

Statistical Mechanical Approaches to Deep Learning

Lecture 1 (Module I, part 1): the approach, the toolbox, and the kernel limit

Zohar Ringel

The Hebrew University of Jerusalem

Fields Institute mini-course, October 2, 2026

Abstract

The first hour sets out the approach of the course and a minimal toolbox: statistical mechanics read as probability theory, Gaussian and path integrals, and Langevin dynamics as a way of sampling Gibbs distributions over weights. The second hour introduces the first universal phenomenon, generalization despite interpolation, and treats it in the limit where it is most puzzling — infinitely wide networks. The equilibrium of noisy training is a Bayesian posterior over weights. A functional delta function moves it to function space, where the weights enter only through a functional prior. At infinite width this prior is a Gaussian process, and the trained network is described by a Gaussian field theory whose predictor is Gaussian-process regression. A curve-fitting example at vanishing ridge shows double descent: the danger zone is a number of parameters close to the number of data points, and adding many more parameters is harmless.

Contents

About these notes	2
Hour 1: The approach and the toolbox	2
1 The approach	2
2 Statistical mechanics as probability theory	3
3 Gaussian integrals and path integrals	6
4 Langevin dynamics and equilibrium	9
Hour 2: Generalization despite interpolation, and the kernel limit	11
5 The first universal phenomenon	11
6 The Bayesian perspective in function space	13
7 Problems (optional homework)	16

About these notes

These are notes for the first lecture of the mini-course. The first two lectures (Module I, infinite-width networks) follow, to a large extent, the review of Ringel et al. (2025), referred to below as *the Review*; pointers such as (Review §2.1) refer to its sections. We assume fluency with training neural networks and no background in statistical physics.

Status labels. Physics derivations and theorems look alike on the page, and the audience is mixed. Key results therefore carry one of four labels: **Exact** an identity derived in these notes; **Theorem** a rigorous result in the literature, cited; **Controlled: small parameter** exact in a stated limit, with corrections that are computable in principle; **Approximation** an uncontrolled step, supported by numerics or by rigorous results in special cases.

Homework. Optional problems are collected in Section 7; those marked (*) are longer or computational.

Hour 1: The approach and the toolbox

1 The approach

1.1 Universal phenomena

Deep learning is, for now, an experimental science. Its most useful empirical regularities are those that recur across architectures, datasets and optimizers. Borrowing a term from critical phenomena — where the same exponents describe a magnet and a liquid regardless of microscopic details — we call such regularities *universal*. The program of this course is to identify universal phenomena and then explain them, preferably with theories that are themselves insensitive to microscopic details. Examples:

- **Generalization despite interpolation.** Networks with far more parameters than training points, able to fit random labels, generalize well on real labels, and often improve as parameters are added (Neyshabur et al., 2014; Zhang et al., 2017; Belkin et al., 2019; Nakkiran et al., 2020).
- **Spectral (simplicity) bias.** Smooth, low-frequency components of a target are learned first (Rahaman et al., 2019) and from fewer samples (Canatar et al., 2021; Cohen et al., 2021).
- **Feature-learning transitions.** Abrupt changes in what a network represents. In grokking, generalization appears long after the training set is fitted (Power et al., 2022), as Fourier-feature circuits form (Nanda et al., 2023); in simple settings this is a first-order phase transition (Rubin et al., 2024).
- **Scaling laws.** Test loss falls as a power law in data, parameters and compute (Hestness et al., 2017; Kaplan et al., 2020).
- **Hyperparameter transfer.** Optimal hyperparameters — learning rate, initialization scale, batch size — follow simple scaling rules as models grow, so values tuned on a small model carry over to a large one. The rules are stable across compute and data scales and, mutatis

mutandis, across architectures (Yang et al., 2022; Bordelon et al., 2024; DeepSeek-AI, 2024; Bjorck et al., 2025).

1.2 General formalisms first, toy models second

In condensed-matter physics one rarely builds a bespoke theory for each material. One starts from a general formalism — statistical field theory — and specializes it to toy models (the Ising model, ϕ^4 theory, and so on) chosen to isolate a mechanism. The formalism carries information that no single toy model does: which symmetries matter, which approximations are controlled, which details are irrelevant. As a result, theory can *surprise* us by predicting phenomena that were not put in by hand, rather than only rationalizing what has already been observed.

Physics has many examples. The Aharonov–Bohm effect was derived from the Schrödinger equation with electromagnetic potentials (Aharonov and Bohm, 1959) before it was observed: electrons passing on either side of a solenoid acquire a relative phase although the magnetic field vanishes everywhere along their paths, and the phase depends only on how many times a path winds around the solenoid. This and many other topological effects were first understood through formalism — Schrödinger’s equation, field theory and algebraic topology. Topological insulators, for instance, were predicted (Kane and Mele, 2005; Bernevig et al., 2006) before they were observed (König et al., 2007). Black holes came out of the equations of general relativity (Oppenheimer and Snyder, 1939) decades before astronomers found them. Elementary particles are organized by representation theory: Wigner identified particles with irreducible representations of the Poincaré group, labeled by mass and spin (Wigner, 1939), and the SU(3) classification of hadrons predicted the Ω^- baryon before its discovery (Barnes et al., 1964). Representation theory returns in Lecture 2, where it organizes the eigenvalues of kernels.

A general formalism has a second advantage: it gives theorists a common language in which tools and ideas can be shared, instead of being rediscovered under different notations. The effective ridge of Lecture 2 is a case in point (Review §2.4.3).

In this course the formalism is statistical field theory for the distribution of a network’s outputs $f(x)$ (the Review), and the toy models are architectures and data distributions for which the resulting field theory is tractable: fully connected networks on the sphere, permutation-symmetric sequences, modular arithmetic, and others.

1.3 Timescales

We assume we have of order ten years to develop this science. If the relevant timescale is much shorter, the leverage lies in policy and governance rather than in the slow accumulation of scientific understanding.

2 Statistical mechanics as probability theory

2.1 Boltzmann distributions

A statistical-mechanical system consists of microstates s (discrete or continuous), an energy function $E(s)$, and a temperature $T > 0$. In equilibrium the probability of a microstate is

$$P(s) = \frac{e^{-E(s)/T}}{Z}, \quad Z = \sum_s e^{-E(s)/T}, \quad (2.1)$$

with sums replaced by integrals for continuous s . Nothing here is specific to physics: any distribution with full support can be written this way, with $E = -T \log P + \text{const}$. For instance, a Bayesian posterior $P(\theta | \mathcal{D}) = P(\mathcal{D} | \theta)P(\theta)/P(\mathcal{D})$ is of the form (2.1) with $T = 1$, $E(\theta) = -\log P(\mathcal{D} | \theta) - \log P(\theta)$, and $Z = P(\mathcal{D})$, the evidence. In Section 4 the correspondence becomes dynamical: noisy gradient descent samples such distributions.

2.2 Partition functions and probabilities

Z is more than a normalization constant. Let $m(s)$ be any function of the microstate — a *macroscopic* observable, or macrostate. Its distribution is a ratio of partition functions,

$$P(m) = \frac{Z(m)}{Z}, \quad Z(m) = \sum_s \delta_{m,m(s)} e^{-E(s)/T}, \quad (2.2)$$

with the Kronecker delta replaced by a Dirac delta for continuous m . We will use (2.2) repeatedly with the weights θ of a network as microstates and its outputs f as the macrostate: the delta function turns a distribution over weights into a distribution over functions (Sections 3.5 and 6).

2.3 Sources, expectation values and cumulants

Add *source terms* J_i coupled to observables $O_i(s)$:

$$Z[\mathbf{J}] = \sum_s \exp \left(-\frac{E(s)}{T} + \sum_i J_i O_i(s) \right). \quad (2.3)$$

Differentiating the logarithm generates expectation values and correlations:

$$\left. \frac{\partial \log Z}{\partial J_i} \right|_{\mathbf{J}=0} = \langle O_i \rangle, \quad \left. \frac{\partial^2 \log Z}{\partial J_i \partial J_j} \right|_{\mathbf{J}=0} = \langle O_i O_j \rangle - \langle O_i \rangle \langle O_j \rangle, \quad (2.4)$$

and in general the n -th derivative is the n -th joint cumulant (in physics language, the connected n -point function). Equivalently,

$$\log Z[\mathbf{J}] - \log Z[0] = \log \langle e^{\sum_i J_i O_i} \rangle \quad (2.5)$$

is the cumulant generating function (CGF) of the O_i under the Gibbs measure; for probabilists, the log-Laplace transform, hence a convex function of $\mathbf{J}$. The inverse temperature is itself a source, coupled to $-E$: with $\beta = 1/T$,

$$-\frac{\partial \log Z}{\partial \beta} = \langle E \rangle, \quad \frac{\partial^2 \log Z}{\partial \beta^2} = \text{Var}(E). \quad (2.6)$$

2.4 Free energy and entropy

The free energy is

$$F \equiv -T \log Z. \quad (2.7)$$

Start from Shannon's definition of the entropy of the Gibbs distribution,

$$S = - \sum_s P(s) \log P(s), \quad (2.8)$$

and simply track the formula. By (2.1), $-\log P(s) = E(s)/T + \log Z$, so

$$S = \frac{\langle E \rangle}{T} + \log Z = \frac{\langle E \rangle - F}{T} \iff F = \langle E \rangle - TS. \quad (2.9)$$

The thermodynamic definition of entropy gives the same object. Since $\partial_T \log Z = \langle E \rangle / T^2$,

$$-\frac{\partial F}{\partial T} = \log Z + T \partial_T \log Z = \log Z + \frac{\langle E \rangle}{T} = S. \quad \text{Exact} \quad (2.10)$$

Two limits are worth keeping in mind (for a finite set of microstates). As $T \rightarrow \infty$, P is uniform and $S = \log(\text{number of microstates})$. As $T \rightarrow 0$, P is uniform on the minimizers of E , and $S \rightarrow \log(\text{number of minimizers})$. The second limit is the relevant one for overparametrized networks: very many weight configurations reach zero training loss, and the question becomes which functions — macrostates, [Section 2.5](#) — they realize most often. (For continuous microstates the $T \rightarrow 0$ limit also weighs each minimizer by the curvature of E around it, favoring flat minima; see [Section 4.3](#).)

2.5 Macrostates: energy versus entropy

Suppose the energy depends on the microstate only through the macrostate, $E(s) = E(m(s))$. Then (2.2) gives $Z(m) = \Omega(m) e^{-E(m)/T}$ with $\Omega(m)$ the number of microstates realizing m , and

$$P(m) = \frac{e^{-F(m)/T}}{Z}, \quad F(m) = E(m) - TS(m), \quad S(m) = \log \Omega(m). \quad \text{Exact} \quad (2.11)$$

This is consistent with the previous subsection: conditioned on m , the microstates are uniformly distributed over $\Omega(m)$ states, and the Shannon entropy of that conditional distribution is $\log \Omega(m) = S(m)$. The most probable macrostate minimizes $F(m)$. Energy favors macrostates of low E ; entropy favors macrostates realized by many microstates; the temperature sets the exchange rate between the two. For continuous microstates with a reference measure $\rho(s) ds$, $\Omega(m) = \int ds \rho(s) \delta(m - m(s))$ is a volume (a density in m) and $S(m)$ its logarithm. [Problem 1](#) works this out for n spins in a magnetic field.

Dictionary		
Statistical mechanics	Probability	This lecture
microstate s	outcome	weights θ
reference measure $\rho(s)$	base measure	weight prior $\pi(\theta)$ (weight decay)
energy $E(s)/T$	$-\log$ density (+ const)	data loss/ T
temperature T	—	gradient noise = ridge κ^2
partition function Z	normalizer	evidence $P(\mathcal{D})$
$\log Z[\mathbf{J}]$	cumulant generating function	generating function of output statistics
macrostate $m(s)$	pushforward variable	network outputs $f_\theta(x)$
entropy $S(m) = \log \Omega(m)$	log-density of the pushforward	log of the prior over functions
$F(m) = E(m) - TS(m)$	$-T \log$ (marginal)	$-T \log$ (posterior over functions)

3 Gaussian integrals and path integrals

3.1 Multivariate Gaussians

We write $\mathcal{N}[\boldsymbol{\mu}, \Sigma; \mathbf{w}] = (2\pi)^{-d/2} \det(\Sigma)^{-1/2} \exp(-\frac{1}{2}(\mathbf{w} - \boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{w} - \boldsymbol{\mu}))$ for the density of $\mathbf{w} \in \mathbb{R}^d$ (Review §1.3). The one identity to remember is *square completion*:

$$\int d^d w \mathcal{N}[0, \Sigma; \mathbf{w}] e^{\mathbf{J}^\top \mathbf{w}} = e^{\frac{1}{2} \mathbf{J}^\top \Sigma \mathbf{J}}, \quad \text{Exact} \quad (3.1)$$

which follows from $-\frac{1}{2} \mathbf{w}^\top \Sigma^{-1} \mathbf{w} + \mathbf{J}^\top \mathbf{w} = -\frac{1}{2}(\mathbf{w} - \Sigma \mathbf{J})^\top \Sigma^{-1}(\mathbf{w} - \Sigma \mathbf{J}) + \frac{1}{2} \mathbf{J}^\top \Sigma \mathbf{J}$ and a shift of the integration variable (a shift of contour for complex $\mathbf{J}$). In the language of Section 2, the CGF of a centered Gaussian is exactly quadratic: the mean vanishes, the covariance is Σ , and all higher cumulants vanish. Conversely, a random vector is Gaussian if and only if its CGF is quadratic. With $\mathbf{J} = i\mathbf{t}$ we get the characteristic function $e^{-\frac{1}{2} \mathbf{t}^\top \Sigma \mathbf{t}}$. For an unnormalized Gaussian,

$$\int d^d w e^{-\frac{1}{2} \mathbf{w}^\top A \mathbf{w} + \mathbf{J}^\top \mathbf{w}} = (2\pi)^{d/2} \det(A)^{-1/2} e^{\frac{1}{2} \mathbf{J}^\top A^{-1} \mathbf{J}}. \quad (3.2)$$

Differentiating (3.1) four times at $\mathbf{J} = 0$ gives the Wick (Isserlis) theorem,

$$\langle w_i w_j w_k w_l \rangle = \Sigma_{ij} \Sigma_{kl} + \Sigma_{ik} \Sigma_{jl} + \Sigma_{il} \Sigma_{jk}, \quad (3.3)$$

and in general a $2n$ -th moment is the sum over all $(2n-1)!!$ ways of pairing the indices.

3.2 Marginals and conditionals

Marginalizing a Gaussian keeps the corresponding sub-block of the covariance. Conditioning is the formula behind Gaussian-process regression: for a centered Gaussian vector $(\mathbf{u}, \mathbf{v})$ with blocks $\Sigma_{uu}, \Sigma_{uv}, \Sigma_{vv}$,

$$\mathbf{u} | \mathbf{v} \sim \mathcal{N}[\Sigma_{uv} \Sigma_{vv}^{-1} \mathbf{v}, \Sigma_{uu} - \Sigma_{uv} \Sigma_{vv}^{-1} \Sigma_{vu}]. \quad \text{Exact} \quad (3.4)$$

(Problem 3: derive this by square completion.)

3.3 From vectors to functions

A Gaussian process $f \sim \text{GP}(0, K)$ is a random function such that for every finite set of inputs $x_1, \dots, x_n$ the vector $(f(x_1), \dots, f(x_n))$ is $\mathcal{N}[0, K_n]$ with $(K_n)_{ij} = K(x_i, x_j)$. For any symmetric positive semi-definite kernel such a process exists (Kolmogorov extension theorem). Theorem

Physicists write the same object as a path integral (Review §1.8),

$$P[f] = \frac{e^{-S_{\text{GP}}[f]}}{Z}, \quad S_{\text{GP}}[f] = \frac{1}{2} \int d\mu_x d\mu_{x'} f(x) K^{-1}(x, x') f(x'), \quad Z = \int \mathcal{D}f e^{-S_{\text{GP}}[f]}. \quad (3.5)$$

Here $d\mu_x = p(x) dx$ is a probability measure on inputs — later, the data distribution — and K^{-1} is the inverse of K as an operator on functions, defined below. A path integral is shorthand for a limit of finite-dimensional integrals, obtained by cutting off the number of degrees of freedom. Two cutoffs are natural here.

- (i) *A grid.* Place n points $x_1, \dots, x_n$ distributed according to μ . The field on the grid is a vector with law $\mathcal{N}[0, K_n]$. Since $\frac{1}{n} \sum_i \rightarrow \int d\mu_x$, one finds $(K_n^{-1})_{ij} \approx n^{-2} K^{-1}(x_i, x_j)$, and $\frac{1}{2} \mathbf{f}_n^\top K_n^{-1} \mathbf{f}_n$ tends to $S_{\text{GP}}[f]$ as $n \rightarrow \infty$ (Review §1.8).

- (ii) *A mode cutoff.* View K as the integral operator $(Kg)(x) = \int d\mu_{x'} K(x, x')g(x')$ on $L^2(\mu)$. By Mercer's theorem (K continuous and positive semi-definite, μ a finite Borel measure with compact support; the expansion converges absolutely and uniformly on $\text{supp } \mu \times \text{supp } \mu$, and in general only there) [Theorem](#)

$$\begin{aligned} \int d\mu_{x'} K(x, x') \phi_k(x') &= \lambda_k \phi_k(x), \quad \lambda_k \geq 0, \quad \int d\mu_x \phi_k \phi_q = \delta_{kq}, \\ K(x, x') &= \sum_k \lambda_k \phi_k(x) \phi_k(x'). \end{aligned} \quad (3.6)$$

Keeping the modes $k \leq \Lambda$, $f = \sum_{k \leq \Lambda} f_k \phi_k$, the action becomes $S_{\text{GP}}[f] = \frac{1}{2} \sum_{k \leq \Lambda} f_k^2 / \lambda_k$, with $K^{-1}(x, x') = \sum_k \lambda_k^{-1} \phi_k(x) \phi_k(x')$ and $\mathcal{D}f = \prod_{k \leq \Lambda} df_k$. The Gaussian process becomes a product of independent one-dimensional Gaussians, $f_k \sim \mathcal{N}[0, \lambda_k]$ (the Karhunen–Loève expansion). The quadratic form $\sum_k f_k^2 / \lambda_k$ is the squared norm $\|f\|_K^2$ of f in the reproducing kernel Hilbert space (RKHS) of K . It does not depend on the choice of μ among measures with the same support (Review §1.6), as befits a prior.

In physics the cutoff is usually placed in Fourier space, as in (ii). It is an empirical observation that for “good” theories — ones in which nothing too pathological happens at high frequencies — reasonable observables become cutoff independent as the cutoff goes to infinity, and that carrying out the computation directly in the path integral leads to the same asymptotic results. For the Gaussian process this can be checked by hand: with the mode cutoff, $\langle f(x)f(x') \rangle = \sum_{k \leq \Lambda} \lambda_k \phi_k(x) \phi_k(x') \rightarrow K(x, x')$ by Mercer's theorem, which is what the continuum computation of [Section 3.4](#) gives directly. Not every quantity is a reasonable observable: the action itself, evaluated on a typical sample, grows with the cutoff, $\langle S_{\text{GP}} \rangle = \Lambda/2$, and its continuum value is infinite.

For mathematicians: what is rigorous. Gaussian theories are on firm ground. A Gaussian process is a Gaussian measure on a space of functions, and the RKHS of K is its Cameron–Martin space: shifting the measure by h gives an equivalent measure if h lies in the RKHS, and a mutually singular one otherwise ([Bogachev, 1998](#)). The RKHS itself has measure zero: typical samples have $S_{\text{GP}} = \infty$, which is the divergence of $\langle S_{\text{GP}} \rangle = \Lambda/2$ seen above. So $e^{-S_{\text{GP}}} \mathcal{D}f$ has no meaning on its own, but ratios do. A source tilts the measure into the same Gaussian with mean $m = \int dx' K(\cdot, x') \alpha(x')$, an RKHS element, and $\log Z[\alpha]/Z[0] = \frac{1}{2} \|m\|_K^2$. With finitely many data points, the posterior has density $\propto \exp(-\sum_{\mu} (f(x_{\mu}) - y_{\mu})^2 / 2\kappa^2)$ with respect to the prior, so Bayes' rule over functions is rigorous ([Stuart, 2010](#)).

Interacting theories are harder. Constructive field theory built ϕ^4 in two and three dimensions ([Glimm and Jaffe, 1987](#)); in four, the scaling limits of lattice ϕ^4 are Gaussian ([Aizenman and Duminil-Copin, 2021](#)), and four-dimensional Yang–Mills theory remains open. One productive recent route is the idea of [Section 4](#): stochastic quantization ([Parisi and Wu, 1981](#)) realizes e^{-S}/Z as the stationary law of a Langevin equation for the field, and the theory of singular SPDEs ([Hairer, 2014](#); [Gubinelli et al., 2015](#)) controls that equation, yielding, for instance, a new construction of ϕ_3^4 ([Gubinelli and Hofmanová, 2021](#)). A probabilistic construction of Liouville conformal field theory, based on Gaussian multiplicative chaos, has been carried as far as a proof of the DOZZ formula for its three-point function ([Kupiainen et al., 2020](#)).

In deep learning the cutoffs are physical — finite width, finite data — so the finite-dimensional integrals always exist, and the questions concern their limits.

3.4 Sources and functional derivatives

Add a source $\alpha(x)$ coupled to the field through the plain measure dx ,

$$Z[\alpha] = \int \mathcal{D}f e^{-S_{\text{GP}}[f] + \int dx \alpha(x) f(x)}. \quad (3.7)$$

Since $\int dx \alpha(x) f(x)$ is a Gaussian variable with variance $\int dx dx' \alpha(x) K(x, x') \alpha(x')$, (3.1) gives

$$\frac{Z[\alpha]}{Z[0]} = \left\langle e^{\int dx \alpha(x) f(x)} \right\rangle = \exp \left(\frac{1}{2} \int dx dx' \alpha(x) K(x, x') \alpha(x') \right). \quad \text{Exact} \quad (3.8)$$

With the functional derivative defined by $\delta\alpha(x')/\delta\alpha(x) = \delta(x - x')$,

$$\left. \frac{\delta \log Z}{\delta \alpha(x)} \right|_{\alpha=0} = \langle f(x) \rangle = 0, \quad \left. \frac{\delta^2 \log Z}{\delta \alpha(x) \delta \alpha(x')} \right|_{\alpha=0} = \langle f(x) f(x') \rangle = K(x, x'). \quad (3.9)$$

This is (2.4) with the discrete index i promoted to a continuous label x .

A word on measures. Following the Review, the action is built with the data measure $d\mu_x$, while sources couple through dx . With this choice $\delta/\delta\alpha(x)$ returns $\langle f(x) \rangle$ itself. A source written as $\int d\mu_x \alpha(x) f(x)$ would instead produce $p(x)\langle f(x) \rangle$, and (3.8) would carry $d\mu_x d\mu_{x'}$ in place of $dx dx'$.

3.5 Functional delta functions and auxiliary fields

(a) Pushforwards. Let f_θ be a parametric family of functions (e.g. a neural network) and $\pi(\theta)$ a distribution over parameters. The induced law of the outputs on points $x_1, \dots, x_n$ is a ratio of partition functions, as in (2.2). Writing $\delta(u) = \int \frac{dt}{2\pi} e^{itu}$,

$$P_0(f_1, \dots, f_n) = \int d\theta \pi(\theta) \prod_{i=1}^n \delta(f_i - f_\theta(x_i)) = \int \prod_{i=1}^n \frac{dt_i}{2\pi} e^{i \sum_i t_i f_i} \left\langle e^{-i \sum_i t_i f_\theta(x_i)} \right\rangle_\pi. \quad \text{Exact} \quad (3.10)$$

The *auxiliary variables* t_i are conjugate to the outputs, and the last factor is the characteristic function of the outputs. If it is Gaussian, $e^{-\frac{1}{2} \mathbf{t}^\top K_n \mathbf{t}}$, the $\mathbf{t}$ -integral returns $\mathcal{N}[0, K_n; \mathbf{f}]$. This is how the Gaussian-process limit of wide networks will enter (Section 6.3).

The functional version, with the auxiliary field coupled through the data measure as in the Review (§3.1), is

$$\delta[f - g] = \int \mathcal{D}\tilde{f} e^{i \int d\mu_x \tilde{f}(x) (f(x) - g(x))}, \quad (3.11)$$

which reduces, for any discretization, to a product of ordinary delta functions times an irrelevant constant. This convention matches the action (3.5). If f_θ is Gaussian,

$$\left\langle e^{-i \int d\mu_x \tilde{f}(x) f_\theta(x)} \right\rangle_\pi = e^{-\frac{1}{2} \int d\mu_x d\mu_{x'} \tilde{f}(x) K(x, x') \tilde{f}(x')}, \quad (3.12)$$

and integrating out $\tilde{f}$ in $\int d\theta \pi(\theta) \delta[f - f_\theta]$ returns $e^{-S_{\text{GP}}[f]}$ with exactly its measures. When a source written with dx meets an auxiliary field written with $d\mu_x$, integrating out the auxiliary field turns $\alpha(x)$ into $\alpha(x)/p(x)$ (Review §3.4). In Section 6 the functional delta moves the Bayesian posterior over weights to function space; it is also the starting point of Lecture 2.

(b) Worked example: Gaussian-process regression. Take a prior under which the test output and the P training outputs are jointly Gaussian,

$$(f_*, \mathbf{f}) \sim \mathcal{N}[0, \hat{K}], \quad \hat{K} = \begin{pmatrix} K_{**} & \mathbf{k}_*^\top \\ \mathbf{k}_* & K_{\mathcal{D}} \end{pmatrix}, \quad (3.13)$$

$$K_{**} = K(x_*, x_*), \quad (\mathbf{k}_*)_\mu = K(x_*, x_\mu), \quad (K_{\mathcal{D}})_{\mu\nu} = K(x_\mu, x_\nu),$$

and a Gaussian likelihood with noise variance κ^2 . The posterior partition function, with a source α on the test output, is

$$Z[\alpha] = \int df_* d^P f \mathcal{N}[0, \hat{K}; (f_*, \mathbf{f})] \exp\left(\alpha f_* - \frac{1}{2\kappa^2} \sum_{\mu=1}^P (f_\mu - y_\mu)^2\right). \quad (3.14)$$

We evaluate it in three steps, each an instance of (3.1).

1. *Auxiliary variables* linearize the likelihood: $e^{-u^2/(2\kappa^2)} = \frac{\kappa}{\sqrt{2\pi}} \int dt e^{-\kappa^2 t^2/2 + it u}$, one t_μ per training point.
2. *Integrate out the outputs* using (3.1) with $\mathbf{J} = (\alpha, i\mathbf{t})$: $\langle e^{\alpha f_* + i\mathbf{t}^\top \mathbf{f}} \rangle = \exp\left(\frac{1}{2}\alpha^2 K_{**} + i\alpha \mathbf{k}_*^\top \mathbf{t} - \frac{1}{2}\mathbf{t}^\top K_{\mathcal{D}} \mathbf{t}\right)$.
3. *Integrate out $\mathbf{t}$* using (3.2):

$$Z[\alpha] \propto e^{\frac{1}{2}\alpha^2 K_{**}} \int d^P t e^{-\frac{1}{2}\mathbf{t}^\top (K_{\mathcal{D}} + \kappa^2 I) \mathbf{t} - i\mathbf{t}^\top (\mathbf{y} - \alpha \mathbf{k}_*)}$$

$$\propto \exp\left(\frac{1}{2}\alpha^2 K_{**} - \frac{1}{2}(\mathbf{y} - \alpha \mathbf{k}_*)^\top (K_{\mathcal{D}} + \kappa^2 I)^{-1} (\mathbf{y} - \alpha \mathbf{k}_*)\right). \quad (3.15)$$

Reading off the first two cumulants as in (2.4),

$$\langle f_* \rangle = \mathbf{k}_*^\top (K_{\mathcal{D}} + \kappa^2 I)^{-1} \mathbf{y}, \quad \text{Var}(f_*) = K_{**} - \mathbf{k}_*^\top (K_{\mathcal{D}} + \kappa^2 I)^{-1} \mathbf{k}_*, \quad \text{Exact} \quad (3.16)$$

and all higher cumulants vanish. These are the standard GPR formulas (Rasmussen and Williams, 2006). As a byproduct, $\log Z[0] = -\frac{1}{2}\mathbf{y}^\top (K_{\mathcal{D}} + \kappa^2 I)^{-1} \mathbf{y} - \frac{1}{2} \log \det(K_{\mathcal{D}} + \kappa^2 I) + \text{const}$ is, up to a κ -dependent constant, the log marginal likelihood: the free energy $-\log Z[0]$ (at $T = 1$) is minus the log evidence.

4 Langevin dynamics and equilibrium

4.1 Langevin dynamics

Noisy gradient descent on a loss $L(\theta)$ is modeled by the Langevin equation (Welling and Teh, 2011) (Review §1.4)

$$d\theta_t = -\eta \nabla L(\theta_t) dt + \sqrt{2\eta T} dW_t, \quad (4.1)$$

with W_t a standard Brownian motion, learning rate η and temperature T . At $T = 0$ it is gradient flow.

4.2 The Gibbs distribution is stationary

The density $\rho_t(\theta)$ of θ_t obeys the Fokker–Planck equation (Gardiner, 2009)

$$\partial_t \rho = \eta \nabla \cdot (\rho \nabla L + T \nabla \rho) \equiv -\nabla \cdot j. \quad (4.2)$$

A stationary state with vanishing probability current j requires $\rho \nabla L + T \nabla \rho = 0$, that is $\nabla \log \rho = -\nabla L/T$:

$$\rho_\infty(\theta) = \frac{e^{-L(\theta)/T}}{Z}. \quad \text{Exact} \quad (4.3)$$

If L is smooth and confining (weight decay makes it so), ρ_∞ is the unique stationary density and $\rho_t \rightarrow \rho_\infty$ (Pavliotis, 2014, ch. 4); rates require more, e.g. a Poincaré inequality for ρ_∞ . Theorem Moreover, the free energy of Section 2, evaluated on ρ_t ,

$$F[\rho] = \int d\theta \rho L + T \int d\theta \rho \log \rho = \langle L \rangle_\rho - T S[\rho], \quad \frac{dF[\rho_t]}{dt} = -\eta \int d\theta \rho_t |\nabla(L + T \log \rho_t)|^2 \leq 0, \quad (4.4)$$

decreases along the dynamics. In fact $F[\rho] = -T \log Z + T \text{KL}(\rho \| \rho_\infty)$, so F is minimized exactly at (4.3), where it equals the free energy $-T \log Z$. Exact

Two caveats. First, convergence can be slow: rates can be exponentially small in (barrier height)/ T , so equilibrium is a statement about long times. For wide networks, network outputs and pre-activations are found to relax on timescales comparable to that of the loss (Naveh et al., 2021) (Review §1.4). Second, the gradient noise of SGD is neither isotropic nor state-independent; (4.1) is a model of it, not a derivation. Approximation

4.3 Equilibrium is Bayesian inference

Take the loss to be the unnormalized squared error plus a weight decay proportional to T (Review §2.1),

$$L(\theta) = L_{\mathcal{D}}(\theta) + T \sum_{\alpha} \frac{\theta_{\alpha}^2}{2\sigma_{\alpha}^2}, \quad L_{\mathcal{D}}(\theta) = \frac{1}{2} \sum_{\mu=1}^P (f_{\theta}(x_{\mu}) - y_{\mu})^2. \quad (4.5)$$

Then (4.3) factorizes,

$$\rho_{\infty}(\theta) \propto \underbrace{\exp\left(-\frac{1}{2\kappa^2} \sum_{\mu=1}^P (f_{\theta}(x_{\mu}) - y_{\mu})^2\right)}_{\text{likelihood}} \underbrace{\prod_{\alpha} \mathcal{N}[0, \sigma_{\alpha}^2; \theta_{\alpha}]}_{\text{prior } \pi(\theta)}, \quad \kappa^2 \equiv T. \quad (4.6)$$

The equilibrium of noisy training is the Bayesian posterior of a network whose outputs are corrupted by Gaussian noise of variance $\kappa^2 = T$, with an independent Gaussian prior on the weights. Weight decay is the prior; temperature is the noise level, which will play the role of a ridge parameter. Two remarks:

- The weight-decay coefficient T/σ_{α}^2 scales with T so that the prior stays fixed. Choosing σ_{α}^2 equal to the initialization variances makes the prior the initialization distribution.
- In the overparametrized regime, as $T \rightarrow 0$ the posterior converges to the prior conditioned on fitting the data, formally $\propto \pi(\theta) \delta(f_{\theta}(X) - \mathbf{y})$. On the zero-loss manifold this has density $\pi(\theta)/\sqrt{\det J_{\theta} J_{\theta}^T}$, with J_{θ} the Jacobian of the training outputs (co-area formula), so flatter

interpolants weigh more. In function space the limit is simply the prior over functions conditioned on $f(X) = \mathbf{y}$ (Section 6): which interpolating function dominates is decided by the entropy of Section 2.5.

Hour 2: Generalization despite interpolation, and the kernel limit

5 The first universal phenomenon

5.1 A curve-fitting puzzle

The folklore of every lab class: never fit 10 data points with 100 parameters. Figure 1 puts it to the test. Ten noisy samples of a smooth function on $[-1, 1]$ are fitted with a polynomial of M coefficients, written in the orthonormal Legendre basis with the coefficients weighted by a power law,

$$f(x) = \sum_{\ell=0}^{M-1} c_{\ell} \sqrt{\lambda_{\ell}} \psi_{\ell}(x), \quad \lambda_{\ell} = (\ell + 1)^{-3/2}, \quad \psi_{\ell} = \sqrt{2\ell + 1} P_{\ell}, \quad (5.1)$$

where P_{ℓ} are the Legendre polynomials, so that $\int_{-1}^1 \frac{dx}{2} \psi_{\ell} \psi_{\ell'} = \delta_{\ell\ell'}$. The coefficients minimize

$$\sum_{\mu=1}^P (f(x_{\mu}) - y_{\mu})^2 + \kappa^2 \sum_{\ell=0}^{M-1} c_{\ell}^2. \quad (5.2)$$

We focus on vanishing ridge, $\kappa^2 \rightarrow 0^+$. The fit is then the least-squares fit when $M < P$, the unique interpolant when $M = P$, and, when $M > P$, the interpolant with the smallest $\sum_{\ell} c_{\ell}^2$ (the minimum-norm interpolant). The weights decide which coefficients are cheap.

The coefficients can be eliminated, and what remains is a kernel. With $\Phi_{\mu\ell} = \sqrt{\lambda_{\ell}} \psi_{\ell}(x_{\mu})$, the minimizer of (5.2) is $\mathbf{c} = (\Phi^T \Phi + \kappa^2 I)^{-1} \Phi^T \mathbf{y} = \Phi^T (\Phi \Phi^T + \kappa^2 I)^{-1} \mathbf{y}$, so

$$\hat{f}(x) = K_M(x, X) (K_M(X, X) + \kappa^2 I)^{-1} \mathbf{y}, \quad K_M(x, x') = \sum_{\ell=0}^{M-1} \lambda_{\ell} \psi_{\ell}(x) \psi_{\ell}(x'). \quad \text{Exact} \quad (5.3)$$

Here $X = (x_1, \dots, x_P)$ collects the training inputs, $K_M(x, X)$ is the row vector $(K_M(x, x_{\mu}))_{\mu}$, and $K_M(X, X)$ is the $P \times P$ kernel matrix. This is the GPR mean (3.16) with prior $c_{\ell} \sim \mathcal{N}[0, 1]$ and noise variance κ^2 . As $\kappa^2 \rightarrow 0^+$ with $M \geq P$ it becomes kernel interpolation, $\hat{f}(x) = K_M(x, X) K_M(X, X)^{-1} \mathbf{y}$: the minimum-norm interpolant is the ridgeless kernel regressor. The eigenfunctions of K_M with respect to the uniform measure $d\mu_x = dx/2$ are the ψ_{ℓ} , and its eigenvalues are the power law λ_{ℓ} : the Mercer decomposition (3.6), read backwards. The number of coefficients enters only as the cutoff M on this spectral sum. Since $\sum_{\ell} \lambda_{\ell} = \zeta(3/2) \approx 2.6$ is finite, K_M has a limit as $M \rightarrow \infty$, a trace-class kernel: infinitely many parameters define a perfectly good prior.

Three observations, all at vanishing ridge:

- (i) At the interpolation threshold, $M = P$, the curve through the noisy points oscillates wildly (panel a). This is the textbook picture of overfitting.
- (ii) With ten times more coefficients than data points, $M = 101$, the minimum-norm fit is smooth and close to the truth (panel b).

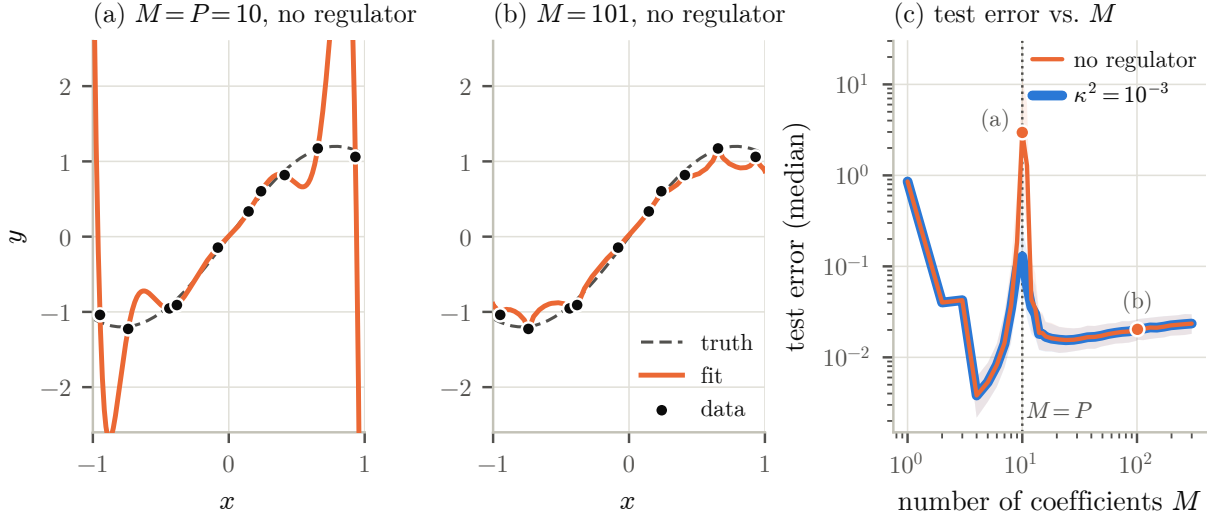


Figure 1: Double descent in curve fitting. $P = 10$ noisy samples ($\sigma_\epsilon = 0.1$) of $y(x) = 1.2 \sin 2x$, one drawn uniformly in each tenth of $[-1, 1]$, are fitted with M power-law-weighted Legendre coefficients (5.1), i.e. by kernel regression with K_M of (5.3). (a) $M = P$, no regulator: the unique interpolant. (b) $M = 101$, no regulator: the minimum-norm interpolant, on the same data. (c) Test error — the mean squared deviation from the noiseless target over $[-1, 1]$ — against M . Lines: median over 2000 draws of the data; bands: 25–75% range. Orange: vanishing ridge; blue: $\kappa^2 = 10^{-3}$. Dots mark panels (a) and (b).

- (iii) The test error against M shows double descent (Belkin et al., 2019; Nakkiran et al., 2020), a phenomenon with a long prehistory in the statistical mechanics of learning (Loog et al., 2020). The error falls to about 0.004 at $M = 4$, rises to a sharp peak of about 3 at $M = P$, and falls again to a plateau of about 0.02, twice the noise variance, that drifts only slowly between $M = 20$ and $M = 300$ (panel c). A small ridge reduces the peak to a bump and changes little elsewhere.

So the folklore is half right: the danger zone is a number of parameters close to the number of data points, with no regulator. Adding parameters far beyond that is harmless. Why? We return to this once we have a theory of spectral bias.

5.2 The phenomenon in deep learning

Neural networks are routinely trained with far more parameters than data points. They can fit randomly relabeled data (Zhang et al., 2017), so a small hypothesis class cannot be what makes them succeed; yet on real labels they generalize, and past the interpolation threshold test error typically keeps improving as width grows (Neyshabur et al., 2014; Belkin et al., 2019; Nakkiran et al., 2020). The pattern holds across architectures and datasets: a universal phenomenon in the sense of Section 1. For linear regression, when the minimum-norm interpolant is nevertheless consistent (“benign overfitting”) is characterized, up to constants, by two effective ranks of the feature covariance (Bartlett et al., 2020).

For mathematicians. Uniform-convergence bounds over the whole hypothesis class are vacuous at these parameter counts. Whether any uniform-convergence argument can explain the phenomenon is debated: Nagarajan and Kolter (2019) give settings where even algorithm-dependent, two-sided uniform convergence fails, and Negrea et al. (2020) recover it for de-randomized surrogates of the learned predictor. PAC-Bayes bounds, which control the risk of a *distribution* Q over weights relative to a prior π , can be non-vacuous for stochastic networks (Dziugaite and Roy, 2017). For Catoni-type bounds the optimal Q is a Gibbs posterior, $Q \propto \pi e^{-\beta L_{\mathcal{D}}}$ (Catoni, 2007): the equilibrium distribution of Section 4 with $\beta = 1/T$, for which $F[\rho] = \langle L_{\mathcal{D}} \rangle_{\rho} + T \text{KL}(\rho \parallel \pi) + \text{const.}$

5.3 Strategy

We take the paradox to its extreme: networks of infinite width, hence infinitely many parameters, trained with the noisy dynamics of Section 4. Their equilibrium is a Bayesian posterior over weights. Section 6 moves it to function space, where the number of parameters drops out, and shows that at infinite width the trained network is described by a Gaussian field theory whose predictor is kernel regression — the structure we found in the curve-fitting example.

6 The Bayesian perspective in function space

By Section 4.3, noisy training equilibrates to a Bayesian posterior over weights. We now move this posterior to function space, where the weights survive only through a prior over functions, and then let the width go to infinity.

6.1 From weights to functions: the functional prior

Start from the Boltzmann equilibrium (4.3) with the loss (4.5), written as a partition function over weights,

$$Z = \int d\theta e^{-L(\theta)/T} \propto \int d\theta \pi(\theta) e^{-L_{\mathcal{D}}(\theta)/T}, \quad \pi(\theta) = \prod_{\alpha} \mathcal{N}[0, \sigma_{\alpha}^2; \theta_{\alpha}]. \quad (6.1)$$

The data term depends on the weights only through the function the network computes. To make this explicit, insert a path integral over outputs together with the functional delta function (3.11), $1 \propto \int \mathcal{D}f \delta[f - f_{\theta}]$:

$$Z \propto \int \mathcal{D}f \int d\theta \pi(\theta) \delta[f - f_{\theta}] e^{-L_{\mathcal{D}}(\theta)/T}. \quad (6.2)$$

On the support of the delta function $f_{\theta}(x_{\mu}) = f(x_{\mu})$, so the data term leaves the weight integral, and what remains of the weight integral is a functional of f alone:

$$Z \propto \int \mathcal{D}f e^{-E[f]/T} P_0[f], \quad P_0[f] = \int d\theta \pi(\theta) \delta[f - f_{\theta}], \quad \text{Exact} \quad (6.3)$$

where $E[f] = \frac{1}{2} \sum_{\mu} (f(x_{\mu}) - y_{\mu})^2$ is the training loss written in terms of f . We call $P_0[f]$ the *functional prior*. It is the law of the function computed by a network whose weights are drawn from π : the pushforward (3.10) of the weight prior, now on all inputs at once. Removing $\int \mathcal{D}f$ and dividing by Z gives the posterior over functions, $P[f] = e^{-E[f]/T} P_0[f]/Z$. This is Bayes' rule in function space, with a Gaussian likelihood of variance $\kappa^2 = T$ (Review §2.1). Nothing has been approximated, and the result holds at any width. The architecture, the parametrization and the number of parameters now enter only through P_0 .

6.2 Energy versus entropy

Equation (6.3) is [Section 2.5](#) verbatim: the microstates are weights, the macrostate is the function f , and the energy $E[f]$ depends on the weights only through f . Hence

$$P[f] \propto \exp\left(-\frac{E[f] - T S[f]}{T}\right), \quad S[f] = \log P_0[f], \quad (6.4)$$

where $S[f]$ is the log-volume, under π , of weight configurations producing f . The Review notes (§2.3) that replacing the Gaussian weights by uniform weights of equal variance leaves P_0 unchanged at infinite width, provided the fan-in of every layer diverges (at fixed input dimension, keep the first layer Gaussian). In that sense S is a Boltzmann entropy: the logarithm of a count of weight configurations. The energy is minimized by any function that interpolates the data. Among the many interpolating functions, the entropy selects those realized by the most weight configurations. This is the regulator we were looking for.

6.3 Infinite width: the functional prior is a Gaussian process

[Neal \(1996\)](#) showed that the functional prior of a wide network is Gaussian. Take one hidden layer, with prior variances inversely proportional to fan-in as in the Review (§1.5),

$$f_\theta(x) = \sum_{i=1}^N a_i \sigma(\mathbf{w}_i \cdot x), \quad a_i \sim \mathcal{N}[0, \sigma_a^2/N], \quad \mathbf{w}_i \sim \mathcal{N}[0, \sigma_w^2 I_d/d], \quad (6.5)$$

all independent. Writing the delta function in (6.3) as in (3.11),

$$P_0[f] = \int \mathcal{D}\tilde{f} e^{i \int d\mu_x \tilde{f}(x) f(x)} \left\langle e^{-i \int d\mu_x \tilde{f}(x) f_\theta(x)} \right\rangle_\pi. \quad (6.6)$$

For a fixed auxiliary field $\tilde{f}$, the exponent in the average is a sum of N i.i.d. terms,

$$\int d\mu_x \tilde{f}(x) f_\theta(x) = \sum_{i=1}^N a_i g(\mathbf{w}_i), \quad g(\mathbf{w}) = \int d\mu_x \tilde{f}(x) \sigma(\mathbf{w} \cdot x), \quad (6.7)$$

each of mean zero and variance $\frac{\sigma_a^2}{N} \mathbb{E}_{\mathbf{w}}[g(\mathbf{w})^2]$. By the central limit theorem the sum is Gaussian as $N \rightarrow \infty$, so its characteristic function tends to

$$\left\langle e^{-i \int d\mu_x \tilde{f}(x) f_\theta(x)} \right\rangle_\pi \longrightarrow e^{-\frac{1}{2} \int d\mu_x d\mu_{x'} \tilde{f}(x) K(x, x') \tilde{f}(x')}, \quad (6.8)$$

$$K(x, x') = \sigma_a^2 \mathbb{E}_{\mathbf{w}}[\sigma(\mathbf{w} \cdot x) \sigma(\mathbf{w} \cdot x')].$$

Integrating out $\tilde{f}$ by square completion, as in [Section 3.5](#), gives

$$P_0[f] \propto e^{-S_{\text{GP}}[f]}, \quad S_{\text{GP}}[f] = \frac{1}{2} \int d\mu_x d\mu_{x'} f(x) K^{-1}(x, x') f(x'). \quad \text{Theorem} \quad (6.9)$$

The functional prior of an infinitely wide network is the Gaussian process of [Section 3.3](#), with the *neural network Gaussian process* (NNGP) kernel K . At finite N the higher cumulants of P_0 are of order $1/N$ ([Problem 6](#)).

For deep networks, [Lee et al. \(2018\)](#) took the widths to infinity one layer at a time and [Matthews et al. \(2018\)](#) proved the simultaneous limit. The mechanism is the same central limit theorem, applied

layer by layer (Review §1.5): each pre-activation is a sum of N terms that are i.i.d. conditional on the previous layer, and their empirical covariance concentrates. For a fully connected network with H hidden layers, $h^{(1)} = W^{(1)}x + b^{(1)}$, $h^{(l+1)} = W^{(l+1)}\sigma(h^{(l)}) + b^{(l+1)}$ and output $f = h^{(H+1)}$, with weights of variance $\sigma_w^2/\text{fan-in}$ and biases of variance σ_b^2 , this gives the recursion

$$K^{(1)}(x, x') = \sigma_w^2 \frac{x \cdot x'}{d} + \sigma_b^2, \quad K^{(l+1)}(x, x') = \sigma_w^2 \langle \sigma(u) \sigma(v) \rangle + \sigma_b^2, \quad (6.10)$$

where (u, v) is a centered Gaussian pair with covariance $\begin{pmatrix} K^{(l)}(x, x) & K^{(l)}(x, x') \\ K^{(l)}(x', x) & K^{(l)}(x', x') \end{pmatrix}$, and $K = K^{(H+1)}$. For ReLU and erf activations the Gaussian averages are known in closed form (Cho and Saul, 2009; Williams, 1997) (Problem 5).

At infinite width the entropy in (6.4) is $S[f] = -S_{\text{GP}}[f] + \text{const}$: the entropic cost of a function is half its squared RKHS norm. Inserting (6.9) into (6.3), the trained network is described by a Gaussian field theory,

$$Z = \int \mathcal{D}f e^{-S_{\text{GP}}[f] - E[f]/T}, \quad \frac{E[f]}{T} = \frac{1}{2\kappa^2} \sum_{\mu=1}^P (f(x_\mu) - y_\mu)^2, \quad \text{Controlled: } 1/N \quad (6.11)$$

with $\kappa^2 = T$. This is the field theory of Gaussian-process regression (Review §2.3): the prior of Section 3.3, conditioned on noisy observations at the training points.

6.4 From the field theory to the GP predictor

The action in (6.11) is quadratic in f , so the posterior is again a Gaussian process, and we need its mean and covariance. Both terms are quadratic forms under $d\mu_x$ once point evaluation is written as $f(x_\mu) = \int d\mu_x e_\mu(x) f(x)$, with $e_\mu(x) = \delta(x - x_\mu)/p(x)$:

$$S_{\text{GP}}[f] + \frac{E[f]}{T} = \frac{1}{2} \int d\mu_x d\mu_{x'} f(x) A(x, x') f(x') - \int d\mu_x b(x) f(x) + \text{const}, \quad (6.12)$$

$$A(x, x') = K^{-1}(x, x') + \frac{1}{\kappa^2} \sum_{\mu=1}^P e_\mu(x) e_\mu(x'), \quad b(x) = \frac{1}{\kappa^2} \sum_{\mu=1}^P y_\mu e_\mu(x).$$

Completing the square as in (3.2), with sums over components replaced by $\int d\mu_x$, the posterior mean is $\int d\mu_{x'} A^{-1}(x, x') b(x')$ and the posterior covariance is $A^{-1}(x, x')$, with the inverse taken under $d\mu_x$. Without data, $A = K^{-1}$ and this returns the prior covariance K .

Inverting A means inverting an operator on functions — an infinite matrix. But the data change it only by a term of rank P , and the Woodbury identity, valid for invertible A_0 and C ,

$$(A_0 + UCV)^{-1} = A_0^{-1} - A_0^{-1}U(C^{-1} + VA_0^{-1}U)^{-1}VA_0^{-1}, \quad \text{Exact} \quad (6.13)$$

trades it for a $P \times P$ inversion. Take $A_0 = K^{-1}$ and $C = \kappa^{-2}I_P$; let U map $v \in \mathbb{R}^P$ to the function $\sum_\mu v_\mu e_\mu$, and let V map f to $(f(x_\mu))_\mu$. Then $A_0^{-1}U$ is the row of functions $K(\cdot, x_\mu)$, $VA_0^{-1}U = K_{\mathcal{D}}$, and

$$A^{-1}(x, x') = K(x, x') - K(x, X)(K_{\mathcal{D}} + \kappa^2 I)^{-1}K(X, x'). \quad \text{Exact} \quad (6.14)$$

Function space has been traded for data space: the posterior covariance between any two points involves only the kernel between those points and the training set. The mean follows in one line, $\int d\mu_{x'} A^{-1}(x, x') b(x') = \kappa^{-2} A^{-1}(x, X) \mathbf{y} = \kappa^{-2} K(x, X) [I - (K_{\mathcal{D}} + \kappa^2 I)^{-1} K_{\mathcal{D}}] \mathbf{y}$. At a test point x_* ,

$$\langle f(x_*) \rangle = \mathbf{k}_*^\top (K_{\mathcal{D}} + \kappa^2 I)^{-1} \mathbf{y}, \quad \text{Var } f(x_*) = K_{**} - \mathbf{k}_*^\top (K_{\mathcal{D}} + \kappa^2 I)^{-1} \mathbf{k}_*, \quad \kappa^2 = T. \quad (6.15)$$

These are the GPR formulas (3.16), which Section 3.5 obtained by restricting to x_* and the training points from the start. Here they follow from a field theory defined on all inputs at once, which is the form that Lecture 2 averages over datasets. On the grid cutoff of Section 3.3, with x_* and the training points among the grid points, every step above is an identity between finite matrices, and the result does not depend on the grid at all. As $T \rightarrow 0$ the mean becomes kernel interpolation with the NNGP kernel, $\mathbf{k}_*^\top K_{\mathcal{D}}^{-1} \mathbf{y}$: the structure of the curve-fitting example (5.3), with K in place of K_M .

For fully connected networks on the sphere, K is a dot-product kernel, whose eigenfunctions are spherical harmonics; this is the setting of Lecture 2. What this picture leaves out: corrections of order $1/N$ at finite width, and regimes in which the network learns its own kernel. Both come later in the course.

For interpretability. In the GP limit the data leave the hidden layers untouched: the posterior statistics of every hidden representation equal their prior statistics, up to corrections that vanish as $N \rightarrow \infty$ at fixed P , and all the learning is in how fixed random features are combined. This makes the GP limit a precise null model. “Feature learning” means hidden representations whose statistics move away from the prior, and the interpretable features found in trained networks lie, by definition, outside this lecture.

7 Problems (optional homework)

1. **Spins, sources and entropy.** (a) A single spin $s \in \{-1, +1\}$ has energy $E(s) = -hs$. Compute Z , F , $\langle s \rangle$ and $\text{Var}(s)$ from $\log Z$, treating h/T as a source. Verify that $-\partial F/\partial T$ equals the Shannon entropy of $P(s)$. (b) For n independent spins, $E = -h \sum_j s_j$ depends on the configuration only through $m = \sum_j s_j$. Show that $\Omega(m) = \binom{n}{(n+m)/2}$. Using Stirling’s formula, minimize $F(m) = -hm - T \log \Omega(m)$ at large n and show that $m/n \rightarrow \tanh(h/T)$, in agreement with (a).
2. **Wick’s theorem.** Derive (3.3) by differentiating (3.1), and show that $\langle w^{2n} \rangle = (2n-1)!! v^n$ for $w \sim \mathcal{N}[0, v]$. How many terms does $\langle w_1 w_2 \cdots w_6 \rangle$ have?
3. **Gaussian-process regression, three ways.** (a) Derive the conditional law (3.4) by completing the square in $\mathbf{u}$. (b) Apply it to the pair $(f_*, \mathbf{y})$ with $\mathbf{y} = f(X) + \epsilon$, $\epsilon \sim \mathcal{N}[0, \kappa^2 I]$, and recover (3.16). (c) Show that if the likelihood in (3.14) is normalized as a density in $\mathbf{y}$, then $Z[0] = \mathcal{N}[0, K_{\mathcal{D}} + \kappa^2 I; \mathbf{y}]$, the marginal likelihood. (d) Prove the Woodbury identity (6.13) by multiplying its right-hand side by $A_0 + UCV$. On a grid that contains x_* and the training points, use it to rederive (6.15) from (6.11), and check that the grid drops out.
4. **Langevin dynamics.** (a) Solve the Ornstein–Uhlenbeck process $d\theta = -\eta k \theta dt + \sqrt{2\eta T} dW$, i.e. (4.1) with $L = \frac{1}{2} k \theta^2$. Show that θ_t stays Gaussian, find its mean and variance, and check that they relax to (4.3). What are the relaxation times? (b) Verify that the probability current of (4.2) vanishes at (4.3), and derive the expression for $dF[\rho_t]/dt$ in Section 4. (c) For a linear model $f_\theta(x) = \theta \cdot \psi(x)$ with the loss (4.5) and $\sigma_\alpha = \sigma$, show that the equilibrium mean prediction is kernel ridge regression with kernel $\sigma^2 \psi(x) \cdot \psi(x')$ and ridge T .
5. **Kernels in closed form.** Let (u, v) be a centered Gaussian pair with variances q_u, q_v , covariance c , and $\cos \vartheta = c/\sqrt{q_u q_v}$. (a) Show that $\langle \max(u, 0) \max(v, 0) \rangle = \frac{\sqrt{q_u q_v}}{2\pi} [\sin \vartheta + (\pi - \vartheta) \cos \vartheta]$ and

- $\langle H(u) H(v) \rangle = \frac{\pi - \vartheta}{2\pi}$, with H the Heaviside step function (the derivative of ReLU) (Cho and Saul, 2009). (b) Show that $\langle \text{erf}(u) \text{erf}(v) \rangle = \frac{2}{\pi} \arcsin \frac{2c}{\sqrt{(1+2q_u)(1+2q_v)}}$ (Williams, 1997). (c) Using (6.10), write the NNGP kernel of a one-hidden-layer ReLU network with $\sigma_b = 0$, for inputs normalized to $|x|^2 = d$, as a function of $x \cdot x'/d$.
- 6. Neal’s limit, up close.** Consider the one-hidden-layer network (6.5). (a) Show that, conditionally on the hidden weights $\mathbf{w}_i$, the functional prior is exactly Gaussian, with the empirical kernel $\hat{K}_N(x, x') = \frac{\sigma_a^2}{N} \sum_i \sigma(\mathbf{w}_i \cdot x) \sigma(\mathbf{w}_i \cdot x')$, and that $\hat{K}_N \rightarrow K$ by the law of large numbers. (b) Without conditioning, show that the fourth cumulant of $f_\theta(x)$ is $\frac{3\sigma_a^4}{N} \text{Var}_{\mathbf{w}} [\sigma(\mathbf{w} \cdot x)^2]$. The functional prior is Gaussian only at $N = \infty$; its $1/N$ corrections are where finite-width theory starts.
- 7. The power-law Legendre kernel.** (a) Prove the identity $(\Phi^\top \Phi + \kappa^2 I)^{-1} \Phi^\top = \Phi^\top (\Phi \Phi^\top + \kappa^2 I)^{-1}$ used in (5.3), and show that the ψ_ℓ are eigenfunctions of K_M with eigenvalues λ_ℓ under $d\mu_x = dx/2$. (b) Show that $\|K_M - K_\infty\|_1 = \sum_{\ell \geq M} \lambda_\ell \approx 2M^{-1/2}$, while $K_M(1, 1)$ grows like $4\sqrt{M}$: K_∞ is trace class yet unbounded at the endpoints. (Hint: $\psi_\ell(\pm 1)^2 = 2\ell + 1$.) (c) (*) Reproduce Fig. 1c. Repeat with i.i.d. uniform inputs instead of one per bin, and with exponentially decaying weights $\lambda_\ell = e^{-\ell/2}$. When does the second descent survive?

References

- Yakir Aharonov and David Bohm. Significance of electromagnetic potentials in the quantum theory. *Physical Review*, 115(3):485–491, 1959.
- Michael Aizenman and Hugo Duminil-Copin. Marginal triviality of the scaling limits of critical 4D Ising and ϕ_4^4 models. *Annals of Mathematics*, 194(1):163–235, 2021.
- V. E. Barnes et al. Observation of a hyperon with strangeness minus three. *Physical Review Letters*, 12(8):204–206, 1964.
- Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.
- Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.
- B. Andrei Bernevig, Taylor L. Hughes, and Shou-Cheng Zhang. Quantum spin Hall effect and topological phase transition in HgTe quantum wells. *Science*, 314(5806):1757–1761, 2006.
- Johan Bjorck, Alon Benham, Vishrav Chaudhary, Furu Wei, and Xia Song. Scaling optimal LR across token horizons. In *International Conference on Learning Representations (ICLR)*, 2025.
- Vladimir I. Bogachev. *Gaussian Measures*, volume 62 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 1998.
- Blake Bordelon, Lorenzo Noci, Mufan Bill Li, Boris Hanin, and Cengiz Pehlevan. Depthwise hyperparameter transfer in residual networks: Dynamics and scaling limit. In *International Conference on Learning Representations (ICLR)*, 2024.
- Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature Communications*, 12:2914, 2021.
- Olivier Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *IMS Lecture Notes Monograph Series*. Institute of Mathematical Statistics, 2007.
- Youngmin Cho and Lawrence K. Saul. Kernel methods for deep learning. In *Advances in Neural Information Processing Systems*, volume 22, 2009.

- Omry Cohen, Or Malka, and Zohar Ringel. Learning curves for overparametrized deep neural networks: A field theory perspective. *Physical Review Research*, 3(2):023034, 2021.
- DeepSeek-AI. DeepSeek LLM: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.
- Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence*, 2017.
- Crispin Gardiner. *Stochastic Methods: A Handbook for the Natural and Social Sciences*. Springer Series in Synergetics. Springer, fourth edition, 2009.
- James Glimm and Arthur Jaffe. *Quantum Physics: A Functional Integral Point of View*. Springer, New York, 2nd edition, 1987.
- Massimiliano Gubinelli and Martina Hofmanová. A PDE construction of the Euclidean ϕ_3^4 quantum field theory. *Communications in Mathematical Physics*, 384(1):1–75, 2021.
- Massimiliano Gubinelli, Peter Imkeller, and Nicolas Perkowski. Paracontrolled distributions and singular PDEs. *Forum of Mathematics, Pi*, 3:e6, 2015.
- Martin Hairer. A theory of regularity structures. *Inventiones Mathematicae*, 198(2):269–504, 2014.
- Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.
- Charles L. Kane and Eugene J. Mele. Quantum spin Hall effect in graphene. *Physical Review Letters*, 95: 226801, 2005.
- 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.
- Markus König, Steffen Wiedmann, Christoph Brüne, Andreas Roth, Hartmut Buhmann, Laurens W. Molenkamp, Xiao-Liang Qi, and Shou-Cheng Zhang. Quantum spin Hall insulator state in HgTe quantum wells. *Science*, 318(5851):766–770, 2007.
- Antti Kupiainen, Rémi Rhodes, and Vincent Vargas. Integrability of Liouville theory: Proof of the DOZZ formula. *Annals of Mathematics*, 191(1):81–166, 2020.
- Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as Gaussian processes. In *International Conference on Learning Representations*, 2018.
- Marco Loog, Tom Viering, Alexander Mey, Jesse H. Krijthe, and David M. J. Tax. A brief prehistory of double descent. *Proceedings of the National Academy of Sciences*, 117(20):10625–10626, 2020.
- Alexander G. de G. Matthews, Jiri Hron, Mark Rowland, Richard E. Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, 2018.
- Vaishnavh Nagarajan and J. Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.
- Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations*, 2020.
- Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations*, 2023.
- Gadi Naveh, Oded Ben David, Haim Sompolinsky, and Zohar Ringel. Predicting the outputs of finite deep neural networks trained with noisy gradients. *Physical Review E*, 104(6):064301, 2021.
- Radford M. Neal. Priors for infinite networks. In *Bayesian Learning for Neural Networks*, pages 29–53. Springer, 1996.

- Jeffrey Negrea, Gintare Karolina Dziugaite, and Daniel M. Roy. In defense of uniform convergence: Generalization via derandomization with an application to interpolating predictors. In *International Conference on Machine Learning*, 2020.
- Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.
- J. Robert Oppenheimer and Hartland Snyder. On continued gravitational contraction. *Physical Review*, 56(5):455–459, 1939.
- Giorgio Parisi and Yong-Shi Wu. Perturbation theory without gauge fixing. *Scientia Sinica*, 24(4):483, 1981.
- Grigorios A. Pavliotis. *Stochastic Processes and Applications: Diffusion Processes, the Fokker–Planck and Langevin Equations*, volume 60 of *Texts in Applied Mathematics*. Springer, 2014.
- Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.
- Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred A. Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, 2019.
- Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.
- Zohar Ringel, Noa Rubin, Edo Mor, Moritz Helias, and Inbar Seroussi. Applications of statistical field theory in deep learning. *arXiv preprint arXiv:2502.18553*, 2025.
- Noa Rubin, Inbar Seroussi, and Zohar Ringel. Grokking as a first order phase transition in two layer networks. In *International Conference on Learning Representations*, 2024.
- Andrew M. Stuart. Inverse problems: A Bayesian perspective. *Acta Numerica*, 19:451–559, 2010.
- Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, pages 681–688, 2011.
- Eugene Wigner. On unitary representations of the inhomogeneous Lorentz group. *Annals of Mathematics*, 40(1):149–204, 1939.
- Christopher K. I. Williams. Computing with infinite networks. In *Advances in Neural Information Processing Systems*, volume 9, 1997.
- Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs V: Tuning large neural networks via zero-shot hyperparameter transfer. *arXiv preprint arXiv:2203.03466*, 2022.
- Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.