

Can Stationary Distributions of Scale-Invariant Neural Networks Be Described by the Thermodynamics of an Ideal Gas?

Markov lab seminar, 27.03.2026

Ildus Sadrtidinov

Ekaterina Lobacheva

Ivan Klimov

Mikhail Burtsev

Mikhail I. Katsnelson

Dmitry Vetrov

<https://arxiv.org/abs/2511.07308>



01

Motivation & Background

Why thermodynamics for neural network training?

Thermodynamics & scale-invariant networks

Motivation

Why use thermodynamics to understand neural network training?

THE CHALLENGE

Weights don't converge

SGD keeps wandering in weight space, no single 'optimal' solution

Distribution matters

Generalization depends on where the stationary distribution lives

Hyperparameters interact

LR, weight decay, batch size couple in non-obvious, empirical ways

STATISTICAL MECHANICS REINTERPRETATION

Gradient noise $\leftrightarrow$ Temperature

Stochastic gradient noise acts as thermal fluctuations

Stationary distr $\leftrightarrow$ Equilibrium

Weights at thermodynamic equilibrium of the system

$T, p, V \leftrightarrow$ Hyperparameters

Macroscopic state variables map onto LR, weight decay, parameter norm

GOAL

Derive a rigorous, quantitative thermodynamic framework directly from SDE analysis — not just an analogy, but exact predictions.

Background: Thermodynamics

State variables, free energies, and the Gibbs distribution

State Variables

T — **Temperature**: energy of random fluctuations

p — **Pressure**: external compressive force

V — **Volume**: size of the system

U — **Internal energy**: kinetic + potential energy of particles

S — **Entropy**: measure of disorder in the system

Thermodynamic Potentials

Helmholtz: $F = U - TS$ (fixed T, V)

Gibbs: $G = F + pV$ (fixed T, p)

At equilibrium: appropriate potential is minimized

Ideal Gas Law

When particles do not interact, the following equation holds:

$$pV = RT$$

Gibbs (Boltzmann) Distribution

At thermal equilibrium, state i is occupied with probability:

$$p_i \propto e^{-E_i/T} \quad U = \mathbb{E}[E_i] \quad S = \mathbb{E}[-\log p_i]$$

Maxwell Relations

Cross-partial symmetry of the thermodynamic potentials:

$$\frac{\partial^2 F}{\partial V \partial T} = -\frac{\partial S}{\partial V} = -\frac{\partial p}{\partial T} \quad \frac{\partial^2 G}{\partial p \partial T} = -\frac{\partial S}{\partial p} = \frac{\partial V}{\partial T}$$

Adiabatic Process

No heat exchange; entropy stays constant:

$$pV^\gamma = \text{const} \quad \gamma = C_p/C_v$$

Background: Scale-Invariant Networks

Normalization layers, effective learning rate, stationary distributions

Scale Invariance

Networks with BatchNorm / LayerNorm / Weight Norm are scale-invariant

The loss $L(w)$ depends only on the direction $\bar{w} = w/\|w\|$ — weights effectively live on a sphere

$$L(\alpha w) = L(w) \quad \forall \alpha > 0$$

Gradient Properties (from scale invariance)

Gradient are orthogonal to the weights:

$$\nabla L(w)^\top w = 0$$

Gradient scales as $1/r$:

$$\nabla L(w) = \nabla L(\bar{w}) / \|w\|$$

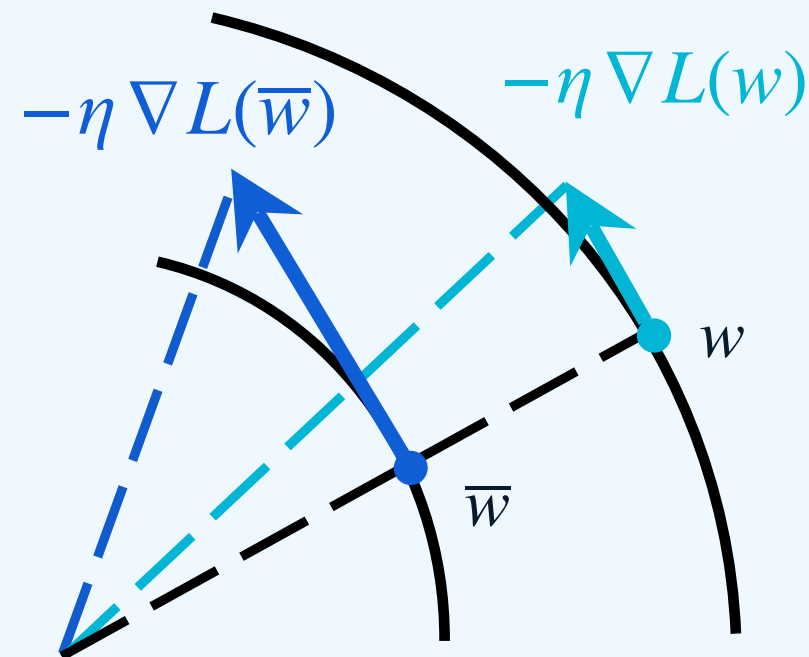
Noise covariance scales as $1/r^2$:

$$\Sigma_w = \Sigma_{\bar{w}} / \|w\|^2$$

Effective Learning Rate (ELR)

SGD with LR η acts on the unit sphere with:

$$\eta_{\text{eff}} = \eta / \|w\|^2$$



02

Theoretical Framework

Three protocols · Thermodynamic variables · Ideal gas law

Assumptions

Mathematical foundations for the stationary-distribution analysis

A1 Gaussian Gradient Noise

Mini-batch gradient noise is approximately Gaussian:

$$\nabla L_B(w) \approx \nabla L(w) + \Sigma_w^{1/2} \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I_d)$$

A2 Scale Invariance

The (mini-batch) loss is invariant to rescaling of weights:

$$L_B(\alpha w) = L_B(w) \quad \forall \alpha > 0$$

A3 SDE Approximation

Discrete SGD is well approximated by a continuous SDE for small η, λ :

$$dW_t = -\eta \nabla L(W_t) dt - \lambda W_t dt + \eta \Sigma_{W_t}^{1/2} dB_t$$

A4 Isotropic Noise — key for closed-form results

Gradient noise is isotropic in the tangent space of the sphere:

$$\Sigma_{\bar{w}} = \sigma^2 P_{\bar{w}}, \quad P_{\bar{w}} = I_d - \bar{w} \bar{w}^\top$$

NOTE: A4 holds exactly for the isotropic (VMF) model — relaxed to $\text{Tr}(\Sigma_{W_t}) = \text{const}$ for neural network experiments

Three Training Protocols

Different hyperparameter conventions — same thermodynamic structure

P1

Fixed Sphere

Projected SGD on $S^{d-1}(r)$ with $\eta_{\text{eff}} = \text{const}$

$$\rho^*(\bar{w}) \propto \exp\left(-\frac{L(\bar{w})}{T}\right), \quad T = \frac{\eta_{\text{eff}} \sigma^2}{2}, \quad r^* = r_0 \text{ (fixed)}$$

P2

Fixed ELR

Weight decay λ , learning rate $\eta = \eta_{\text{eff}} \|w\|^2$, $\eta_{\text{eff}} = \text{const}$

$$\rho^*(\bar{w}) \propto \exp\left(-\frac{L(\bar{w})}{T}\right), \quad T = \frac{\eta_{\text{eff}} \sigma^2}{2}, \quad r^* = \sqrt{\frac{\eta_{\text{eff}} \sigma^2 (d-1)}{2\lambda}}$$

P3

Fixed LR

Weight decay λ , learning rate $\eta = \text{const}$

$$\rho^*(\bar{w}) \propto \exp\left(-\frac{L(\bar{w})}{T}\right), \quad T = \sqrt{\frac{\eta \lambda \sigma^2}{2(d-1)}}, \quad r^* = \sqrt[4]{\frac{\eta \sigma^2 (d-1)}{2\lambda}}$$

For $p = \lambda$ and $V = (r^*)^2/2$, all three protocols yield: $pV = RT$, $R = (d-1)/2$

Thermodynamic Framework

Identifying T , p , V and the free energies

Thermodynamics	Optimizational Analog	Expression
Temperature T	Noise level (LR $\times$ grad. variance)	$T = \frac{\eta_{\text{eff}} \sigma^2}{2}$ (or $T = \sqrt{\frac{\eta \lambda \sigma^2}{2(d-1)}}$)
Pressure p	Weight decay	$p = \lambda$
Volume V	Half squared norm	$V = \frac{1}{2} \ w\ ^2$
Gas constant R	Dimension factor	$R = \frac{d-1}{2}$
Internal energy U	Expected loss	$U = \mathbb{E}_{\rho^*(w)}[L(w)]$
Entropy S	Shannon entropy of ρ^*	$S = \mathbb{E}_{\rho^*(w)}[-\log \rho^*(w)]$

03

Empirical Validation

VMF isotropic noise model · Neural networks

Empirical Validations V1 – V4

Four independent tests of the thermodynamic framework

V1 Stationary Radius

Does r^* follow the ideal gas law prediction?

$$r^* \propto (\eta/\lambda)^\alpha, \quad \alpha = 1/2 \text{ (fixed ELR)} \quad \text{or} \quad \alpha = 1/4 \text{ (fixed LR)}$$

→ *Ideal gas law determines equilibrium norm*

V2 Thermodynamic Potential Minimization

Is p^* the minimizer of the thermodynamic free energy?

$$\eta^*, \lambda^* = \operatorname{argmin}_{(\eta_i, \lambda_i)} \left\{ U_i - T^* S_i + p^* V_i \right\}$$

→ *Confirms the thermodynamic equilibrium interpretation*

V3 Maxwell Relation

Does the cross-partial identity of G hold?

$$\left. \frac{\partial S}{\partial \log \eta} \right|_\lambda - \left. \frac{\partial S}{\partial \log \lambda} \right|_\eta = \frac{d-1}{2} \quad (\text{fixed LR})$$

→ *Testable without access to the full distribution*

V4 Adiabatic Process

Is entropy conserved along $pV^\gamma = \text{const}$ curves?

$$S = \text{const} \quad \text{along} \quad pV^\gamma = \text{const}, \quad \gamma = C_p/C_v$$

→ *Validates heat capacity ratio $\gamma = 2$ from the isotropic model*

Von-Mises Fisher (VMF) isotropic noise model

Toy example for thermodynamic framework verification

$$L(w) = 1 + \mu^\top w / \|w\|, \quad \mu \in \mathbb{R}^d, \|\mu\| = 1$$

← *scale-invariant loss*

$$w_{k+1} = w_k - \eta \left(\nabla L(w_k) + \frac{\sigma}{\|w_k\|} P_{\hat{w}_k} \varepsilon_k + \lambda w_k \right), \quad \varepsilon_k \sim \mathcal{N}(0, I_d)$$

Why this loss?

SDE theory predicts VMF stationary distribution:

$$\rho_{\bar{w}}^* \propto \exp(-\mu^\top \bar{w} / T)$$

Closed-form U and S — enable direct comparison with the simulation

Hyperparameters

$d = 3, \sigma = 1$ (isotropic noise std)

$\eta, \lambda \in [10^{-3}, 10^{-1}]$ — 17 log-spaced values each

$2 \cdot 10^6$ iterations per run, μ sampled once and shared
kNN entropy estimator

Theoretical predictions (VMF asymptotics, $T \rightarrow 0$):

$$U = \frac{d-1}{2} T$$

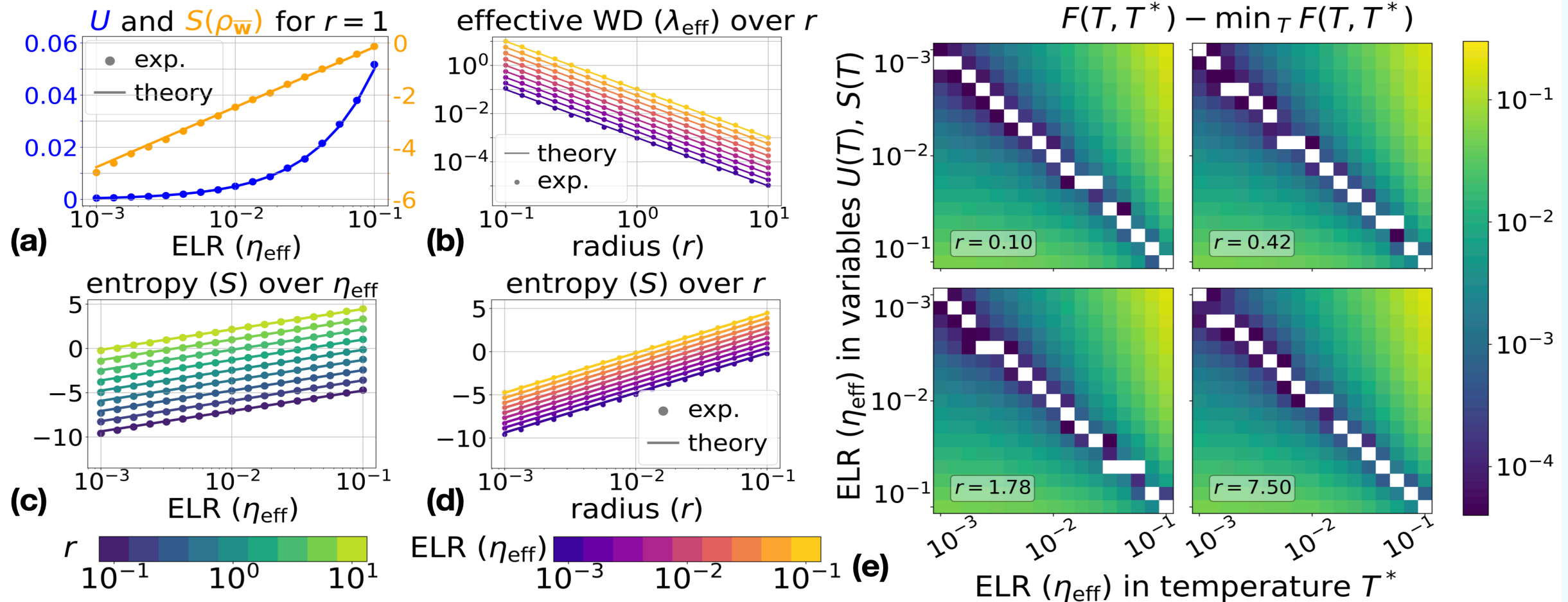
$$S(\rho_{\bar{w}}) = \frac{d-1}{2} \log(2\pi e T)$$

$$S = S(\rho_{\bar{w}}) + (d-1) \log r^*$$

Isotropic Noise Model — Fixed Sphere Protocol

VMF simulation on S^2 ($d = 3$) — Protocol 1: fixed radius r and ELR η_{eff}

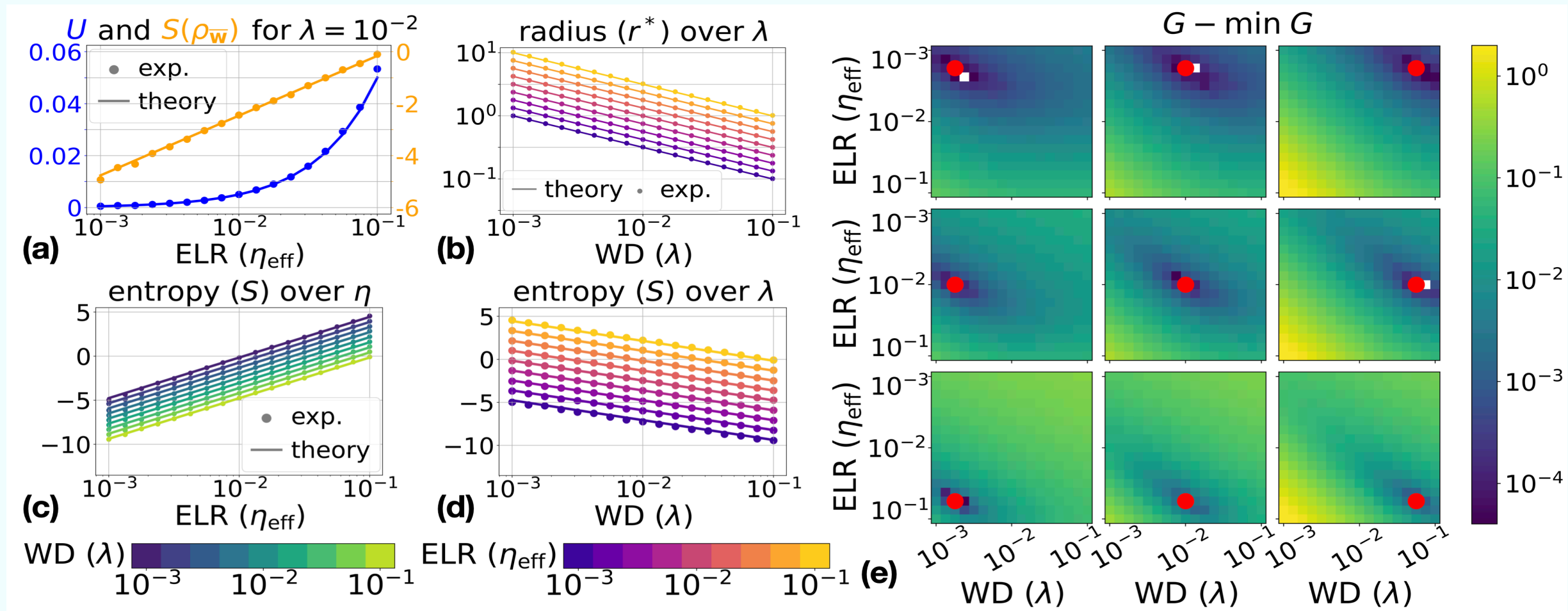
VMF model on fixed sphere. Subfigures a–d: numerical vs. theory. Subfigure e: Helmholtz energy minimization (V2).



Isotropic Noise Model — Fixed ELR Protocol

VMF simulation on S^2 ($d = 3$) — Protocol 2: fixed ELR η_{eff} and WD λ

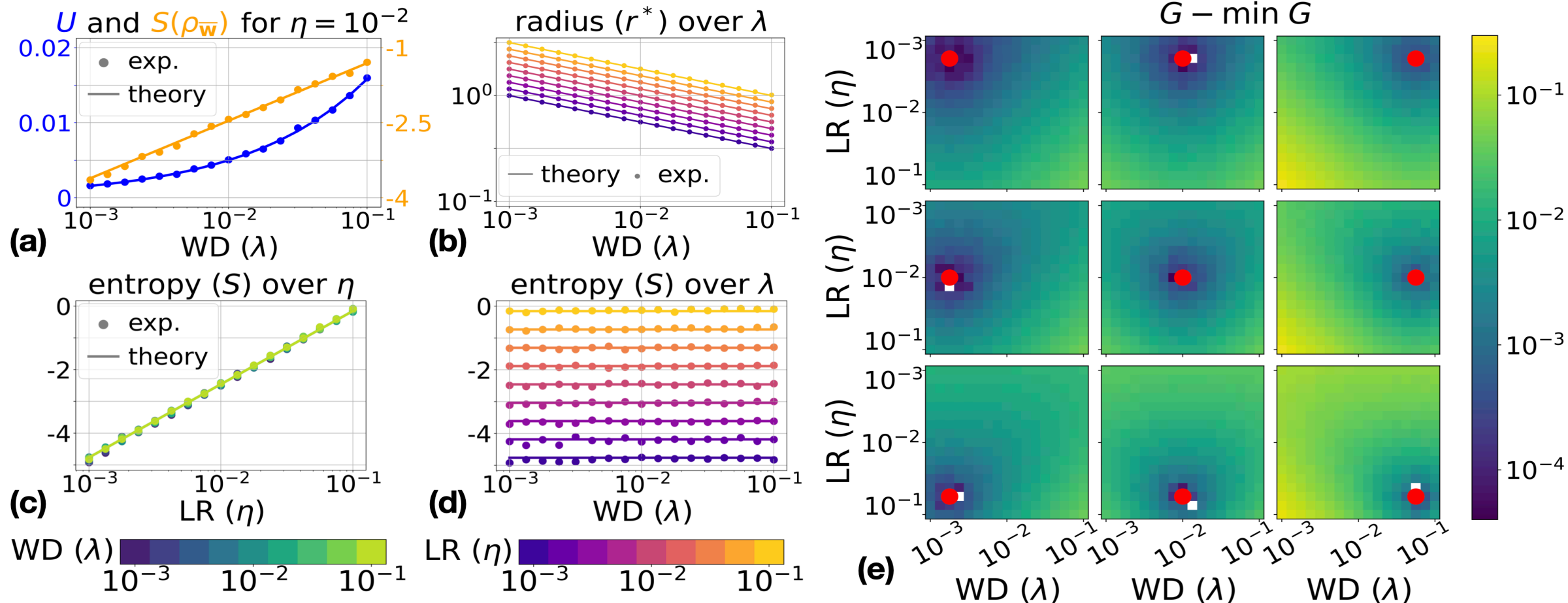
VMF model with fixed ELR. Subfigures a–d: numerical vs. theory. Subfigure e: Gibbs energy minimization (V2).



Isotropic Noise Model — Fixed LR Protocol

VMF simulation on S^2 ($d = 3$) — Protocol 3: fixed LR η and WD λ

VMF model with fixed LR. Subfigures a–d: numerical vs. theory. Subfigure e: Gibbs energy minimization (V2).



From Isotropic Model to Neural Networks

What changes — and what survives

Anisotropic gradient noise

$$\Sigma_w \neq \sigma^2 P_w$$

Does NOT break the thermodynamic analogy

Changes the potential to implicit function Φ :

$$\rho^*(w) \propto \exp(-\Phi(w)/T)$$

Φ captures noise anisotropy — not directly observable

Non-constant $\text{Tr } \Sigma_{\bar{w}}$ (σ^2 varies)

$$\sigma^2 = \frac{1}{d-1} \text{Tr } \Sigma_{\bar{w}} \neq \text{const}$$

DOES break the ideal gas law

Main problem: when σ^2 varies, there is no well-defined stationary radius r^*

In practice $\sigma^2 = \sigma^2(\eta\lambda)$ stabilizes at stationary

✓ CAN verify

V1 (stationary radius) + V3 (Maxwell relation)

Verified on ResNet-18, accounting for variable σ^2

$$r^* = \sqrt[4]{\frac{\eta\sigma(\eta\lambda)^2(d-1)}{2\lambda}} \quad \left. \frac{\partial S}{\partial \log \eta} \right|_{\lambda} - \left. \frac{\partial S}{\partial \log \lambda} \right|_{\eta} = \frac{d-1}{2}$$

✗ CANNOT verify directly

V2 (potential minimization) + V4 (adiabatic process)

Require knowledge of Φ — not directly observable

V2 also requires an unbiased entropy estimate — kNN estimate has bias $\mathcal{O}(N^{-2/d})$ for sample size N

Neural Network Experiments

ResNet-18 on CIFAR-10

Architecture

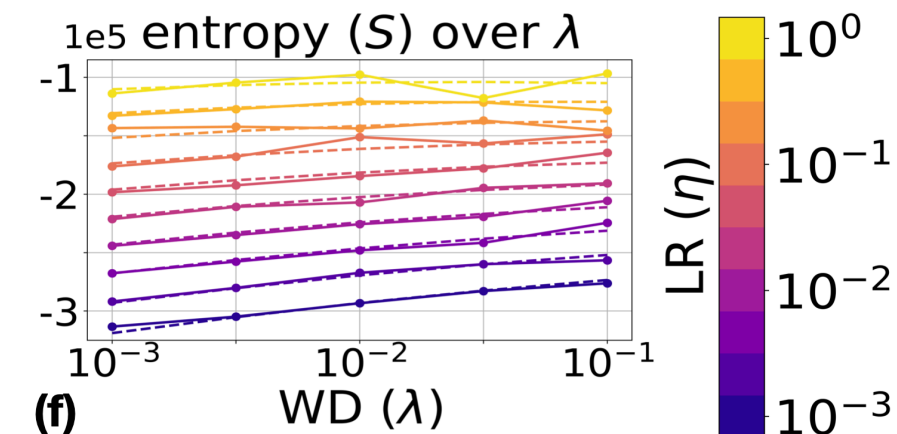
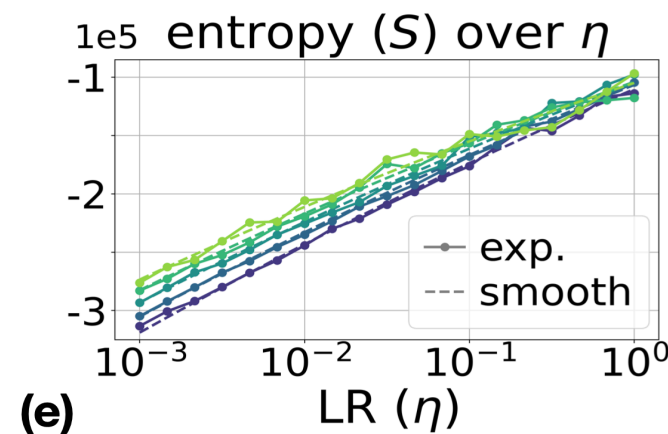
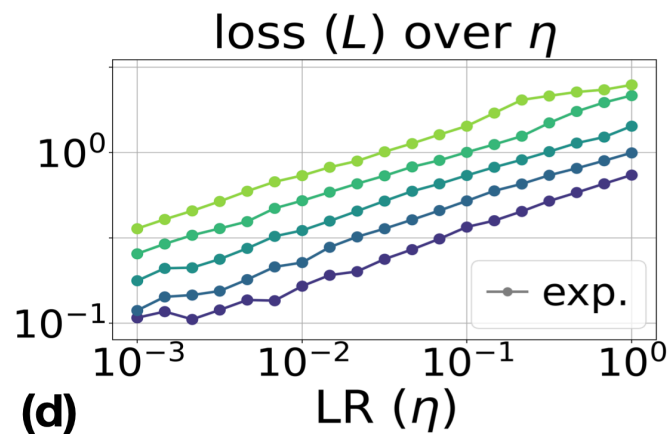
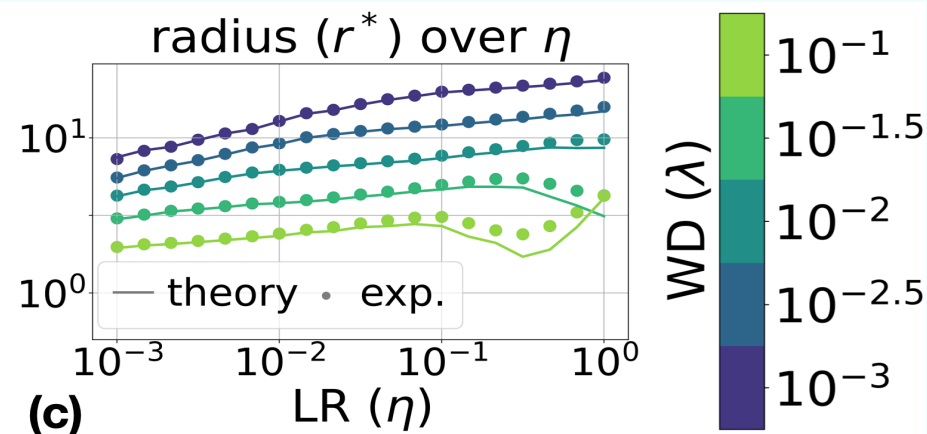
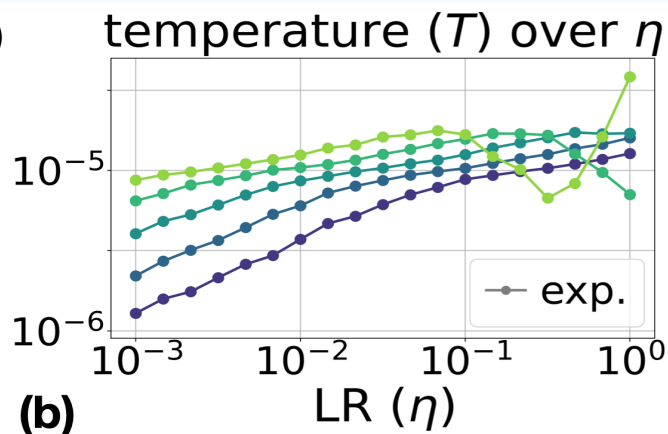
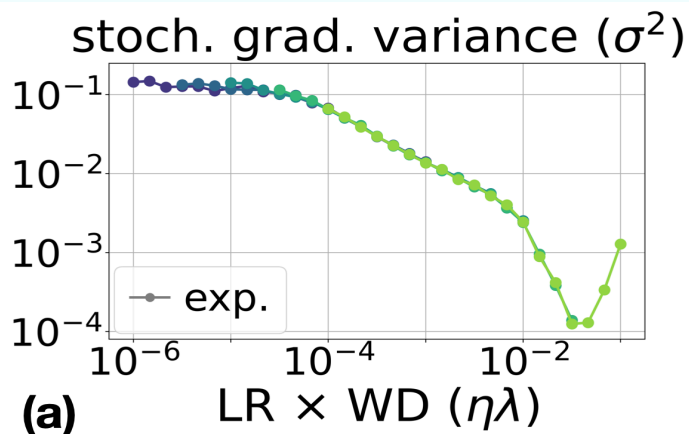
ResNet-18 (thin), $k = 4$, $d = 43692$

LR / WD range

$\eta \in [10^{-3}, 10^0]$, $\lambda \in [10^{-3}, 10^{-1}]$

Maxwell relations

Absolute error < 2.5 %



04

Discussion & Conclusion

Extension to real gases and Adam · Practical implications

Discussion

Extensions, limitations, and practical implications

Extending the ideal gas analogy

Real-gas analogy

$V = Z(p, T) \cdot RT/p$, where $Z(p, T)$ is the compressibility factor

Natural extension; mirrors $(r^*)^2/2 = \text{Tr} \Sigma_{\bar{w}} \cdot \eta_{\text{eff}}/4\lambda$

Non-scale-invariant networks

The definition $V = \|w\|^2/2$ relies on deterministic norm evolution

Must redefine V for models without full scale-invariance

SGD with momentum / Adam

More complex optimizers change effective temperature

Most common in practice — relevant to real training

Practical implications

LR scheduling

Maxwell relations $\Rightarrow$ we can control entropy decay rate

Regulate entropy decay: avoid premature convergence to sharp minima (associated with poor generalization)

Weight averaging (SWA)

T governs energy-entropy trade-off

LRs optimal for weight averaging often exceed convergence threshold¹; Maxwell relations quantify effect

Conclusion

SGD on scale-invariant networks obeys the ideal gas law.

- 1 Unified SDE derivation of stationary distributions for 3 training protocols
- 2 Thermodynamic mapping: $T \leftrightarrow$ stochastic gradient noise, $p \leftrightarrow$ weight decay, $V \leftrightarrow$ parameter norm
- 3 Ideal gas law $pV = RT$ derived rigorously from first principles for isotropic noise model
- 4 V1–V4 validated: radius, free energy, Maxwell relation, adiabatic process
- 5 Confirmed on ResNet-18 / CIFAR-10 with $< 2.5\%$ quantitative error for Maxwell relations

Thermodynamic reasoning gives exact, testable predictions for neural network training — not just inspiration.