Supplementary Information for *Thermodynamic cost of inference and learning in physical neural networks*

Alexei V. Tkachenko

Contents

1 Hamiltonian embedding and local linearization

The network has depth L and layer variables $\mathbf{x}_n \in \mathbb{R}^{d_n}$, with input $\mathbf{x}_0 = \mathbf{u}$ and output $\mathbf{x}_L = \mathbf{y}$. The feedforward map is

$$\mathbf{x}_n = f_n(\mathbf{z}_n), \qquad \mathbf{z}_n = W_n \mathbf{x}_{n-1} + \mathbf{b}_n, \qquad n = 1, \dots, L. \tag{S1}$$

The physical energy is

$$\mathcal{H}(\mathbf{x}; \boldsymbol{\theta}) = \frac{\kappa}{2} \sum_{n=1}^{L} \|\boldsymbol{\epsilon}_n\|^2, \qquad \boldsymbol{\epsilon}_n = \mathbf{x}_n - f_n(W_n \mathbf{x}_{n-1} + \mathbf{b}_n). \tag{S2}$$

For fixed $\mathbf{x}_0$ the feedforward solution has $\boldsymbol{\epsilon}_n = 0$ for all n and therefore $\mathcal{H} = 0$; we denote this zero-energy state $\mathbf{x}^*(\mathbf{u}; \boldsymbol{\theta})$.

Linearizing near $\mathbf{x}^*$ gives

$$\boldsymbol{\delta}\mathbf{x}_n = J_n \boldsymbol{\delta}\mathbf{x}_{n-1} + \boldsymbol{\delta}\boldsymbol{\epsilon}_n, \qquad J_n = \mathrm{diag}[f_n'(\mathbf{z}_n)] W_n, \tag{S3}$$

with downstream Green function

$$\widehat{G}_{\ell \to m} = J_m J_{m-1} \cdots J_{\ell+1} \ (\ell < m), \qquad \widehat{G}_{m \to m} = I, \tag{S4}$$

so a residual fluctuation injected at layer ℓ appears at layer m as $\boldsymbol{\delta}\mathbf{x}_m = \widehat{G}_{\ell \to m} \boldsymbol{\delta}\boldsymbol{\epsilon}_\ell$.

At temperature T the residual fluctuations are independent Gaussians in the harmonic approximation, $\langle \boldsymbol{\delta}\boldsymbol{\epsilon}_\ell \boldsymbol{\delta}\boldsymbol{\epsilon}_q^T \rangle = (k_\mathrm{B} T/\kappa) \delta_{\ell q} I$, so the covariance of activity fluctuations at layer m is

$$C_m = \frac{k_\mathrm{B} T}{\kappa} \mathcal{C}_m, \qquad \mathcal{C}_m = \sum_{\ell=1}^{m} \widehat{G}_{\ell \to m} \widehat{G}_{\ell \to m}^T. \tag{S5}$$

The inverse covariance C_m^{-1} is the Fisher metric for displacements of the layer-m mean, and $\mathcal{C}_m$ is its dimensionless part. The $\ell = m$ term of Eq. (S5) is the identity, since $\widehat{G}_{m \to m} = I$, so

$$\mathcal{C}_m = I + \sum_{\ell < m} \widehat{G}_{\ell \to m} \widehat{G}_{\ell \to m}^T \ \succeq \ I \tag{S6}$$

is positive definite for every activation pattern: each physical layer carries its own residual fluctuation, and vanishing f_n' can only remove upstream contributions, never the direct one. No active-subspace restriction or pseudo-inverse is needed, and d_m is the full layer dimension throughout. Rank deficiency would arise only in a modified model in which some coordinates carry no independent residual noise.

2 Exact constancy of the free network's free energy

Clamp $\mathbf{x}_0 = \mathbf{u}$ and let the remaining layer variables range over $\mathbf{x}_n \in \mathbb{R}^{d_n}$. Define the residual map

$$\Phi : \ (\mathbf{x}_1, \dots, \mathbf{x}_L) \ \longmapsto \ (\boldsymbol{\epsilon}_1, \dots, \boldsymbol{\epsilon}_L), \qquad \boldsymbol{\epsilon}_n = \mathbf{x}_n - f_n(W_n \mathbf{x}_{n-1} + \mathbf{b}_n). \tag{S7}$$

Φ is a bijection of $\prod_n \mathbb{R}^{d_n}$ onto itself: given $\{\boldsymbol{\epsilon}_n\}$ the layer variables are recovered uniquely and recursively by $\mathbf{x}_n = \boldsymbol{\epsilon}_n + f_n(W_n \mathbf{x}_{n-1} + \mathbf{b}_n)$ starting from the clamped $\mathbf{x}_0$. Its Jacobian matrix has blocks

$$\frac{\partial \boldsymbol{\epsilon}_n}{\partial \mathbf{x}_m} = \begin{cases} I, & m = n, \\ -J_n, & m = n-1, \\ 0, & \text{otherwise}, \end{cases} \tag{S8}$$

i.e. it is block lower-triangular with identity diagonal blocks, so $|\det \partial\boldsymbol{\epsilon}/\partial\mathbf{x}| = 1$ identically. No smoothness or invertibility of the activation functions is required beyond measurability, and the result holds for arbitrary nonlinear f_n, including saturating and rectified ones.

The partition function of the free network therefore factorizes,

$$Z(\mathbf{u};\boldsymbol{\theta}) = \int e^{-\beta\mathcal{H}(\mathbf{x};\boldsymbol{\theta})} \prod_{n=1}^{L} \mathrm{d}\mathbf{x}_n = \int e^{-\beta\frac{\kappa}{2}\sum_n \|\boldsymbol{\epsilon}_n\|^2} \prod_{n=1}^{L} \mathrm{d}\boldsymbol{\epsilon}_n = \prod_{n=1}^{L} \left(\frac{2\pi k_\mathrm{B}T}{\kappa}\right)^{d_n/2}, \tag{S9}$$

and the equilibrium free energy

$$F_\mathrm{free} = -k_\mathrm{B}T \ln Z = \frac{k_\mathrm{B}T}{2} \sum_{n=1}^{L} d_n \ln \frac{\kappa}{2\pi k_\mathrm{B}T} \tag{S10}$$

is a constant: independent of the input $\mathbf{u}$, of all weights and biases, and exact at every temperature. Equation (S10) is not a harmonic or saddle-point estimate; it is the full equilibrium free energy.

The same conclusion follows without computing any integral, by splitting $F = \langle\mathcal{H}\rangle - TS$. In residual coordinates the Gibbs measure is a fixed isotropic Gaussian, so $\langle\mathcal{H}\rangle = \frac{\kappa}{2}\langle\sum_n \|\boldsymbol{\epsilon}_n\|^2\rangle = \frac{1}{2}\big(\sum_n d_n\big)k_\mathrm{B}T$ by equipartition, independent of input and parameters. And the map Φ is volume preserving, so it leaves the differential entropy unchanged; S is that of the fixed Gaussian. Both terms are separately constant, which makes the mechanism transparent: changing the input deforms the zero-energy manifold without stretching phase space anywhere.

Two consequences are used in the main text. First, quasi-static change of the input connects states of equal free energy, so the reversible work vanishes identically and $E_\mathrm{inf}^\mathrm{min} = 0$: not merely zero dissipation, but zero net work. Second, $\partial F_\mathrm{free}/\partial\boldsymbol{\theta} = 0$, so the free network exerts no force on the parameters and the clamped free energy of Eq. (S37) alone is the physical loss, with no free-minus-clamped subtraction required.

If κ varies between coordinates, $\kappa \to \kappa_i$ in Eq. (S10) and the conclusion is unchanged provided the stiffnesses are not themselves trainable. If the physical coordinates are confined to a bounded range rather than to $\mathbb{R}^{d_n}$, the change of variables no longer maps the domain onto itself exactly and F_free acquires an input dependence suppressed by $\exp(-\kappa\ell^2/2k_\mathrm{B}T)$, with ℓ the distance from the operating point to the confining boundary.

2.1 Monostability and the origin of the Landauer bound

The reversibility of quasi-static inference is not special to Eq. (S2). Consider any machine computing by relaxation on a free-energy landscape $F(\mathbf{x};\mathbf{u})$ under external control $\mathbf{u}$. If the equilibrium state is unique and varies continuously along the control path, the system tracks equilibrium under slow driving, the work is the state function ΔF, and the dissipated part $W - \Delta F$ vanishes as $\tau \to \infty$. A finite quasi-static cost requires the system to be left in a state that is not the equilibrium one, which requires the landscape to support several states whose lifetimes exceed the protocol duration. Digital logic is built on exactly this: a bit is a pair of metastable wells engineered to be long-lived, and erasing it merges two basins, which is where $k_\mathrm{B}T\ln 2$ enters. Landauer's bound therefore constrains multistable memory, not computation as such.

For the embedding used here, monostability follows from $\mathcal{H} = \frac{\kappa}{2}\boldsymbol{\eta}^T A^T A\boldsymbol{\eta}$ with A block lower-bidiagonal with unit diagonal blocks: A is invertible, $A^TA \succ 0$, and the compatible state is unique for every input and every parameter set. The construction thus belongs to the monostable family, and Eq. (S10) adds to it the stronger property $\Delta F \equiv 0$. A general monostable machine dissipates nothing quasi-statically but may require reversible work to be supplied and later recovered; this one requires none.

2.2 Fluctuational correction to the clamped free energy

Clamping the output breaks the factorization of Eq. (S9): the last residual is no longer a free integration variable. Performing the Gaussian integral over the hidden coordinates around the constrained minimum gives

$$F_\alpha = \mathcal{H}_{\min}(\boldsymbol{\theta};\alpha) + \frac{k_{\mathrm{B}}T}{2}\ln\det\frac{K_\alpha}{2\pi k_{\mathrm{B}}T}, \qquad K_\alpha = \kappa\, A_c^T A_c, \tag{S11}$$

where A_c is the compatibility operator restricted to the hidden coordinates $\mathbf{x}_1,\dots,\mathbf{x}_{L-1}$, i.e. the rectangular matrix mapping hidden displacements to all L residuals. Unlike the free case, the fluctuational term depends on the parameters. Its structure, however, is constrained. Writing $\tilde{A}$ for the square unit-lower-bidiagonal block acting among the hidden layers, one finds $\det\tilde{A} = 1$ and

$$\det\big(A_c^T A_c\big) = \det\big(I + J_L\, S\, J_L^T\big)\,, \qquad S = \big[(\tilde{A}^T\tilde{A})^{-1}\big]_{L-1,L-1}, \tag{S12}$$

so the entropic term depends on $\boldsymbol{\theta}$ only through the layer Jacobians, the gains, and never through the strain. It therefore cannot rotate the error signal; it adds to the physical loss a data-independent (for fixed activation pattern) gain penalty.

For the two-layer network of the main text, $K = \kappa(I + W_2^T W_2)$ and the entropic force on the output weights is

$$\nabla_{W_2}\left[\frac{k_{\mathrm{B}}T}{2}\ln\det(I + W_2^T W_2)\right] = k_{\mathrm{B}}T\, W_2\big(I + W_2^T W_2\big)^{-1} \xrightarrow{\|W_2\|\ll 1} k_{\mathrm{B}}T\, W_2\,: \tag{S13}$$

the fluctuational correction is a thermal weight decay of strength $k_{\mathrm{B}}T$, a Laplace-type Occam factor penalizing large gains. It is subleading to the strain gradient whenever the strain per coordinate exceeds $k_{\mathrm{B}}T$, which is the same condition $\Gamma \gtrsim 1$ that governs the fixed-covariance approximation and the Bures suppression, Eq. (S29).

Two practical consequences for the learning rule of the main text. First, the parameter gradient at finite temperature is $\partial F_\alpha/\partial\boldsymbol{\theta} = \langle\partial\mathcal{H}/\partial\boldsymbol{\theta}\rangle$, an exact envelope-theorem statement; the difference between $\langle\boldsymbol{\sigma}_n\rangle$ and $\boldsymbol{\sigma}_n$ evaluated at the constrained minimum is precisely the gradient of the entropic term above. Second, the operational advantage over free-minus-clamped schemes is unchanged but should be stated correctly: subtracting F_{free} would not have removed the entropic term, since F_{free} is constant; the advantage is that the free phase exerts no parameter force at all, so nothing need be measured in it.

3 Finite-rate inference: the optimal-transport bound

3.1 Master inequality

The overdamped dynamics of the main text, with mobility μ and fluctuation–dissipation noise at temperature T, has the Fokker–Planck form

$$\partial_t\rho = -\nabla\cdot(\rho\,\mathbf{v}), \qquad \mathbf{v} = \mu\,(-\nabla\mathcal{H} - k_{\mathrm{B}}T\nabla\ln\rho)\,, \tag{S14}$$

where $\rho(\mathbf{x},t)$ is the joint density of all physical coordinates and $\mathbf{v}$ is the local mean velocity. The irreversible work over a protocol of duration τ is the kinetic action

$$E_{\mathrm{inf}} = \frac{1}{\mu}\int_0^\tau \mathrm{d}t\int \rho\,\|\mathbf{v}\|^2\,\mathrm{d}\mathbf{x}. \tag{S15}$$

The Benamou–Brenier theorem [1] states that the minimum of $\int_0^\tau \mathrm{d}t \int \rho \|\mathbf{v}\|^2 \mathrm{d}\mathbf{x}$ over all pairs $(\rho, \mathbf{v})$ satisfying the continuity equation with fixed endpoints ρ_0, ρ_τ equals $\mathcal{W}_2^2(\rho_0, \rho_\tau)/\tau$, where $\mathcal{W}_2$ is the Wasserstein-2 distance,

$$\mathcal{W}_2^2(\rho_0, \rho_\tau) = \inf_{\pi \in \Pi(\rho_0, \rho_\tau)} \int \|\mathbf{x} - \mathbf{x}'\|^2 \, \mathrm{d}\pi(\mathbf{x}, \mathbf{x}'), \tag{S16}$$

the infimum running over couplings with the given marginals. Hence

$$E_{\mathrm{inf}} \;\geq\; \frac{\mathcal{W}_2^2(\rho_0, \rho_\tau)}{\mu\tau} \;=\; \frac{1}{\mu\tau}\, \mathcal{W}_2^2(\rho_0, \rho_\tau), \qquad \tau_0 = (\mu\kappa)^{-1}. \tag{S17}$$

Equation (S17) is the minimum-dissipation form of the thermodynamic speed limit [2–6]; the relation of the transport and information-geometric formulations is reviewed in Ref. [7]. Three features are used in the main text.

(i) No slow-driving assumption. Equation (S17) constrains the endpoints only, so it holds arbitrarily far from quasistatic operation. In the deterministic slow-driving limit $\rho \to \delta(\mathbf{x} - \mathbf{x}^*)$, $\mathbf{v} \to \dot{\mathbf{x}}^*$ and Eq. (S15) reduces to the Rayleigh form

$$P_{\mathrm{inf}} = \frac{1}{\mu} \sum_{n=1}^{L} \|\dot{\mathbf{x}}_n^*\|^2 \,. \tag{S18}$$

(ii) No Gaussian assumption. The harmonic approximation enters only when $\mathcal{W}_2$ is evaluated, not when the bound is derived.

(iii) Achievability. Equality holds for the constant-speed displacement interpolation, so the bound is an attainable target rather than a generic inequality.

3.2 Gaussian evaluation

For thermal endpoint states $\rho_0 = \mathcal{N}(\mathbf{x}_i^*, \Sigma_i)$ and $\rho_\tau = \mathcal{N}(\mathbf{x}_{i+1}^*, \Sigma_{i+1})$ over the full set of physical coordinates, the Givens–Shortt formula [8] gives

$$\mathcal{W}_2^2 = \sum_{n=1}^{L} \|\Delta\mathbf{x}_n^*\|^2 + \mathcal{B}^2(\Sigma_i, \Sigma_{i+1}), \qquad \mathcal{B}^2 = \mathrm{Tr}\Big[\Sigma_i + \Sigma_{i+1} - 2\big(\Sigma_{i+1}^{1/2}\Sigma_i\Sigma_{i+1}^{1/2}\big)^{1/2}\Big], \tag{S19}$$

with $\Delta\mathbf{x}_n^*$ the shift of the layer-n feedforward state. The mean-shift term is a bare Euclidean displacement, independent of the covariance; the Bures term $\mathcal{B}^2$ is non-negative and vanishes iff $\Sigma_i = \Sigma_{i+1}$. It is precisely the covariance-mismatch contribution neglected in the entropic treatment, and it is retained here as an additive cost rather than discarded.

3.3 Relaxation to the information-geometric form

Fix an observed layer m. On the linearized manifold the same displacement can be described from any earlier layer,

$$\Delta\mathbf{x}_m = \widehat{G}_{\ell\to m}\Delta\mathbf{x}_\ell, \qquad \ell \leq m. \tag{S20}$$

Summing over $\ell = 1, \ldots, m$ gives $\sum_{\ell=1}^{m} \widehat{G}_{\ell\to m}\Delta\mathbf{x}_\ell = m\,\Delta\mathbf{x}_m$. Taking the scalar product with $\mathcal{C}_m^{-1}\Delta\mathbf{x}_m$ and applying Cauchy–Schwarz,

$$\left(m\,\Delta\mathbf{x}_m^T\mathcal{C}_m^{-1}\Delta\mathbf{x}_m\right)^2 = \left(\sum_{\ell=1}^{m}\Delta\mathbf{x}_\ell^T\widehat{G}_{\ell\to m}^T\mathcal{C}_m^{-1}\Delta\mathbf{x}_m\right)^2 \tag{S21}$$

$$\leq \left(\sum_{\ell=1}^{m}\|\Delta\mathbf{x}_\ell\|^2\right)\left(\sum_{\ell=1}^{m}\Delta\mathbf{x}_m^T\mathcal{C}_m^{-1}\widehat{G}_{\ell\to m}\widehat{G}_{\ell\to m}^T\mathcal{C}_m^{-1}\Delta\mathbf{x}_m\right) \tag{S22}$$

$$= \left(\sum_{\ell=1}^{m}\|\Delta\mathbf{x}_\ell\|^2\right)\Delta\mathbf{x}_m^T\mathcal{C}_m^{-1}\Delta\mathbf{x}_m, \tag{S23}$$

where the last line uses the definition (S5) of $\mathcal{C}_m$. Cancelling the common positive factor,

$$\sum_{\ell=1}^{m}\|\Delta\mathbf{x}_\ell\|^2 \;\geq\; m^2\,\Delta\mathbf{x}_m^T\mathcal{C}_m^{-1}\Delta\mathbf{x}_m. \tag{S24}$$

This holds for every m, so each layer supplies a valid lower bound and the strongest may be retained. With $C_m^{-1} = (\kappa/k_\mathrm{B}T)\mathcal{C}_m^{-1}$ and the Gaussian Kullback–Leibler divergence for a mean shift at fixed covariance [9],

$$D_\mathrm{KL}^{(m)} = \tfrac{1}{2}\left\langle\Delta\mathbf{x}_m^T C_m^{-1}\Delta\mathbf{x}_m\right\rangle_\mathrm{tr}, \tag{S25}$$

and dropping the non-negative Bures term of Eq. (S19),

$$E_\mathrm{inf} \;\geq\; \frac{1}{\mu\tau}\,\mathcal{W}_2^2 \;\geq\; \frac{1}{\mu\tau}\sum_{n=1}^{L}\|\Delta\mathbf{x}_n^*\|^2 \;\geq\; \frac{2k_\mathrm{B}T\tau_0}{\tau}\max_{1\leq m\leq L} m^2\,D_\mathrm{KL}^{(m)}. \tag{S26}$$

Every step is a relaxation, so the transport bound is never weaker; the endpoint evaluation assumes only piecewise linearity within activation sectors, with the path-dependent Fisher length as the rigorous object across sector changes, as discussed below. The fidelity of the main text is defined by $\Gamma_\mathrm{inf} \equiv (2/L^2 d)\max_m m^2 D_\mathrm{KL}^{(m)}$ with d the network width, which makes Eq. (S26) identical to $E_\mathrm{inf} \geq dk_\mathrm{B}T\Gamma_\mathrm{inf}\tau_L/\tau$ by construction.

Two remarks. First, on where the maximum sits. Under gain normalization $\mathcal{C}_m \simeq m\,I$, so $D_\mathrm{KL}^{(m)} \simeq (\kappa/2k_\mathrm{B}T)\|\Delta\mathbf{x}_m\|^2/m$ and $m^2 D_\mathrm{KL}^{(m)} \propto m\|\Delta\mathbf{x}_m\|^2$, which increases with depth when per-layer displacements are comparable: the accumulated noise grows only linearly in m while the counting factor grows quadratically, so the maximum is generically attained at or near the output layer m^*, and the counting factor $(m^*/L)^2$ implicit in Γ_inf is of order unity. The maximization is retained rather than fixed at $m = L$ because this reasoning fails when the displacement is concentrated upstream, or when the output width is small and unrepresentative of the network, in which case an interior layer gives the stronger bound.

Second, the entropic bound can equivalently be obtained through a thermodynamic-length argument: defining $\mathcal{L}_m = \int_0^\tau[\dot{\mathbf{x}}_m^T C_m^{-1}\dot{\mathbf{x}}_m]^{1/2}\mathrm{d}t$ and applying Cauchy–Schwarz in time gives $\int\dot{\mathbf{x}}_m^T C_m^{-1}\dot{\mathbf{x}}_m\mathrm{d}t \geq \mathcal{L}_m^2/\tau$; because the length of any path in a Riemannian metric is at least the geodesic distance between its endpoints, and for a fixed-covariance Gaussian family that distance is the Mahalanobis separation, $\mathcal{L}_m^2 \geq 2D_\mathrm{KL}^{(m)}$. The result is a genuine inequality, not a constant-speed estimate. For finite displacements the endpoint form applies as stated within a fixed activation pattern, where the network is linear and C_m is constant along the path; across pattern changes the path-dependent length is the rigorous object and the endpoint Kullback–Leibler form is its fixed-covariance evaluation. For the two-layer model simulated in the main text the interlayer map is linear and the covariances are input independent, so the endpoint form is exact.

Finally, the factor L^2 in Eq. (S26) and the factor L^2 in $\tau_{\rm inf}$ have different mathematical origins in the same bidiagonal structure, layer counting and the singular-value spectrum of A respectively, but both are controlled by a single physical condition: layer gains of order unity. That common condition is what makes $\tau_{\rm inf} \simeq \tau_L$, and when it fails both fail together, though differently on the two sides of unit gain: for $g > 1$ a single amplifying layer dominates the sum and $\lambda_{\min}(A^T A)$ collapses exponentially, while for $g < 1$ the relaxation time saturates and it is the signal that attenuates, as detailed below.

3.4 Tightness and dynamical isometry

Equality in Eq. (S24) requires $\Delta\mathbf{x}_\ell \propto \widehat{G}^T_{\ell\to m}\mathcal{C}_m^{-1}\Delta\mathbf{x}_m$ for every $\ell \le m$ with a common constant. If the partial Green functions are mutually orthogonal, $\widehat{G}_{\ell\to m}\widehat{G}^T_{\ell\to m} = I$ for all ℓ, then $\mathcal{C}_m = mI$, all $\|\Delta\mathbf{x}_\ell\|$ are equal, and Eq. (S24) is saturated. This is the condition of dynamical isometry [10, 11].

Away from isometry, the two-sided operator bound $\lambda^{-1}_{\max}(\mathcal{C}_m)\, I \preceq \mathcal{C}_m^{-1} \preceq \lambda^{-1}_{\min}(\mathcal{C}_m)\, I$ gives

$$\lambda_{\min}(\mathcal{C}_m)\,\frac{k_{\rm B}T}{\kappa}\,2D^{(m)}_{\rm KL} \;\le\; \|\Delta\mathbf{x}_m\|^2 \;\le\; \lambda_{\max}(\mathcal{C}_m)\,\frac{k_{\rm B}T}{\kappa}\,2D^{(m)}_{\rm KL}. \tag{S27}$$

Equation (S27) is the Gaussian instance of the Talagrand-type transport–entropy inequality [12]: transport cost and relative entropy agree up to the fluctuation-amplification factor of the network. Since $\lambda_{\max}(\mathcal{C}_m)$ grows with depth whenever layer gains exceed unity, the entropic bound degrades in exactly the regime in which the physical cost grows.

For reference, in the well-conditioned diffusive case $\mathcal{C}_m \sim mI$ with comparable per-layer displacements, $\sum_n \|\Delta\mathbf{x}^*_n\|^2 \sim L\|\Delta\mathbf{x}\|^2$ and $\max_m m^2 \cdot 2D^{(m)}_{\rm KL} \sim L(\kappa/k_{\rm B}T)\|\Delta\mathbf{x}\|^2$: the two bounds coincide up to order-unity factors, and the main-text scaling $E_{\rm inf} \gtrsim k_{\rm B}T d\Gamma_{\rm inf}\tau_L/\tau$ is unchanged. What changes is its logical status and its domain of validity.

3.5 Bures term and activation switching

In a piecewise-linear network the covariance changes between two inputs only through the activation pattern, and only through nonlinearities that sit *between* physical layers, where they propagate thermal fluctuations via the Jacobians $J_n = \mathrm{diag}[f_n'(\mathbf{z}_n)]W_n$. A nonlinearity applied to a clamped variable propagates nothing: in the two-layer model simulated in the main text, the rectifier acts on the clamped input, so the output covariance is exactly $C_y = (k_{\rm B}T/\kappa)(I + W_2W_2^T)$, independent of the input, and the Bures term vanishes identically. The transport bound for that model therefore reduces to its mean-shift part with no relaxation at all.

For an architecture in which a rectified layer separates two physical layers, switching hidden unit j toggles the corresponding entry of $\mathrm{diag}[f']$ and shifts the downstream covariance by the rank-one amount $\pm(k_{\rm B}T/\kappa)\,\mathbf{w}_j\mathbf{w}_j^T$, with $\mathbf{w}_j$ the j-th column of the downstream weight matrix. For a rank-one perturbation comparable to the covariance scale, the Bures term in Eq. (S19) is of order $(k_{\rm B}T/\kappa)\|\mathbf{w}_j\|^2$, so through Eq. (S17) each switched unit contributes

$$\Delta E_{\rm Bures} \sim k_{\rm B}T\,\frac{\tau_0}{\tau}\,\|\mathbf{w}_j\|^2 \tag{S28}$$

to the minimum dissipation. This is a finite-rate analogue of Landauer's cost — of order $k_{\rm B}T$ per reconfigured degree of freedom — which vanishes as $\tau \to \infty$ and is invisible to the mean-shift entropic treatment.

Dropping $\mathcal{B}^2$ from the lower bound requires no smallness assumption, since the term is non-negative and its removal is a further relaxation. Its size nevertheless controls how much is given

away. The Bures term is a thermal quantity, responding only to changes of the covariance, which scales as $k_{\mathrm{B}}T/\kappa$; the mean-shift term is a signal quantity of order $\|\Delta\mathbf{x}^*\|^2$. With $\|\Delta\mathbf{x}_L\|^2 \sim d\,L\,\Gamma_{\mathrm{inf}}k_{\mathrm{B}}T/\kappa$ from the definition of the fidelity, and at most d switched units contributing $\sim k_{\mathrm{B}}T/\kappa$ each,

$$\frac{\mathcal{B}^2}{\sum_n \|\Delta\mathbf{x}_n^*\|^2} \lesssim \frac{1}{L^2\,\Gamma_{\mathrm{inf}}}: \tag{S29}$$

the discarded term is suppressed by the very fidelity that must be large for the network to compute reliably. The estimate uses an output-layer displacement and therefore applies in the gain-normalized regime where the binding layer sits at or near the output, $m^* \simeq L$ and $d_{m^*} \simeq d$; the term vanishes identically for a linear network or for any transition that leaves the activation pattern unchanged, and exactly for the architecture simulated in the main text.

3.6 Saturation under input-only control

The Benamou–Brenier optimum assumes free control of the driving potential, whereas here only the input $\mathbf{u}$ is controlled. For a linear network, $\mathbf{x}^*$ is affine in $\mathbf{u}$ and the covariance is $\mathbf{u}$-independent, so the Bures term vanishes and linear interpolation $\mathbf{u}(t) = (1-s)\mathbf{u}_i + s\mathbf{u}_{i+1}$, $s = t/\tau$, produces constant-velocity motion of the mean, which is the McCann displacement interpolation. In the slow-driving limit, therefore, linear input switching realizes the optimal transport path and Eq. (S17) is saturated to leading order in τ_0/τ. For nonlinear networks the accessible optimum is the geodesic length of the path constrained to the curved zero-energy manifold, which exceeds the unconstrained Wasserstein chord; the difference is a curvature cost of the architecture and is not evaluated here.

4 Depth-controlled relaxation time

Let $\boldsymbol{\eta}_n = \mathbf{x}_n - \mathbf{x}_n^*$ be the lag from the instantaneous feedforward state at fixed input, with $\boldsymbol{\eta}_0 = 0$ (the symbol $\mathbf{u}$ is reserved for the input throughout). Linearizing the residuals gives $\boldsymbol{\epsilon}_n = \boldsymbol{\eta}_n - J_n\boldsymbol{\eta}_{n-1}$. Collecting the hidden and output lags into a vector $\boldsymbol{\eta}$,

$$\boldsymbol{\epsilon} = A\boldsymbol{\eta}, \qquad \mathcal{H} = \frac{\kappa}{2}\boldsymbol{\eta}^T A^T A\boldsymbol{\eta}, \tag{S30}$$

with A block lower-bidiagonal and unit diagonal blocks. The fixed-input overdamped dynamics is $\dot{\boldsymbol{\eta}} = -\mu\kappa A^T A\boldsymbol{\eta}$, so the slowest relaxation time is exactly

$$\tau_{\mathrm{inf}} = \frac{1}{\mu\kappa\,\lambda_{\min}(A^T A)}. \tag{S31}$$

For a path-like, well-conditioned network whose layer Jacobians have singular values of order unity and generate no additional soft modes, the smallest singular value of A scales as L^{-1}, whence

$$\lambda_{\min}(A^T A) \sim L^{-2}, \qquad \tau_{\mathrm{inf}} \simeq \tau_L \equiv L^2\tau_0. \tag{S32}$$

The two sides of unit gain fail differently. For $g > 1$ the product structure of A^{-1} gives $\lambda_{\min}(A^T A) \sim g^{-2(L-1)}$, and τ_{inf} grows exponentially beyond τ_L: the network becomes exponentially slow. For $g < 1$ the relaxation time instead saturates at a finite value, $\tau_{\mathrm{inf}} \to \tau_0/(1-g)^2$ as $L \to \infty$; what degrades is not the timescale but the signal, whose displacement attenuates as g^L down the chain, collapsing the fidelity Γ_{inf}. The diffusive law $\tau_{\mathrm{inf}} \simeq \tau_L$ thus characterizes the critical, near-isometric chain $g \to 1$, which is also the condition that makes Eq. (S24) tight: gain normalization is the operating point at which neither the timescale nor the signal deteriorates with depth. In a trained or strongly nonlinear network the exact Eq. (S31) defines the relevant relaxation time, while τ_L remains the natural normalization scale.

4.1 Relaxation time of the simulated network

For the two-layer architecture of the main text the spectrum of A^TA can be written in closed form, which fixes the ratio $\tau_{\rm inf}/\tau_L$ for the trained model without any fitting. With physical layers $\mathbf{h} \in \mathbb{R}^{d_1}$ and $\mathbf{y} \in \mathbb{R}^{d_2}$ and the output map linear, the compatibility operator and its Gram matrix are

$$A = \begin{pmatrix} I_{d_1} & 0 \\ -M & I_{d_2} \end{pmatrix}, \qquad A^TA = \begin{pmatrix} I_{d_1} + M^TM & -M^T \\ -M & I_{d_2} \end{pmatrix}, \qquad M \equiv W_2^T, \tag{S33}$$

where the rectified nonlinearity acts on the clamped input and so contributes no coupling between the fluctuating coordinates. Let $M = U\Sigma V^T$ be the singular value decomposition, with singular values s_i, $i = 1, \ldots, d_2$. In the basis that pairs the i-th right singular vector of M with its i-th left singular vector, A^TA becomes block diagonal with 2×2 blocks

$$\begin{pmatrix} 1 + s_i^2 & -s_i \\ -s_i & 1 \end{pmatrix}, \qquad \lambda_i^{\pm} = \frac{2 + s_i^2 \pm s_i\sqrt{s_i^2 + 4}}{2}, \tag{S34}$$

together with $d_1 - d_2$ directions in the hidden layer orthogonal to the row space of M, each contributing $\lambda = 1$. Since $\det A = 1$, each pair satisfies $\lambda_i^+\lambda_i^- = 1$, so every singular value of M produces one stiff and one soft mode, reciprocal to each other; the softest mode belongs to the largest singular value, and

$$\lambda_{\min}(A^TA) = \frac{2 + \sigma_{\max}^2 - \sigma_{\max}\sqrt{\sigma_{\max}^2 + 4}}{2}, \qquad \frac{\tau_{\rm inf}}{\tau_L} = \frac{1}{L^2\,\lambda_{\min}(A^TA)}, \tag{S35}$$

with $\sigma_{\max} = \sigma_{\max}(W_2)$ and $L = 2$. The ratio grows with the output gain: $\tau_{\rm inf}/\tau_L = 0.65$ at $\sigma_{\max} = 1$, 1.5 at $\sigma_{\max} = 2$ and 2.7 at $\sigma_{\max} = 3$, remaining of order unity over the whole range spanned by training.

For the trained model used in Figs. 2 and 3, $\sigma_{\max}(W_2) = 3.21$, so that Eq. (S35) gives $\lambda_{\min}(A^TA) = 0.0816$ and

$$\tau_{\rm inf} = 12.2\,\tau_0 = 3.1\,\tau_L, \tag{S36}$$

in exact agreement with direct numerical diagonalization of A^TA. The exact relaxation time is therefore an order-unity multiple of the diffusive scale, as the gain-normalized estimate predicts. All time axes in the main text are normalized by τ_L, which is the theoretically defined quantity; the measured $\tau_{\rm inf}$ enters only through the ratio of Eq. (S36), and no qualitative feature depends on the distinction.

The full spectrum is shown in Fig. 1. Its structure is that of Eq. (S34): ten stiff modes, ten soft modes and, between them, the $d_1 - d_2 = 10$ hidden directions orthogonal to the row space of M, which relax at exactly τ_0. Because $\lambda_i^+\lambda_i^- = 1$, the relaxation times are mirror-symmetric about τ_0 on a logarithmic scale, each soft mode being the exact reciprocal of a stiff partner; numerically $\tau_{\max}\tau_{\min} = \tau_0^2$ to machine precision. Two features are worth noting. First, the slow modes form a band rather than an isolated eigenvalue: eight of the thirty modes are slower than τ_L, spanning 3.7 to $12.2\,\tau_0$, and the slowest exceeds the second slowest by only 12%. The relaxation is thus genuinely multi-exponential, and $\tau_{\rm inf}$ should be read as the conservative representative of that band, the time after which every mode has relaxed, rather than as the single dominant timescale. Second, the band lies above τ_L rather than straddling it, which is why $\tau_{\rm inf}/\tau_L > 1$ for the trained model: training increases the output gain, and by Eq. (S35) the softest mode slows accordingly.

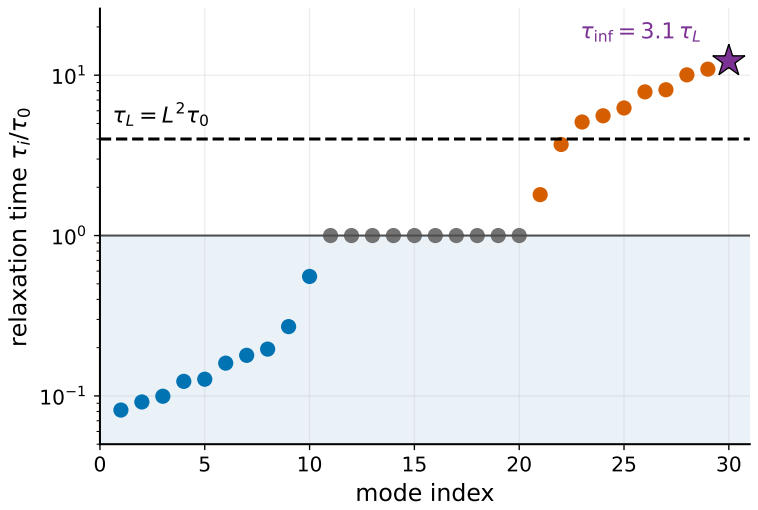

Figure 1: Relaxation spectrum of the trained network, $\tau_i = 1/[\mu\kappa\lambda_i(A^TA)]$ in units of τ_0, ordered from fastest to slowest. Stiff modes (blue) and soft modes (orange) form exact reciprocal pairs about τ_0, one pair per singular value of W_2, while the $d_1 - d_2 = 10$ hidden directions orthogonal to the row space of W_2^T relax at exactly τ_0 (grey). Dashed: the diffusive scale $\tau_L = L^2\tau_0$. Dotted: the slowest relaxation time $\tau_{\rm inf} = 3.1\,\tau_L$, which caps a band of eight modes slower than τ_L.

5 Constrained free energy and backpropagation as force balance

For a training example $\alpha = (\mathbf{u}_\alpha, \mathbf{t}_\alpha)$, impose $\mathbf{x}_0 = \mathbf{u}_\alpha$ and $\mathbf{x}_L = \mathbf{t}_\alpha$. The data-constrained free energy is

$$F_\alpha(\boldsymbol{\theta}) = -k_{\rm B}T \ln \int_{\mathbf{x}_0=\mathbf{u}_\alpha,\,\mathbf{x}_L=\mathbf{t}_\alpha} \exp[-\beta\mathcal{H}(\mathbf{x};\boldsymbol{\theta})] \prod_{n=1}^{L-1} \mathrm{d}\mathbf{x}_n, \tag{S37}$$

reducing at low temperature to the minimum constrained energy. At finite temperature the envelope theorem gives

$$\frac{\partial F_\alpha}{\partial \theta_k} = \left\langle \frac{\partial \mathcal{H}}{\partial \theta_k} \right\rangle_\alpha, \tag{S38}$$

the average being over the constrained equilibrium distribution.

Define the stress

$$\boldsymbol{\sigma}_n = \kappa[f_n(\mathbf{z}_n) - \mathbf{x}_n] \circ f_n'(\mathbf{z}_n) = -\kappa\,\boldsymbol{\epsilon}_n \circ f_n'(\mathbf{z}_n). \tag{S39}$$

Differentiating Eq. (S2) and using the envelope identity gives the exact finite-temperature gradients

$$\frac{\partial F_\alpha}{\partial \mathbf{b}_n} = \langle \boldsymbol{\sigma}_n \rangle_\alpha\,, \qquad \frac{\partial F_\alpha}{\partial W_n} = \langle \boldsymbol{\sigma}_n \mathbf{x}_{n-1}^T \rangle_\alpha\,, \tag{S40}$$

where $\langle\cdots\rangle_\alpha$ is the constrained equilibrium average. Equations (S40) are averages of products, not products of averages; the recursion below is the mechanical equilibrium condition holding pointwise at the constrained minimum, which is the low-temperature form of Eq. (S40), with the difference between $\langle\boldsymbol{\sigma}_n\rangle_\alpha$ and $\boldsymbol{\sigma}_n$ at the minimum given by the entropic term of Sec. 2.2. The hidden-layer force-balance condition is

$$0 = \frac{\partial \mathcal{H}}{\partial \mathbf{x}_n} = \kappa\boldsymbol{\epsilon}_n + W_{n+1}^T\boldsymbol{\sigma}_{n+1}, \qquad n = 1, \ldots, L-1, \tag{S41}$$

and multiplying componentwise by $-f_n'(\mathbf{z}_n)$ gives the adjoint recursion

$$\boldsymbol{\sigma}_n = (W_{n+1}^T \boldsymbol{\sigma}_{n+1}) \circ f_n'(\mathbf{z}_n), \qquad n = L-1, \ldots, 1, \tag{S42}$$

with output-layer stress $\boldsymbol{\sigma}_L = \kappa[f_L(\mathbf{z}_L) - \mathbf{t}_\alpha] \circ f_L'(\mathbf{z}_L)$. Equations (S40) and (S42) identify the backpropagated error with an internal mechanical stress in the constrained network.

6 Residual parameter-annealing dissipation

Let $\overline{F}(\boldsymbol{\theta}) = \langle F_\alpha(\boldsymbol{\theta})\rangle_\alpha$, $\mathbf{g}_\alpha = \nabla_\theta F_\alpha$ and $\overline{\mathbf{g}} = \langle \mathbf{g}_\alpha\rangle_\alpha = \nabla_\theta \overline{F}$. For deterministic annealing under example α, $\dot{\boldsymbol{\theta}} = -M_\theta \mathbf{g}_\alpha$ with positive mobility matrix M_θ, the instantaneous dissipation rate is $\mathbf{g}_\alpha^T M_\theta \mathbf{g}_\alpha$, so

$$\left\langle \dot{Q}_\theta \right\rangle_\alpha = \left\langle \mathbf{g}_\alpha^T M_\theta \mathbf{g}_\alpha \right\rangle_\alpha, \qquad -\frac{\mathrm{d}\overline{F}}{\mathrm{d}t} = \overline{\mathbf{g}}^T M_\theta \overline{\mathbf{g}}. \tag{S43}$$

With the mobility-weighted variance $\mathrm{Var}_{M_\theta}(\mathbf{g}) = \langle \mathbf{g}_\alpha^T M_\theta \mathbf{g}_\alpha\rangle_\alpha - \overline{\mathbf{g}}^T M_\theta \overline{\mathbf{g}}$ and the incoherence ratio

$$\mathcal{R}_\theta = \frac{\mathrm{Var}_{M_\theta}(\mathbf{g})}{\overline{\mathbf{g}}^T M_\theta \overline{\mathbf{g}}}, \tag{S44}$$

one obtains $\mathrm{d}Q_\theta = (1+\mathcal{R}_\theta)\mathrm{d}F_{\mathrm{drop}}$ with $\mathrm{d}F_{\mathrm{drop}} = -\mathrm{d}\overline{F} > 0$, and for a finite trajectory

$$E_\theta \simeq \int_0^{\Delta\overline{F}} [1 + \mathcal{R}_\theta(F)]\, \mathrm{d}F_{\mathrm{drop}} \;\sim\; (1 + \overline{\mathcal{R}}_\theta)\, \Delta\overline{F}. \tag{S45}$$

In numerical comparisons a single order-one prefactor C accounts for protocol discretization and windowing.

The numerical comparison of Fig. 2 bears this out: the measured slow-driving work of the free protocol saturates the transport bound computed from the free feedforward endpoints to within a few percent. The clamped protocol is excluded from that comparison for a structural reason, not a numerical one: clamping changes the endpoint states themselves, from free equilibria to constrained ones, so the relevant transport distance is $\sum_n \|\Delta\mathbf{x}_n^c\|^2$ evaluated between clamped equilibria, with the output displacement running between one-hot targets. That distance is smaller than the free one, and the clamped work correspondingly sits below the free-endpoint bound without any inconsistency.

6.1 Thermal resolvability of the learned state

Near a minimum $\boldsymbol{\theta}^*$ of the averaged loss, $\overline{F} \simeq \overline{F}^* + \frac{1}{2}\delta\boldsymbol{\theta}^T H \delta\boldsymbol{\theta}$ with Hessian $H \succ 0$. At temperature T the equilibrium parameter fluctuations satisfy equipartition, $\langle \delta\boldsymbol{\theta}^T H \delta\boldsymbol{\theta}\rangle = N_\theta k_{\mathrm{B}} T$: the landscape is thermally blurred over an energy range $N_\theta k_{\mathrm{B}} T$, independently of the curvature spectrum. A learned state is resolvable against this blur only if the achieved decrease of the averaged free energy exceeds it, giving the necessary condition

$$\Gamma_{\mathrm{train}} = \frac{\Delta\overline{F}}{N_\theta k_{\mathrm{B}} T} \;\gtrsim\; 1, \tag{S46}$$

and, combining with Eq. (S45),

$$E_\theta \gtrsim \Gamma_{\mathrm{train}} (1 + \overline{\mathcal{R}}_\theta) N_\theta k_{\mathrm{B}} T. \tag{S47}$$

Equation (S46) is a consistency condition for a high-dimensional noisy parameter space, not a theorem about arbitrary landscapes, and it is necessary rather than sufficient: the accuracy actually attainable at a given Γ is task dependent and is quantified empirically in Sec. 8.8. Note also that if only $N_{\mathrm{eff}} < N_\theta$ directions are stiff enough to matter, the threshold in terms of Γ_{train} as defined with the nominal N_θ is correspondingly lower, so Eq. (S46) is conservative.

6.2 Combined bound

Adding the finite-rate data-presentation contribution, and using Eq. (S17) for each presentation,

$$E_{\mathrm{train}} \gtrsim \frac{\kappa\tau_0}{\tau_{\mathrm{feed}}} \sum_{p=1}^{N_{\mathrm{pres}}} \left[\mathcal{W}_2^2\right]_p + \Gamma_{\mathrm{train}}(1+\overline{\mathcal{R}}_\theta) N_\theta k_{\mathrm{B}} T \;\gtrsim\; k_{\mathrm{B}} T \left[N_{\mathrm{pres}}\, d\, \Gamma_{\mathrm{inf}} \frac{\tau_L}{\tau_{\mathrm{feed}}} + \Gamma_{\mathrm{train}}(1+\overline{\mathcal{R}}_\theta) N_\theta \right]. \tag{S48}$$

Only the first term carries a protocol time; the second is irreducible at any speed within the passive parameter dynamics considered. Dataset size does not appear explicitly: it enters the second term through the accumulated incoherence $\overline{\mathcal{R}}_\theta$, which grows as examples are presented in sequence.

7 Evaluation of the estimates in Table 1

All entries of Table 1 of the main text are order-of-magnitude estimates at $T = 300$ K, where $k_{\mathrm{B}}T = 4.1 \times 10^{-21}$ J and one irreversible 16-bit operation carries the Landauer benchmark $\Gamma_{\mathrm{dig}} k_{\mathrm{B}} T \ln 2 = 4.6 \times 10^{-20}$ J with $\Gamma_{\mathrm{dig}} = 16$. The arithmetic below is quoted to one or two significant figures; only the order of magnitude is meaningful, and that is what the table reports.

Landauer benchmark column. Autoregressive inference performs $2N_a$ operations per token, with N_a the number of parameters active per token [13]. For the dense 65-billion-parameter model, $2N_a = 1.3 \times 10^{11}$ operations per token, giving $1.3 \times 10^{11} \times 4.6 \times 10^{-20} \approx 6 \times 10^{-9}$ J per token. Training performs $6N_a$ operations per token; for the mixture-of-experts model with $N_a \approx 3.7 \times 10^{10}$ active parameters [14], this gives $2.2 \times 10^{11} \times 4.6 \times 10^{-20} \approx 1 \times 10^{-8}$ J per token.

Current hardware column. For inference, $2N_a = 1.3 \times 10^{11}$ operations per token executed on a contemporary accelerator: peak efficiency is of order 10^{-12} J per 16-bit operation (several hundred watts against a few hundred teraflops), and autoregressive decoding is memory-bandwidth bound with utilization of a few per cent, giving an effective $\sim 3 \times 10^{-11}$ J per operation and hence a few joules per token; we quote 4 J. For training, the mixture-of-experts run of Ref. [14] reports $\approx 2.8 \times 10^6$ accelerator-hours for 1.5×10^{13} tokens; at ~ 400 W per accelerator this is $\approx 4 \times 10^{12}$ J, or ≈ 0.2 J per token.

Slow limit column. For inference the bound is exactly zero, Eq. (4) of the main text. For training, only the parameter-annealing term survives, $\Gamma_{\mathrm{train}}(1+\overline{\mathcal{R}}_\theta) N_\theta k_{\mathrm{B}} T \approx N_\theta k_{\mathrm{B}} T$ for $\Gamma_{\mathrm{train}}(1+\overline{\mathcal{R}}_\theta) \sim 1$; with $N_\theta = 6.7 \times 10^{11}$ this one-time cost of 2.8×10^{-9} J, amortized over 1.5×10^{13} tokens, gives $\approx 2 \times 10^{-22}$ J per token.

Fastest-useful-speed column. At $\tau \simeq \tau_{\mathrm{inf}}$ the fidelity cancels against the relaxation time it also sets, and the cost reduces to $d\, k_{\mathrm{B}} T$ per inference event. Taking $d \sim 10^6$ for the number of coordinates that move per token gives $10^6 \times 4.1 \times 10^{-21} \approx 4 \times 10^{-15}$ J, quoted as 10^{-14} J in the table. This is conservative in two respects: Eq. (13) takes the maximum over layers rather than the sum, so the derived bound refers to the widest layer alone rather than to all activations, and $\tau_{\mathrm{inf}} \sim \Gamma_{\mathrm{inf}} \tau_L$ is an estimate of the fastest usable speed rather than a bound. The same figure applies to training at maximal presentation rate, where the data-presentation term of Eq. (26) dominates the annealing term by many orders of magnitude.

Physics-based hardware calibration. The per-cell figure of 2 fJ per conditional update quoted in the Discussion [15] corresponds to $2 \times 10^{-15}/(\Gamma_{\rm inf} k_{\rm B}T) \approx 5 \times 10^4$ per-coordinate bounds at $\Gamma_{\rm inf} \sim 10$, against the $\sim 10^7$ by which digital accelerators exceed their own Landauer benchmark (5×10^{-13} J measured against 4.6×10^{-20} J per operation).

Digital columns. The Landauer entries follow Eqs. (37) and (38) of the Methods. For the dense model, $2N_\theta = 1.3 \times 10^{11}$ operations per token gives $1.3 \times 10^{11} \times 4.6 \times 10^{-20} \approx 6 \times 10^{-9}$ J, quoted as 10^{-8} J. For the mixture-of-experts model, $6N_\theta = 2.2 \times 10^{11}$ operations per token gives $\approx 1 \times 10^{-8}$ J per token; over $D = 1.5 \times 10^{13}$ presentations the corresponding total is 1.5×10^5 J, against a physical training cost of order $N_\theta k_{\rm B}T \approx 3 \times 10^{-9}$ J incurred once. The ratio of the two is dominated by D rather than by any prefactor, which is the point made in the Discussion.

8 Numerical implementation

8.1 Data set and architecture

The numerical tests used the handwritten-digits data set from `scikit-learn`, with pixel values divided by 16 so that all input components lay in $[0, 1]$, and a stratified train–test split with test fraction 0.35 and fixed random seed, giving 1168 training and 629 test examples. The architecture was $64 \to 20 \to 10$ with rectified-linear hidden layer and linear output layer, and one-hot targets. With $\mathbf{a}(\mathbf{u}) = \mathrm{ReLU}(W_1\mathbf{u} + \mathbf{b}_1)$, the free-inference and clamped energies were

$$\mathcal{H} = \tfrac{1}{2}\|\mathbf{h} - \mathbf{a}(\mathbf{u})\|^2 + \tfrac{1}{2}\|\mathbf{y} - W_2\mathbf{h} - \mathbf{b}_2\|^2, \qquad \mathcal{H}_{\rm cl} = \tfrac{1}{2}\|\mathbf{h} - \mathbf{a}(\mathbf{u})\|^2 + \tfrac{1}{2}\|\mathbf{t} - W_2\mathbf{h} - \mathbf{b}_2\|^2,$$

and the number of trainable parameters was $N_\theta = 20 \times 64 + 20 + 10 \times 20 + 10 = 1510$.

8.2 Inference dynamics and work

With mobility set to unity in simulation units,

$$\dot{\mathbf{h}} = -(\mathbf{h} - \mathbf{a}) + W_2^T(\mathbf{y} - W_2\mathbf{h} - \mathbf{b}_2), \qquad \dot{\mathbf{y}} = -(\mathbf{y} - W_2\mathbf{h} - \mathbf{b}_2),$$

with the input switched linearly, $\mathbf{u}(t) = (1 - s)\mathbf{u}_i + s\mathbf{u}_{i+1}$, $s = t/\tau$. The input-control work was

$$E_{\rm inf} = \int \frac{\partial \mathcal{H}}{\partial \mathbf{u}} \cdot \mathrm{d}\mathbf{u}, \qquad \frac{\partial \mathcal{H}}{\partial \mathbf{u}} = -W_1^T\left[(\mathbf{h} - \mathbf{a}) \circ \mathrm{ReLU}'(W_1\mathbf{u} + \mathbf{b}_1)\right],$$

normalized by the hidden-state jump scale $E_0 = \frac{1}{2}\langle\|\mathbf{a}(\mathbf{u}_i) - \mathbf{a}(\mathbf{u}_{i-1})\|^2\rangle_i$. Since $\mathcal{H}$ vanishes at both ends of a completed transition, this control work equals the dissipated heat; at $\tau \ll \tau_L$ the transition is not completed within the protocol window and the measured control work falls below the asymptotic τ^{-1} branch, which is the regime in which classification accuracy also collapses. All time axes are normalized by $\tau_L = L^2\tau_0 = 4\tau_0$ for the $L = 2$ physical layers simulated; the exact relaxation time of the trained model is $\tau_{\rm inf} = 3.1\,\tau_L$ (Sec. 4.1). Protocol times $\tau/\tau_L \in [0.05, 25]$ were scanned with $\max(20, \lceil \tau/0.1 \rceil)$ integration steps, the physical state persisting across transitions and the decision read from $\mathbf{y}$ at the end of each switching window. Finite-temperature accuracies were averaged over five noise realizations, and the thermal collapse variable was $E_{\rm inf}/E_T(\tau)$ with $E_T(\tau) = k_{\rm B}T d_h\,\tau_L/\tau$ and $d_h = 20$ the hidden width used as normalization; the output covariance of this architecture is exactly $C_y = (k_{\rm B}T/\kappa)(I + W_2W_2^T)$; the output layer has $d_2 = 10$, while the network width entering the fidelity definition of the main text is $d = \max_m d_m = 20$.

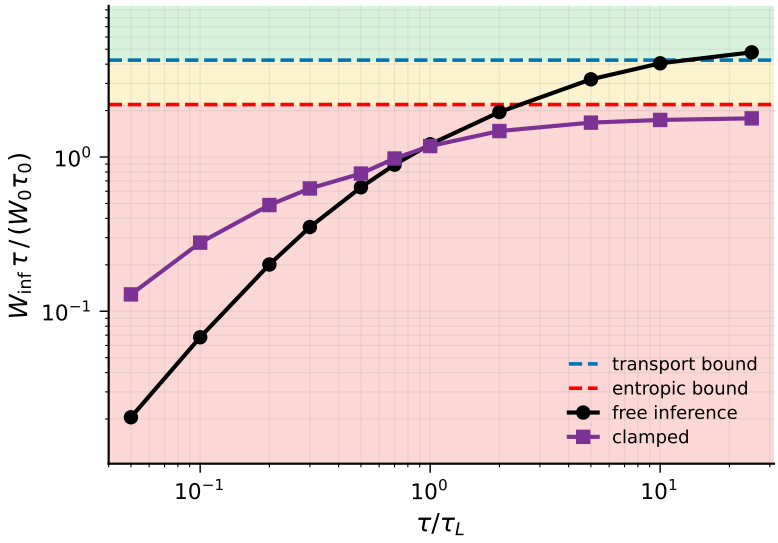

Figure 2: Measured inference work times protocol duration at $T = 0$, for the free (circles) and clamped (squares) protocols, against the transition-averaged transport bound (dashed) and its entropic relaxation (dotted). The free protocol saturates the transport bound to 3.9% on the slow branch; the clamped protocol transitions between clamped equilibria, whose transport distance is smaller, and is therefore not bounded by the dashed line. Work below the lines at $\tau \lesssim \tau_L$ reflects incomplete transitions.

8.3 Evaluation of the transport and entropic bounds

For this architecture the endpoint covariances coincide, so the Bures term vanishes analytically and the transport bound is exactly its mean-shift part, evaluated as $\sum_n \|\Delta \mathbf{x}_n^*\|^2$ from the feedforward states at the endpoints; the information-geometric relaxation follows from Eq. (S26) with the full-rank $\mathcal{C}_L$ of Eq. (S6).

8.4 Numerical test of the transport and entropic bounds

For each of the 628 test-set transitions of the trained classifier, the transport bound was evaluated as $\sum_n \|\Delta \mathbf{x}_n^*\|^2$ from the free feedforward endpoints, and the entropic bound as $\max_m m^2 \Delta \mathbf{x}_m^T \mathcal{C}_m^{-1} \Delta \mathbf{x}_m$ with $\mathcal{C}_1 = I$ and $\mathcal{C}_2 = I + W_2 W_2^T$. Both are pure geometry of the trained model, the temperature cancelling. Their transition-averaged values are 2.55 and 1.32 in units of $\tau_0 W_0$, a ratio of 0.52, and the output layer is the binding one in 44% of transitions, the hidden layer in the remainder, which is why the maximum over layers is retained rather than fixed at the output.

Figure 2 compares both with the measured work. On the slow-driving branch the free-inference $E_{\mathrm{inf}}\tau$ exceeds the transport bound by only 3.9%: input-only control saturates the optimal-transport minimum for this architecture, as the linear-response argument above predicts, and the entropic relaxation gives away a factor of about two, the price of trading motion for distinguishability. For $\tau \lesssim \tau_L$ the measured work falls below both lines because the transitions are not completed, the regime already excluded by the bounds' derivation. The clamped curve sits below the free-endpoint bound for the structural reason given above: its endpoints are clamped equilibria, with a strictly smaller transport distance, so no inequality is violated. Extending the comparison across training checkpoints, where $\lambda_{\max}(\mathcal{C}_L)$ grows as the output weights strengthen, remains a natural further test.

8.5 Training dynamics

Both input and target were switched linearly between consecutive examples. Under the clamped target,

$$\dot{\mathbf{h}} = -(\mathbf{h} - \mathbf{a}) + W_2^T \mathbf{r}_2, \qquad \mathbf{r}_2 = \mathbf{t} - W_2\mathbf{h} - \mathbf{b}_2,$$

and with $\mathbf{r}_1 = (\mathbf{h} - \mathbf{a}) \circ \mathrm{ReLU}'(W_1\mathbf{u} + \mathbf{b}_1)$,

$$\dot{W}_1 = \mu_\theta \mathbf{r}_1 \mathbf{u}^T, \quad \dot{\mathbf{b}}_1 = \mu_\theta \mathbf{r}_1, \quad \dot{W}_2 = \mu_\theta \mathbf{r}_2 \mathbf{h}^T, \quad \dot{\mathbf{b}}_2 = \mu_\theta \mathbf{r}_2.$$

The representative simulations used $\tau_{\mathrm{train}} = 20\tau_0$, $\Delta t = 0.05\tau_0$, $\mu_\theta = 0.01/\tau_0$ and 3000 data presentations, with checkpoints every 25 presentations; these are the values recorded in the cached run summary used for Fig. 3.

8.6 Training work and incoherence

The fast-variable contribution was $\Theta_x = \int \|\dot{\mathbf{h}}\|^2 \mathrm{d}t$ and the residual parameter-annealing dissipation $\Theta_\theta = \int \mu_\theta \|\mathbf{g}_\theta\|^2 \mathrm{d}t$, with $\Theta_{\mathrm{tot}} = \Theta_x + \Theta_\theta$. The incoherence was estimated in moving windows of duration $25\tau_0$,

$$\mathcal{R}_\theta = \frac{\left\langle \|\mathbf{g}_\theta\|^2 \right\rangle - \|\langle \mathbf{g}_\theta \rangle\|^2}{\|\langle \mathbf{g}_\theta \rangle\|^2},$$

and the residual prediction tested by fitting $\Theta_\theta \simeq C \int (1 + \mathcal{R}_\theta) \mathrm{d}F_{\mathrm{drop}}$ with a single prefactor C.

8.7 Finite-temperature parameter dynamics

The stochastic parameter dynamics was $\mathrm{d}\boldsymbol{\theta} = \mu_\theta \mathbf{f}_\theta \mathrm{d}t + \sqrt{2\mu_\theta k_{\mathrm{B}} T \mathrm{d}t}\, \boldsymbol{\xi}$ with $\mathbf{f}_\theta = -\nabla_\theta \mathcal{H}_{\mathrm{cl}}$ and $\boldsymbol{\xi}$ standard normal. The stochastic contribution was accumulated over one presentation and applied as an equivalent Gaussian kick of variance $\sigma^2_{\mathrm{pres}} = 2\mu_\theta k_{\mathrm{B}} T \tau_{\mathrm{eff}}$; for stability at larger $k_{\mathrm{B}}T/E_0$, τ_{eff} and the time step were reduced by the same adaptive factor so that the number of steps per presentation was unchanged. This rescaling reduces the physical evolution time per example at high temperature, so in the envelope runs temperature and presentation duration are not varied independently. A control rerun with the presentation time held fixed at $\tau_{\mathrm{eff}} = \tau_{\mathrm{train}}$ for all temperatures (only the integration step refined, five independent seeds at each of $k_{\mathrm{B}}T/E_0 = 10^{-4}$ and 10^{-3}, spanning the adaptive threshold and ten times above) leaves the trajectories within 0.013 accuracy of the adaptive-protocol envelope of Fig. 3c over the crossover range sampled, within the seed-to-seed spread of ± 0.02; the accuracy–Γ_θ relation is therefore a property of the delivered thermodynamic fidelity, not of the presentation duration.

The plotted fidelity was

$$\Gamma_\theta = \frac{\Theta_\theta}{N_\theta (1 + \langle \mathcal{R}_\theta \rangle) k_{\mathrm{B}} T},$$

the dissipative counterpart of Γ_{train}, the two being related through $\mathrm{d}Q_\theta = (1 + \mathcal{R}_\theta)\mathrm{d}F_{\mathrm{drop}}$ up to the fitted prefactor $C = 2.96$ ($R^2 = 0.973$). Trajectories were averaged over independent noise seeds and smoothed with a centred seven-point moving average in presentation number; edge points without full moving-average support were discarded.

8.8 Envelope construction and accuracy scaling

Finite-temperature training data comprise multiple trajectories at different $k_{\mathrm{B}}T/E_0$. To estimate the minimum residual work needed to reach a given accuracy, smoothed test accuracy was plotted against Γ_θ and the upper-left envelope constructed by binning in Γ_θ and taking, in each bin,

the largest smoothed accuracy reached by any trajectory. The envelope is a compact visual summary of the best accuracy attainable at a given thermodynamic training fidelity, not an additional theoretical assumption; all individual trajectories are shown alongside it.

The envelope was described by

$$A_{\mathrm{test}}(\Gamma_\theta) = a_{\max} + A \ln\left(1 + \frac{\Gamma_0}{\Gamma_\theta + 1}\right), \tag{S49}$$

with all three of A, $a_{\max}$ and Γ_0 determined from the data: Γ_0 is scanned on a logarithmic grid and, at each value, A and $a_{\max}$ follow by linear least squares, the grid point with the smallest residual being retained. No parameter is fixed in advance; in particular the low-Γ_θ plateau is not constrained to the ten-class chance level, and the value $a_{\max} + A\ln(1+\Gamma_0) \approx 0.12$ that the fit returns is an outcome rather than an input. The best fit for the run used in Fig. 3c gives

$$A = -0.0911, \qquad \Gamma_0 = 6.15 \times 10^3, \qquad a_{\max} = 0.9105, \qquad R^2 = 0.995. \tag{S50}$$

The coefficient A is negative because accuracy increases with Γ_θ while the logarithm decreases; the main text and the figure annotation quote $|A|$.

The physically meaningful content of Eq. (S49) is its large-Γ_θ limit. Expanding the logarithm,

$$\varepsilon \equiv a_{\max} - A_{\mathrm{test}} \simeq \frac{|A|\Gamma_0}{\Gamma_\theta}, \qquad \text{i.e.} \qquad E_\theta \gtrsim \frac{|A|\Gamma_0}{\varepsilon} N_\theta (1 + \overline{\mathcal{R}}_\theta) k_{\mathrm{B}} T. \tag{S51}$$

The residual accuracy deficit therefore falls inversely with the thermodynamic training fidelity, and the energy needed to reach deficit ε diverges as $1/\varepsilon$ with a task-dependent prefactor $|A|\Gamma_0$. For the present task $|A| \approx 0.09$ and $\Gamma_0 \approx 6 \times 10^3$, giving $|A|\Gamma_0 \approx 5 \times 10^2$: the necessary condition $\Gamma_\theta \gtrsim 1$ is met long before useful accuracy is reached, and high accuracy requires $\Gamma_\theta \sim 10^2$–10^3. This resolves the apparent discrepancy between the threshold $\Gamma_{\mathrm{train}} \gtrsim 1$ and the location of the observed crossover: the former is necessary, the latter is set by the task-dependent prefactor in Eq. (S51).

References

[1] Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000. doi: 10.1007/s002110050002.

[2] Erik Aurell, Carlos Mejía-Monasterio, and Paolo Muratore-Ginanneschi. Optimal protocols and optimal transport in stochastic thermodynamics. *Physical Review Letters*, 106(25):250601, 2011. doi: 10.1103/PhysRevLett.106.250601.

[3] Andreas Dechant and Yohei Sakurai. Thermodynamic interpretation of Wasserstein distance. *arXiv preprint*, arXiv:1912.08405, 2019. URL https://arxiv.org/abs/1912.08405.

[4] Muka Nakazato and Sosuke Ito. Geometrical aspects of entropy production in stochastic thermodynamics based on Wasserstein distance. *Physical Review Research*, 3(4):043093, 2021. doi: 10.1103/PhysRevResearch.3.043093.

[5] Tan Van Vu and Keiji Saito. Thermodynamic unification of optimal transport: thermodynamic uncertainty relation, minimum dissipation, and thermodynamic speed limits. *Physical Review X*, 13(1):011013, 2023. doi: 10.1103/PhysRevX.13.011013.

[6] Naoto Shiraishi, Ken Funo, and Keiji Saito. Speed limit for classical stochastic processes. *Physical Review Letters*, 121(7):070601, 2018. doi: 10.1103/PhysRevLett.121.070601.

[7] Sosuke Ito. Geometric thermodynamics for the Fokker–Planck equation: stochastic thermodynamic links between information geometry and optimal transport. *Information Geometry*, 2023. doi: 10.1007/s41884-023-00102-3.

[8] Clark R. Givens and Rae Michael Shortt. A class of Wasserstein metrics for probability distributions. *Michigan Mathematical Journal*, 31(2):231–240, 1984. doi: 10.1307/mmj/1029003026.

[9] Solomon Kullback and Richard A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. doi: 10.1214/aoms/1177729694.

[10] Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations (ICLR)*, 2014. arXiv:1312.6120.

[11] Jeffrey Pennington, Samuel S. Schoenholz, and Surya Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[12] Felix Otto and Cédric Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000. doi: 10.1006/jfan.1999.3557.

[13] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[14] DeepSeek-AI. Deepseek-v3 technical report. *Preprint:*, arXiv:2412.19437, 2025.

[15] Andraž Jelinčič, Owen Lockwood, Akhil Garlapati, Peter Schillinger, Isaac L. Chuang, Guillaume Verdon, and Trevor McCourt. An efficient probabilistic hardware architecture for diffusion-like models. *npj Unconventional Computing*, 3:30, 2026. doi: 10.1038/s44335-026-00075-3.