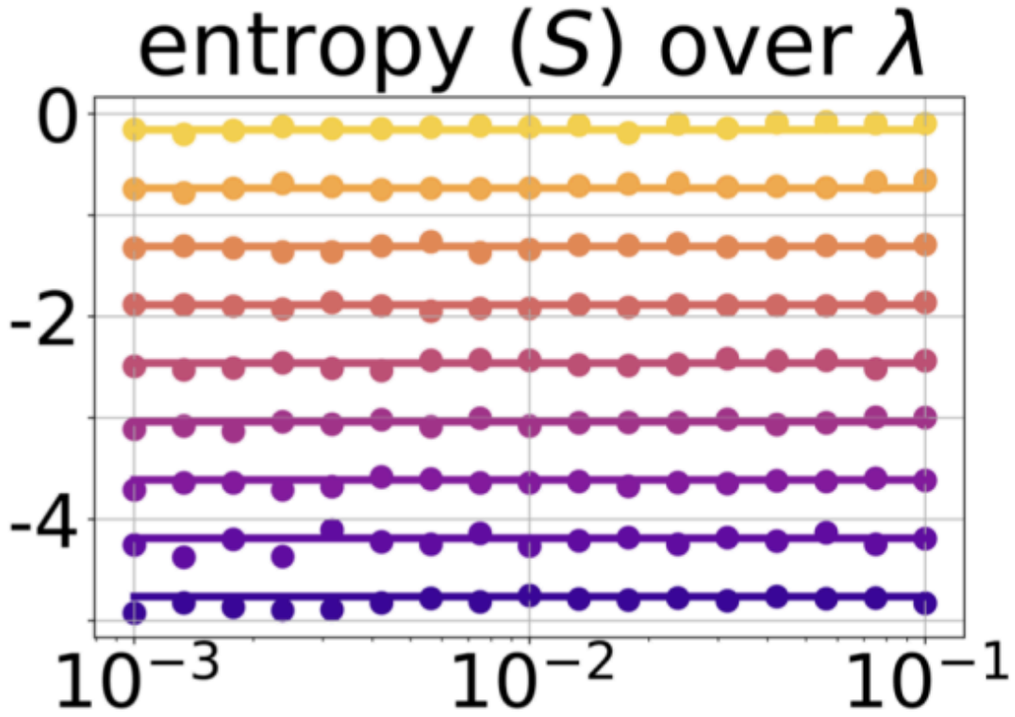
Adiabatic process

- Remember $dQ = TdS$
- No heat exchange $dQ = 0$ means entropy $S = \text{Const}$

Adiabatic process

- Remember $dQ = TdS$

- No heat exchange $dQ = 0$ means entropy $S = \text{Const}$



Colors correspond to different fixed values of pV^γ

Note that resulting distributions are different but their entropies are the same for fixed pV^γ

Training with constant η

- Remember $\eta = \eta_{eff}\|\theta\|^2 \sim TV$

- This corresponds to thermodynamic process with $TV =$ Const

- No analogues in elementary thermodynamics!

Training with constant η

- However we can still solve the differential equations for u, r and get

$$r^2 = \sqrt{\frac{\eta\sigma^2}{\lambda}} \qquad \rho(u) = \frac{1}{C}\exp\left(-\frac{r^2 L(u)}{\eta\sigma^2}\right)$$

- Then temperature is defined as $T = \dfrac{\eta\sigma^2}{r^2} = \sigma\sqrt{\eta\lambda}$

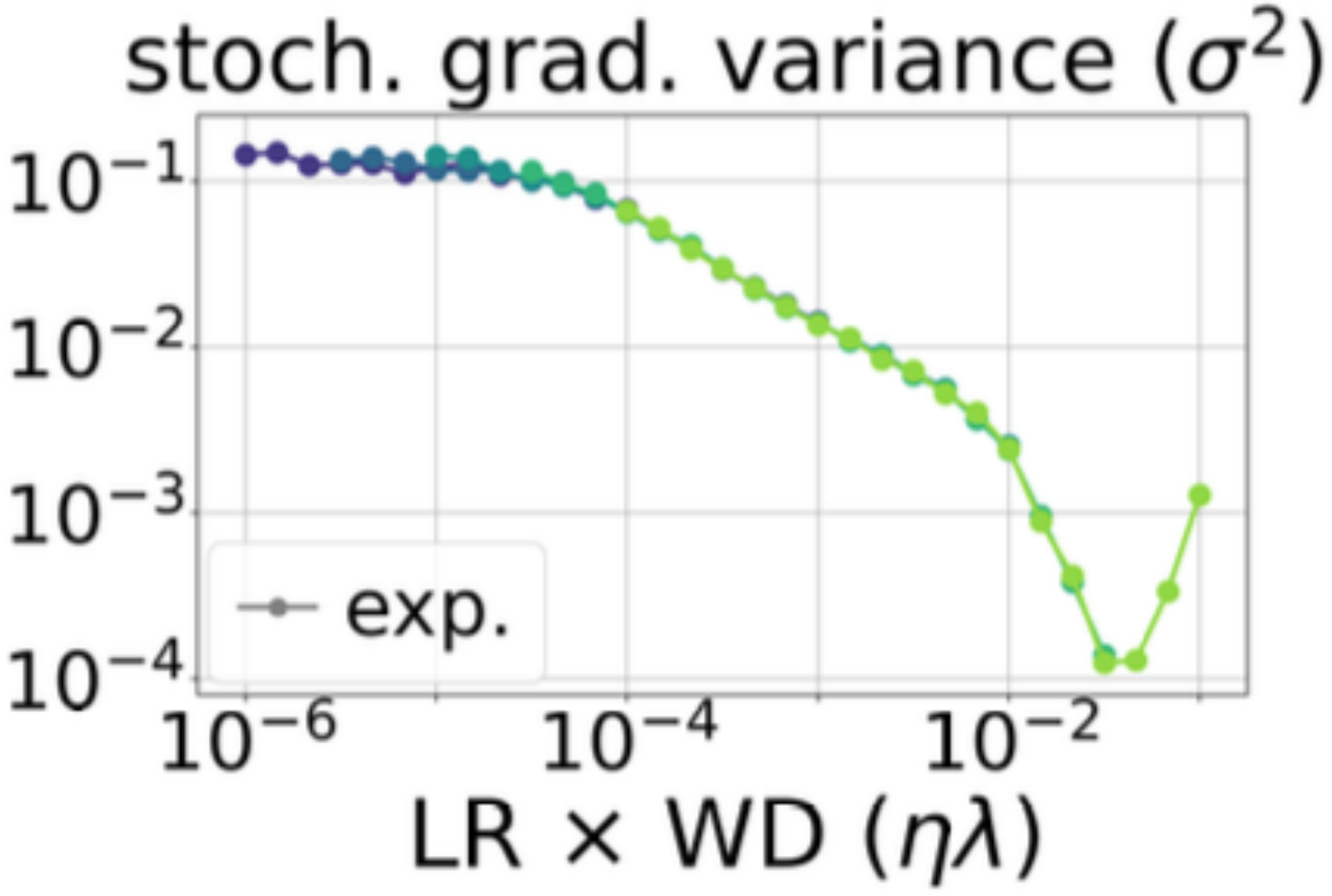
- Note that ideal gas law still holds with such definition of T

$$\frac{\sqrt{\lambda}r^2}{\sqrt{\eta}\sigma} = 1 \iff \frac{pV}{T} = 1$$

Real DNNs

- Noise in stochastic gradients is anisotropic and different in different regions
- No closed form expressions for $\bar{U}_\rho$ any more
- We can still check some relations

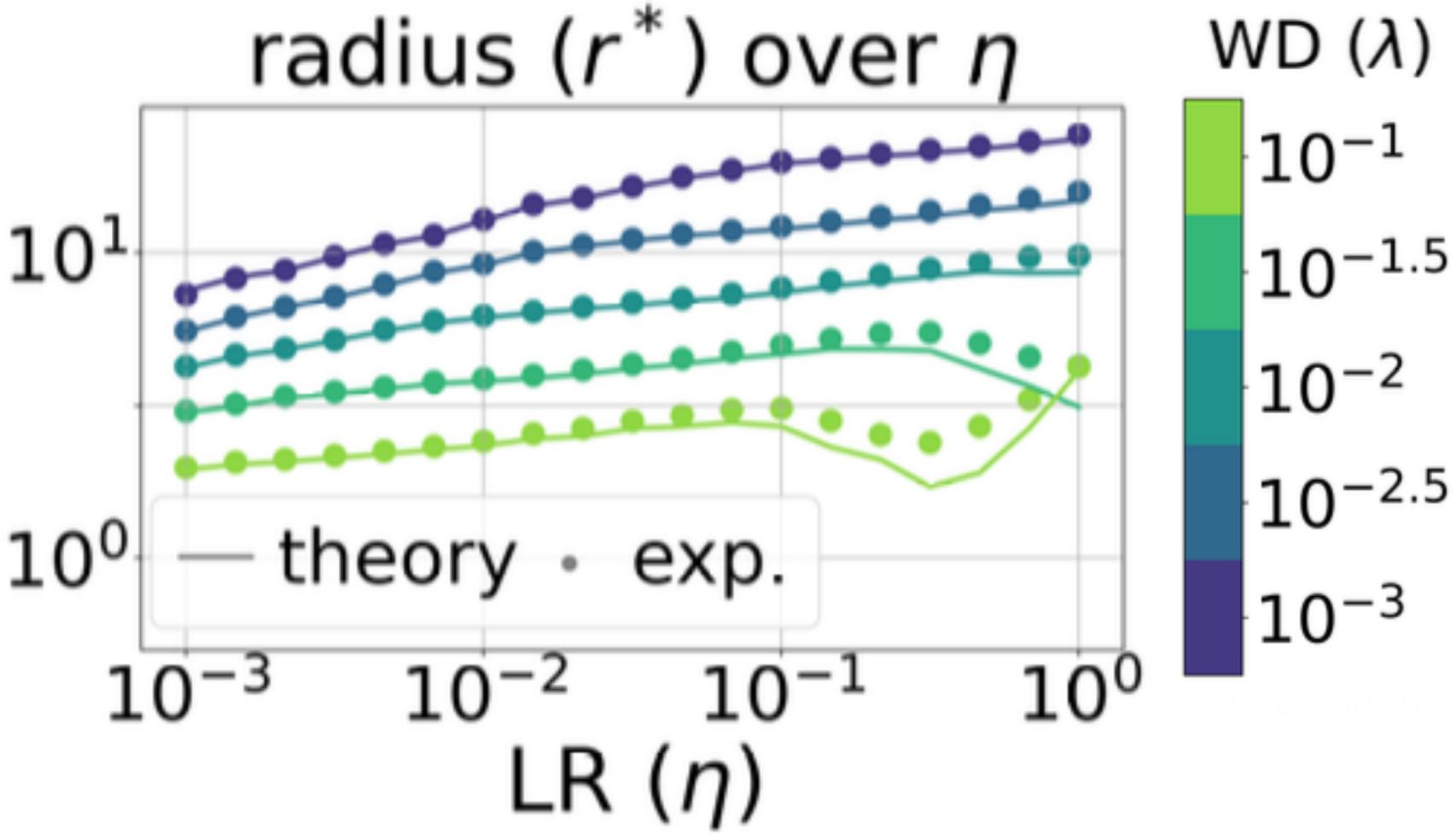
Real DNNs



Covariance trace is no more constant but depends on the product $\eta\lambda$

Implicit control over temperature $T = \sqrt{\eta\lambda}\mathrm{tr}\Sigma$

Real DNNs



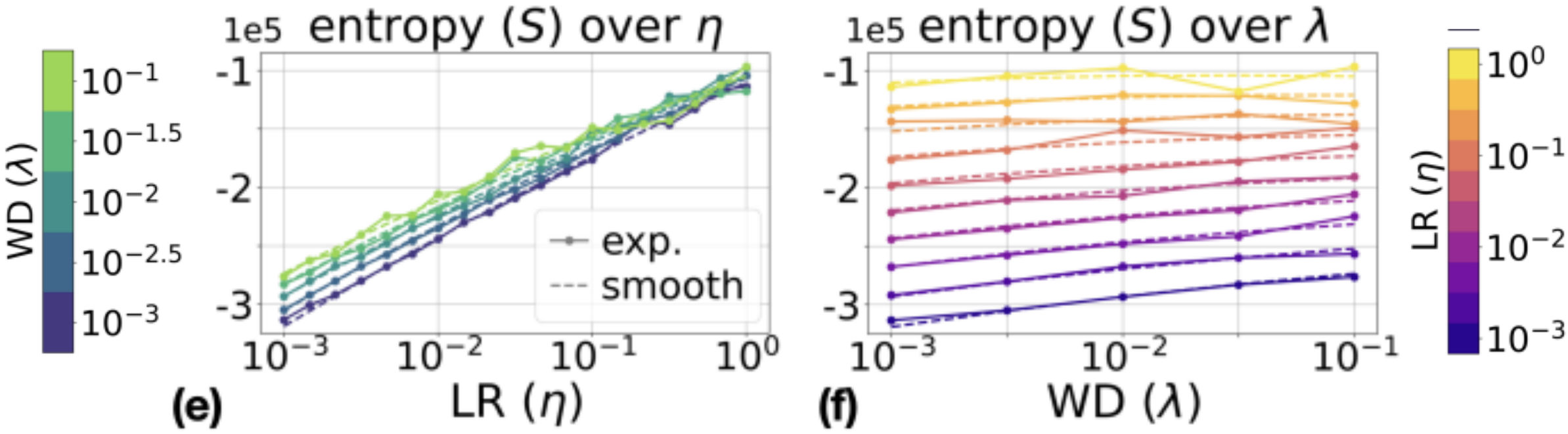
$$r^2 = \sqrt{\frac{\eta \mathbf{tr}\Sigma}{\lambda}}$$

Real DNNs

Maxwell relations for controls η and λ

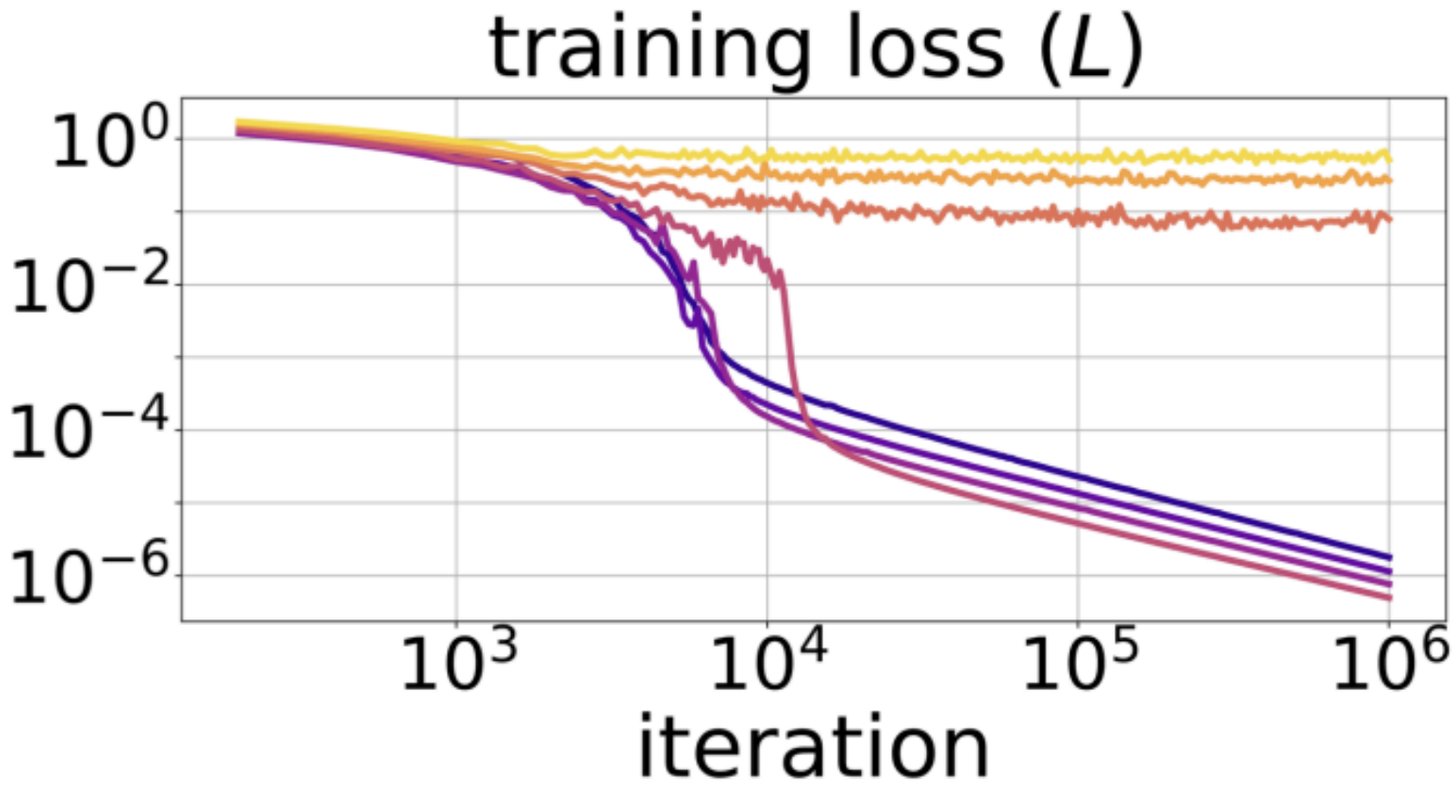
$$\left(\frac{\partial S}{\partial \log \eta}\right)_{\lambda} - \left(\frac{\partial S}{\partial \log \lambda}\right)_{\eta} = \frac{d-1}{2}$$

They are satisfied with absolute error 3%



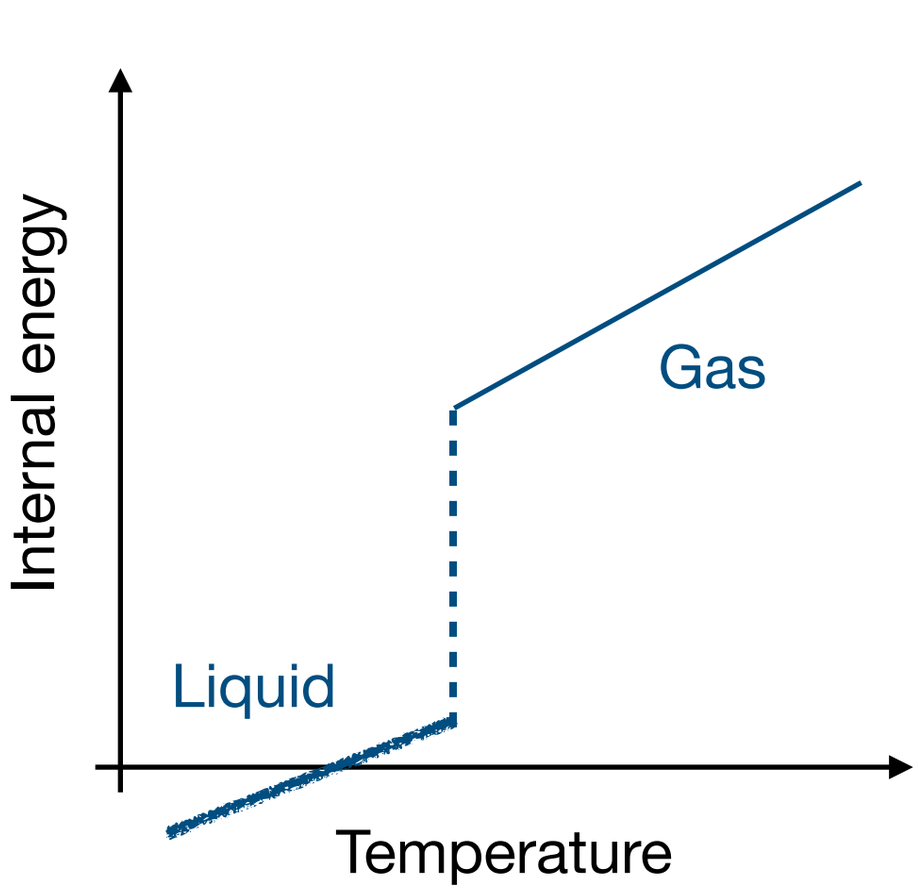
From ideal to real gas

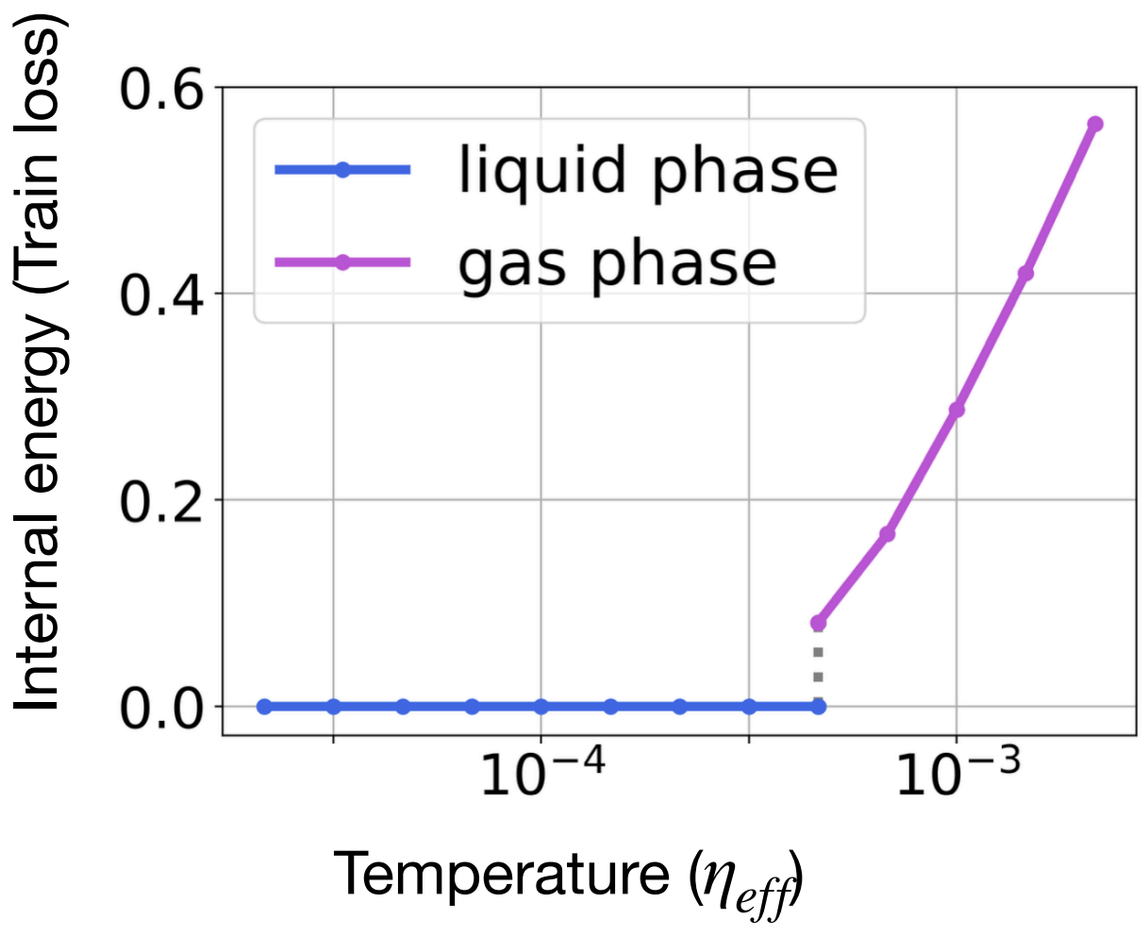
- In overparameterized DNNs $\sigma \sim \|\nabla L(w)\|$
- That's why SGD may converge given small enough fixed η



Phase transitions

- Real gases a phase transition when temperature drops
- Overparameterized DNN exhibit a similar phase transition!





Future work

- Consider different temperatures for different scale-invariant blocks
- Include [small number of] scale-non-invariant parameters
- Consider different optimizers rather than SGD (Adam, Muon, etc.)








Ildus Sadrtdinov · Ekaterina Lobacheva · Ivan Klimov · Mikhail Burtsev · Mikhail Katsnelson · Dmitry Vetrov

CONSTRUCTOR UNIVERSITY · Mila · Université de Montréal · Institute for Mathematical Truth, Beauty and Sciences, London · Radboud Universiteit

Masters in Bremen









MSc program



Ildus Sadrtdinov



Ekaterina Lobacheva



Ivan Klimov



Mikhail Burtsev



Mikhail Katsnelson



Dmitry Vetrov

CONSTRUCTOR UNIVERSITY

Mila

Université de Montréal

Radboud Universiteit