The Local Learning Coefficient

Singular Learning Theory, Lecture 3
Michaelmas Term, 2026

1 Local Learning Coefficient **2**
1.1 Interpretations . . . 5
1.2 Local learning coefficients of deep linear networks . . . 6

2 Sketch Proof of the LLC **8**
2.1 The density of states . . . 8
2.2 Strategy . . . 10
2.3 The Mellin transform and the zeta function . . . 10
2.4 Poles of the zeta function . . . 11
2.5 Inverting the Mellin transform . . . 15
2.6 From the density of states to the volume . . . 16

References and further reading **18**

A Solutions **19**
A.1 Solutions to exercises from Section 1 . . . 20
A.2 Solutions to exercises from Section 2 . . . 27

1 Local Learning Coefficient

We have so far explored different kinds of qualitative degeneracy in the loss landscape. A natural question arises: *how can we quantify the degree of degeneracy at a particular point in parameter space?* In this lecture, we introduce the local learning coefficient, a rich mathematical object that reflects the degree of degeneracy at a parameter, and we explore it from several perspectives.

In this lecture (and in our later treatment of the free energy) we assume that any population loss L is a real analytic function.

The key idea is to measure the *volume* of near-optimal parameters. Consider a local minimum $w^* \in \mathcal{W}$ of the population loss L, and suppose we have a sufficiently small closed ball $B(w^*)$ centred on w^* such that $L(w) \geq L(w^*)$ for all $w \in B(w^*)$. Given a tolerance $\epsilon > 0$, define the **sublevel set**

$$B(w^*, \epsilon) = \{w \in B(w^*) \mid L(w) - L(w^*) < \epsilon\}, \tag{1}$$

consisting of all parameters in the ball whose loss is within ϵ of the minimum. The volume of this set is

$$V(\epsilon) := \mathrm{Vol}(B(w^*, \epsilon)) = \int_{B(w^*,\epsilon)} \mathrm{d}w. \tag{2}$$

We will look at how this volume scales as $\epsilon \to 0$. It will be instructive for us to first calculate this scaling in the non-degenerate case.

Exercise 1.1 (Volume scaling in the regular case)**.** Let $w^* \in \mathcal{W}$ be a local minimum of L, and suppose that the Hessian $\nabla^2 L(w^*)$ is positive definite. Show that

$$V(\epsilon) = c\,\epsilon^{d/2} + o(\epsilon^{d/2}) \quad \text{as } \epsilon \to 0, \tag{3}$$

for some constant $c > 0$ where $d = \dim(\mathcal{W})$. You may assume that, as $\epsilon \to 0$, the volume of $B(w^*, \epsilon)$ agrees to leading order with the volume of the quadratic approximation obtained by replacing L with its second-order Taylor expansion at w^*.
[*Hint:* The volume of the ellipsoid $\{\boldsymbol{x} \in \mathbb{R}^d \mid \boldsymbol{x}^\top A \boldsymbol{x} < 1\}$ is proportional to $(\det A)^{-1/2}$.]

The key takeaway from this exercise is that the dimension of the model controls the leading order asymptotics of the volume scaling. We can investigate this further by calculating the volume scaling in an example where the Hessian is not positive definite.

Exercise 1.2 (Quadratic Valley Volume Scaling)**.** Let $\mathcal{W} = [-R, R]^d$ for some (large) $R \in \mathbb{R}$ and let $L(x_1, \ldots, x_d) = \sum_{i=1}^{d'} x_i^2$.

(a) Sketch the loss landscape when $d = 2$, $d' = 1$.

(b) L achieves its minimum value 0 at multiple points in $\mathcal{W}$. What is the dimension of the space $\mathcal{W}_0 := \{w \in \mathcal{W} \,|\, L(w) = 0\}$?

(c) Calculate the Hessian of L.

(d) Prove that $V(\epsilon) \propto \epsilon^{d'/2}$

[*Hint:* Use the volume of the ellipsoid from Exercise 1.1]

In this example, the degenerate directions do not contribute to the leading exponent of $V(\epsilon)$. We see that the exponent $\frac{d}{2}$ from Exercise 1.1 is replaced with $\frac{d'}{2}$ where d' is the codimension of $\mathcal{W}_0$ i.e. the *effective dimension* of L which ignores the degenerate directions. This suggests that we can capture information about the degree of degeneracy of a loss function by looking at how the volume $V(\epsilon)$ scales as $\epsilon \to 0$. Watanabe (2009) proves a general form for this kind of volume scaling which has been adapted by Lau et al. (2025) into the following definition of the local learning coefficient.

Theorem 1.1 (Local learning coefficient)**.** *Let $w^* \in \mathcal{W}$ be a local minimum of the population loss L. There exists a unique rational number $\lambda(w^*) > 0$, a positive integer $\mu(w^*)$, and a constant $c > 0$ such that as $\epsilon \to 0$,*

$$V(\epsilon) = c\,\epsilon^{\lambda(w^*)}(-\log \epsilon)^{\mu(w^*)-1} + o\big(\epsilon^{\lambda(w^*)}(-\log \epsilon)^{\mu(w^*)-1}\big)\,. \tag{4}$$

*We call $\lambda(w^*)$ the **local learning coefficient (LLC)** at w^*, and $\mu(w^*)$ the **local multiplicity**.*

In the next section we will sketch a proof of this theorem which will require some of the algebro-geometric machinery developed in the previous lecture. First, we explore some properties of the LLC and its multiplicity.

Exercise 1.3 (Two equations for λ)**.**

(a) Prove that for $a > 1$

$$\lambda = \lim_{\epsilon \to 0} \frac{\log(V(a\epsilon)/V(\epsilon))}{\log(a)} \tag{5}$$

(b) Prove that

$$\lambda = \lim_{\epsilon \to 0} \frac{\log V(\epsilon)}{\log \epsilon} \tag{6}$$

Note that from equation (3) we already have that the LLC and local multiplicity in the regular case are $\lambda = \frac{d}{2}$, $\mu = 1$

When $\mu(w^*) = 1$, the formula simplifies to $V(\epsilon) = c\epsilon^{\lambda(w^*)} + o(\epsilon^{\lambda(w^*)})$ which allows us to interpret $\lambda(w^*)$ as the *volume scaling exponent*: increasing the error tolerance by a factor of a increases the volume of near-optimal parameters by a factor of $a^{\lambda(w^*)}$.

Exercise 1.4 (Calculating the LLC)**.** Calculate the LLC and the rank of the Hessian at the global minimum in the following cases:

(a) $L(x) = x^2$

(b) $L(x) = x^{2m}$

(c) $L(x, y) = x^4 + y^4$

(d) $L(x, y) = x^2 + y^4$

Exercise 1.5 (Upper bound on the LLC)**.** In Exercise 1.1, we saw that $\lambda = d/2$ for regular models. In this exercise, you will prove directly from the volume scaling definition that for any local minimum w^* of the population loss, the local learning coefficient satisfies $\lambda(w^*) \leq d/2$.

(a) Let $\Lambda > 0$ be an upper bound on the eigenvalues of the Hessian $\nabla^2 L(w)$ for all $w \in B(w^*)$. Using the Lagrange remainder form of Taylor's theorem, show that for all $w \in B(w^*)$,

$$L(w) - L(w^*) \leq \frac{1}{2}\Lambda \|w - w^*\|^2.$$

[*Hint:* for any real symmetric matrix $H \in \mathbb{R}^{n \times n}$ and vector $v \in \mathbb{R}^n$, $v^\top H v \leq \Lambda \|v\|^2$ where Λ is the largest eigenvalue of H.]

(b) Consider the standard Euclidean ball of radius r centred at w^*, denoted $B_r(w^*) = \{w \in \mathcal{W} \mid \|w - w^*\| < r\}$. Find a radius r (as a function of ϵ and Λ) such that for sufficiently small ϵ,

$$B_r(w^*) \subseteq B(w^*, \epsilon).$$

(c) Recall that the volume of a d-dimensional Euclidean ball of radius r is proportional to r^d. Use your result from part (b) to show that there exists a constant $C > 0$ such that for sufficiently small ϵ,

$$C\epsilon^{d/2} \leq V(\epsilon).$$

(d) Conclude that $\lambda(w^*) \leq d/2$.

We have now shown that the LLC satisfies $\lambda(w^*) \leq d/2$, with equality precisely in the regular (non-degenerate) case. It is also possible (although quite fiddly) to prove that $\frac{r}{2} \leq \lambda(w^*)$ where $r = \text{rank}\nabla^2 L(w^*)$ and as we will see in the next exercise, the LLC captures more information about the geometry than just the Hessian rank.

Exercise 1.6 (LLC vs Hessian Rank)**.** In this exercise we will show that the LLC can detect differences in the local geometry that is not reflected in the rank of the Hessian. Construct two loss functions L_1 and L_2 with minima w_1 and w_2 respectively such that $\text{rank}\,\nabla^2 L_1(w_1) = \text{rank}\,\nabla^2 L_2(w_2)$ but $\lambda(w_1) \neq \lambda(w_2)$.

Exercise 1.7 (The cubically-parameterised loss)**.** We can make a quadratic loss degenerate by changing the parameterisation. Let the *cubically-parameterised loss* be

$$L(\mu) = \tfrac{1}{2}(\mu^3 - \mu_0^3)^2, \qquad \mu \in \mathbb{R} \tag{7}$$

obtained from the ordinary quadratic loss $\ell(\theta) = \frac{1}{2}(\theta - \theta_0)^2$ via the reparameterisation $\theta = \mu^3$, $\theta_0 = \mu_0^3$.

(a) Show that the cubically-parameterised loss is just as expressive as the ordinary quadratic loss: that is, they both have the same set of achievable minima.

(b) Compute the Hessian $L''(\mu_0)$ to show that the loss is degenerate at its minimum when $\mu_0 = 0$, and non-degenerate when $\mu_0 \neq 0$.

(c) Give an explicit formula for $V(\epsilon)$ when $\mu_0 = 0$ and $\mu_0 \neq 0$ separately.

(d) Use your answer to part (c) to find the learning coefficient for $\mu_0 = 0$ and for $\mu_0 \neq 0$. Given that the model is non-degenerate when $\mu_0 \neq 0$, we expect the learning coefficient to be $d/2$ in that case; compare your answer. How does the cubically-parameterised loss at $\mu_0 = 0$ differ from the ordinary quadratic loss?

(e) Instead of taking $\epsilon \to 0$ to get the learning coefficient, fix a small but nonzero value for ϵ, such as $\epsilon = 0.01$. Plot $V(\epsilon)$ as a function of μ_0. As we saw from (d), the learning coefficient changes discontinuously when $\mu_0 = 0$—what happens with $V(\epsilon)$ as μ_0 gets close to zero? What changes if you make ϵ smaller or larger?

Even though the asymptotic *learning coefficient* ($\epsilon \to 0$) only changes when $\mu_0 = 0$ exactly, note how the non-asymptotic *volume* (ϵ finite) is affected in a larger neighbourhood.

1.1 Interpretations

In this section, we survey two alternative perspectives on the LLC, from information theory and geometric measure theory. The algebro-geometric perspective is treated separately in Section 2.4.

Information-theoretic interpretation. The LLC also admits a natural interpretation in terms of information (Lau et al., 2025). The number of bits needed to specify the sublevel set $B(w^*, \epsilon)$ within the ball $B(w^*)$ is

$$-\log_2 \frac{V(\epsilon)}{\mathrm{Vol}(B(w^*))}. \tag{8}$$

For small ϵ and $\mu(w^*) = 1$, this is approximately $-\lambda(w^*)\log_2 \epsilon + O(\log_2 \log_2 \epsilon)$. In particular, the number of *additional* bits needed to halve an already small error tolerance from ϵ to $\epsilon/2$ is

$$-\log_2 \frac{V(\epsilon/2)}{V(\epsilon)} \approx \lambda(w^*). \tag{9}$$

That is, the LLC $\lambda(w^*)$ measures the number of bits required to halve the error near w^*.

Local learning coefficient as a fractal dimension The LLC admits a geometric interpretation: it is a *fractal dimension* of the parameter space, as measured through the lens of the loss function. To make this precise, we recall a standard notion from fractal analysis.

Definition 1.2 (Hölder exponent)**.** Let (X, ρ) be a (pseudo-)metric space and ν a measure on X. The **Hölder exponent** (or **local dimension**) $\alpha(x)$ of ν at a point $x \in X$ is given by

$$\alpha(x) = \lim_{\epsilon \to 0} \frac{\log \nu(B_\epsilon(x))}{\log \epsilon}, \tag{10}$$

where $B_\epsilon(x)$ is the ball of radius ϵ centred at x. Equivalently, $\alpha(x)$ is the unique real number such that

$$\nu(B_\epsilon(x)) = O(\epsilon^{\alpha(x)}) \qquad \text{as } \epsilon \to 0.$$

The Hölder exponent generalises the notion of dimension: when ν is the Lebesgue measure on $\mathbb{R}^d$ and ρ is the Euclidean metric, every point has $\alpha(x) = d$, but for measures concentrated on fractal sets the exponent can be non-integer. Notice that the formula (10) has a similar form to the definition of the local learning coefficient (5). To make this connection precise, we introduce a pseudo-metric induced by the loss function.

Definition 1.3 (Loss pseudo-metric)**.** The **loss pseudo-metric** on $\mathcal{W}$ is defined by

$$\rho(w, u) = |L(w) - L(u)|. \tag{11}$$

Remark 1.4. ρ is only a pseudo-metric as the distance between two distinct points can be zero.

Under this pseudo-metric, the sublevel set (1) is precisely the ϵ-ball around w^*:

$$B(w^*, \epsilon) = \{w \in B(w^*) \mid \rho(w, w^*) < \epsilon\}, \tag{12}$$

so $V(\epsilon) = \nu(B(w^*, \epsilon))$.

Exercise 1.8 (Hölder exponent and learning coefficient)**.** Let $\alpha(w^*)$ be the Hölder exponent of the Lebesgue measure on $\mathcal{W}$ at w^* under the loss pseudo-metric. Show that

$$\lambda(w^*) = \alpha(w^*).$$

1.2 Local learning coefficients of deep linear networks

Having established the local learning coefficient as a measure of degeneracy in the loss landscape, we now compute it explicitly for the two-layer deep linear network introduced in Lecture 1 (Section 3). In Lecture 1 (Section 4), we saw that DLN parameters exhibit varying degrees of degeneracy depending on the ranks of the component matrices (Exercise 4.9 and Remark 4.4 of Lecture 1). The LLC makes this hierarchy quantitative: different ranks correspond to different learning coefficients, confirming that lower-rank parameters are geometrically "simpler."

Consider the two-layer DLN $f_{A,B}(x) = BAx$ with $A \in \mathbb{R}^{H\times M}$ and $B \in \mathbb{R}^{N\times H}$, so the parameter space is $\mathcal{W} = \mathbb{R}^{(M+N)H}$. Suppose the true input–output relationship is $y = B_0A_0x$ where $A_0 \in \mathbb{R}^{H\times M}$, $B_0 \in \mathbb{R}^{N\times H}$, and $r = \operatorname{rank}(B_0A_0)$, and consider the population loss

$$L(A, B) = \mathbb{E}_{x\sim q(x)}\left[\|BAx - B_0A_0x\|^2\right], \tag{13}$$

where $q(x)$ has full-rank covariance. The set of minima of this loss function is $W_0 = \{(A, B) : BA = B_0A_0\}$. Aoyagi and Watanabe (2005) derived an exact formula for the learning coefficient at these minima.

Theorem 1.5 (Aoyagi and Watanabe, 2005)**.** *For the two-layer DLN $f_{A,B}(x) = BAx$ with $A \in \mathbb{R}^{H\times M}$, $B \in \mathbb{R}^{N\times H}$, and true parameter (A_0, B_0) with $r = \operatorname{rank}(B_0A_0)$, at any minimum $w^* \in W_0$ the local learning coefficient λ and its multiplicity μ are given by the following cases:*

1. *If* $N + r \leq M + H, \quad M + r \leq N + H, \quad$ *and* $\quad H + r \leq M + N$*:*

   (a) *If* $M + H + N + r$ *is even, then* $\mu = 1$ *and*

   $$\lambda = \frac{-(H+r)^2 - M^2 - N^2 + 2(H+r)M + 2(H+r)N + 2MN}{8}.$$

   (b) *If* $M + H + N + r$ *is odd, then* $\mu = 2$ *and*

   $$\lambda = \frac{-(H+r)^2 - M^2 - N^2 + 2(H+r)M + 2(H+r)N + 2MN + 1}{8}.$$

2. *If* $M + H < N + r$*, then* $\mu = 1$ *and* $\lambda = \frac{HM - Hr + Nr}{2}$.

3. *If* $N + H < M + r$*, then* $\mu = 1$ *and* $\lambda = \frac{HN - Hr + Mr}{2}$.

4. *If* $M + N < H + r$*, then* $\mu = 1$ *and* $\lambda = \frac{MN}{2}$.

Exercise 1.9 (Learning coefficients of two-layer DLNs)**.** Consider the constant-width two-layer DLN from Exercise 4.9 of Lecture 1: $y = BAx$, where $x \in \mathbb{R}^m$ are the inputs, $y \in \mathbb{R}^m$ are the outputs, and $A, B \in \mathbb{R}^{m \times m}$ are linear transformations. The parameter space is $\mathcal{W} = \mathbb{R}^{2m^2}$ and the parameters are the pair $(A, B) \in \mathcal{W}$.

(a) Aoyagi & Watanabe do not assume the network is constant-width, like we do. Simplify their formula for λ using this assumption.

(b) How does the learning coefficient λ depend on r, the rank of $B_0 A_0$? Compare with the results of Exercise 4.9(d) of Lecture 1.

(c) How does λ compare with the effective parameter count $d/2$ of a regular model? What does this tell us about the role of singularities in the two-layer DLN?

2 Sketch Proof of the LLC

In this section we sketch a proof of Theorem 1.1. We explain where the asymptotic form $V(\epsilon) \approx c\,\epsilon^{\lambda}(-\log\epsilon)^{\mu-1}$ comes from and along the way we obtain an algebro-geometric formula for λ and μ. Throughout, w^* is a local minimum of L with closed ball $B(w^*)$ as in Section 1, and we write

$$K(w) := L(w) - L(w^*) \geq 0, \qquad w \in B(w^*), \tag{14}$$

so that $V(\epsilon) = \mathrm{Vol}\{w \in B(w^*) \mid K(w) < \epsilon\}$.

2.1 The density of states

The volume $V(\epsilon)$ adds up the contributions of all level sets $\{K = t\}$ with $t < \epsilon$. It is convenient to isolate the contribution of a single level set for which we will need the Dirac delta.

The Dirac delta. Informally, we can think of the Dirac delta $\delta(t)$ as a "function" satisfying

$$\delta(t) = \begin{cases} \infty, & t = 0, \\ 0, & t \neq 0, \end{cases} \qquad \int_{-\infty}^{\infty} \delta(t)\,\mathrm{d}t = 1.$$

One way to make this concrete is as the limit as $\eta \to 0$ of the box functions

$$\delta_\eta(t) = \begin{cases} 1/\eta & -\eta < t \leq 0, \\ 0 & \text{otherwise}, \end{cases} \tag{15}$$

which have width η and height $1/\eta$. For any continuous function F,

$$\int_{\mathbb{R}} F(t)\,\delta_\eta(t-a)\;\mathrm{d}t = \frac{1}{\eta}\int_{a-\eta}^{a} F(t)\;\mathrm{d}t \;\xrightarrow{\eta\to 0}\; F(a).$$

We will only use δ inside an integral, so we simply *define* it by this **sifting property**:[1]

$$\int_{\mathbb{R}} F(t)\,\delta(t-a)\;\mathrm{d}t = F(a). \tag{16}$$

Definition 2.1 (Density of states)**.** The **density of states (DOS)** of L at w^* is the function

$$v(t) = \int_{B(w^*)} \delta\big(t - K(w)\big)\;\mathrm{d}w, \qquad t \in \mathbb{R}. \tag{17}$$

To see what this measures, replace δ by the box δ_η. The integrand is $1/\eta$ exactly when $t \leq K(w) < t+\eta$, so

$$v(t) \approx \frac{1}{\eta}\,\mathrm{Vol}\{w \in B(w^*) \mid t \leq K(w) < t+\eta\} = \frac{V(t+\eta) - V(t)}{\eta}. \tag{18}$$

In other words, $v(t)\;\mathrm{d}t$ is the volume of the thin shell of parameters whose loss lies between t and $t+\mathrm{d}t$ (Figure 1). Letting $\eta \to 0$ suggests that $v = \frac{\mathrm{d}V}{\mathrm{d}t}$. Both this and a more general identity follow directly from the sifting property.

[1] The delta is not a genuine function but a *Schwartz distribution* which is a linear functional on smooth test functions. See Watanabe (2009, Chapter 4).

Proposition 2.2 (Properties of the DOS)**.**

(a) For any function F (for which the integrals make sense),

$$\int_0^\infty F(t)\, v(t)\, \mathrm{d}t = \int_{B(w^*)} F\big(K(w)\big)\, \mathrm{d}w. \tag{19}$$

(b) $V(\epsilon) = \int_0^\epsilon v(t)\, \mathrm{d}t$. In other words, $v = \frac{\mathrm{d}V}{\mathrm{d}t}$.

Proof. For (a), substitute (17), swap the order of integration and apply the sifting property

$$\int_{\mathbb{R}} F(t)\, v(t)\, \mathrm{d}t = \int_{B(w^*)} \int_{\mathbb{R}} F(t)\, \delta\big(t - K(w)\big)\, \mathrm{d}t\, \mathrm{d}w = \int_{B(w^*)} F\big(K(w)\big)\, \mathrm{d}w.$$

Since $K \geq 0$ we have $v(t) = 0$ for $t < 0$, so the t-integral can be restricted to $[0, \infty)$.

For (b), let F be the indicator function on $[0, \epsilon)$ giving

$$\int_0^\epsilon v(t)\, \mathrm{d}t = \int_{B(w^*)} \mathbf{1}_{[0,\epsilon)}\big(K(w)\big)\, \mathrm{d}w = \mathrm{Vol}\{w \in B(w^*) \mid K(w) < \epsilon\} = V(\epsilon).$$

□

Coarea Formula We can equivalently write the density of states using the coarea formula

$$v(t) = \int_{\{K=t\}} \frac{\mathrm{d}\sigma(w)}{\|\nabla K(w)\|}, \tag{20}$$

where $\mathrm{d}\sigma$ is the $(d-1)$-dimensional surface measure on the level set $\{K = t\}$. We now sketch a geometric argument for why this formula holds.

By Proposition 2.2, $v(t)\, \mathrm{d}t$ is the volume of the thin shell $\{t \leq K < t + \mathrm{d}t\}$ between the level sets $\{K = t\}$ and $\{K = t + \mathrm{d}t\}$. To compute this volume, cut the shell into thin vertical strips as in Figure 1. The strip at $w \in \{K = t\}$ has as its base a small patch of the $(d-1)$-dimensional level set with area $\mathrm{d}\sigma(w)$. Its height is the distance $h(w)$ across the shell at w.

To find $h(w)$, note that ∇K is normal to the level set and write $\hat{n} = \nabla K / \|\nabla K\|$ for the unit normal. By taylor expanding K to first order we see that

$$t + \mathrm{d}t = K(w + h\,\hat{n}) \approx K(w) + h\, \nabla K(w) \cdot \hat{n} = t + h\, \|\nabla K(w)\|$$

so $h = \frac{\mathrm{d}t}{\|\nabla K(w)\|}$. Each strip is approximately a rectangle so its volume is $h(w)\, \mathrm{d}\sigma(w)$. Summing over the strips we get that the total volume is

$$v(t)\, \mathrm{d}t = \sum_{\text{strips}} \frac{\mathrm{d}t}{\|\nabla K(w)\|}\, \mathrm{d}\sigma(w) \;\longrightarrow\; \mathrm{d}t \int_{\{K=t\}} \frac{\mathrm{d}\sigma(w)}{\|\nabla K(w)\|},$$

which gives the result.

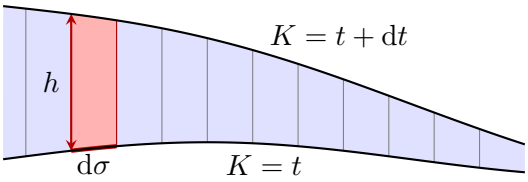


Figure 1: A piece of the shell $\{t \le K < t + \mathrm{d}t\}$, which has total volume $v(t)\,\mathrm{d}t$, cut into strips of base $\mathrm{d}\sigma$ and height $h = \mathrm{d}t/\|\nabla K\|$. The strips are thick where the loss is flat and thin where it is steep.

2.2 Strategy

By Proposition 2.2(b), the behaviour of $V(\epsilon)$ as $\epsilon \to 0$ is controlled by the behaviour of $v(t)$ as $t \to 0$. We will show that

$$v(t) = c'\, t^{\lambda-1}(-\log t)^{\mu-1} + (\text{lower order terms}) \qquad \text{as } t \to 0, \tag{21}$$

and then integrate to get the asymptotic form of V. To proceed we take a detour through an integral transform which will allow us to use tools from complex analysis and algebraic geometry developed in the last lecture.

1. Apply the **Mellin transform** to v to give the **zeta function** $\zeta(z) = \int K(w)^z \,\mathrm{d}w$.
2. Use **resolution of singularities** to show that ζ extends to a meromorphic function on $\mathbb{C}$, and locate its **poles**.
3. Apply the **inverse Mellin transform** to the pole expansion of ζ. The rightmost pole gives the leading term of the state density expansion (21)
4. Integrate v to obtain V.

2.3 The Mellin transform and the zeta function

Definition 2.3 (Mellin transform)**.** The **Mellin transform** of a function $f\colon (0,\infty) \to \mathbb{R}$ is

$$\mathcal{M}[f](z) = \int_0^\infty f(t)\, t^z \,\mathrm{d}t, \tag{22}$$

defined for $z \in \mathbb{C}$ where integral converges.[2]

Taking $F(t) = t^z$ in Proposition 2.2(a), the Mellin transform of the density of states is

$$\mathcal{M}[v](z) = \int_{B(w^*)} K(w)^z \,\mathrm{d}w. \tag{23}$$

Definition 2.4 (Zeta function)**.** The **Zeta function** of L at w^* is

$$\zeta(z) = \int_{B(w^*)} (L(w) - L(w^*))^z \,\mathrm{d}w, \qquad z \in \mathbb{C}. \tag{24}$$

[2]Most references put t^{z-1} in place of t^z. We follow Watanabe (2009), whose convention makes (23) hold without a shift.

Since K is positive and bounded on $B(w^*)$, this integral converges and defines a holomorphic function for $\operatorname{Re} z > 0$. For $\operatorname{Re} z < 0$ the integrand K^z diverges on the set where $K = 0$. Note that the integral may still converge but this needs further analysis.

Example 2.5. Take $K(x) = x^2$ on $B = [-1, 1]$. Then

$$\zeta(z) = \int_{-1}^{1} x^{2z} \, \mathrm{d}x = \frac{2}{2z+1} = \frac{1}{z+\frac{1}{2}},$$

where the integral converges if and only if $\operatorname{Re} z > -\frac{1}{2}$. The right-hand side, however, makes sense for every $z \neq -\frac{1}{2}$: it *extends* ζ to a meromorphic function on $\mathbb{C}$ with a single simple pole at $z = -\frac{1}{2}$. Recall from Exercise 1.4 that $\lambda = \frac{1}{2}$ and $\mu = 1$ for $K(x) = x^2$.

2.4 Poles of the zeta function

To generalise this example we will need Hironaka's resolution of singularities. We will require the notion of a real analytic manifold which is a manifold whose transition maps between charts are real analytic functions.

We will write the zero set of K as

$$\mathcal{W}_0 := \{w \in \mathcal{W} \mid K(w) = 0\}$$

Theorem 2.6 (Resolution of singularities; Hironaka (1964); Watanabe (2009))**.** *Let $K \colon \mathcal{W} \to \mathbb{R}$ be real analytic, with $K \geq 0$ and K not identically zero. Then there is a d-dimensional real analytic manifold $\widetilde{\mathcal{W}}$ and a proper real analytic map $g \colon \widetilde{\mathcal{W}} \to \mathcal{W}$ such that:*

1. *g is an analytic isomorphism from $\widetilde{\mathcal{W}} \setminus g^{-1}(\mathcal{W}_0)$ onto $\mathcal{W} \setminus \mathcal{W}_0$;*
2. *around each point of $g^{-1}(\mathcal{W}_0)$ there are local coordinates $u = (u_1, \ldots, u_d)$ on $\widetilde{\mathcal{W}}$, centred at that point, in which*

$$\begin{aligned} K(g(u)) &= u_1^{2k_1} \cdots u_d^{2k_d}, \\ \left|\det Dg(u)\right| &= b(u) \, |u_1|^{h_1} \cdots |u_d|^{h_d}. \end{aligned} \tag{25}$$

   Here $k_p, h_p \geq 0$ are integers, at least one k_p is positive, and b is real analytic with $b(u) > 0$.

Here $Dg(u)$ is the Jacobian matrix of g written in the local coordinates u. The point of the theorem is that, although the zero set $\mathcal{W}_0$ may be badly singular, after pulling back along g both K and the volume factor $|\det Dg|$ become *monomials* in suitable coordinates. Such an expression is said to be in **normal crossing form**.

Since K is analytic and not identically zero, $\mathcal{W}_0$ has measure zero. Because g is an isomorphism away from $\mathcal{W}_0$, we may therefore change variables $w = g(u)$ in integrals over $\mathcal{W}$, picking up the factor $|\det Dg(u)|$. For polynomial K we can construct g explicitly as a finite sequence of blow-ups.

Theorem 2.7 (Meromorphic continuation)**.** *The zeta function extends to a meromorphic function on $\mathbb{C}$. All its poles are negative rational numbers $-\lambda_1 > -\lambda_2 > \cdots$, of orders $\mu_1, \mu_2, \ldots$, and for every $R > 0$,*

$$\zeta(z) = \sum_{i \colon \lambda_i < R} \sum_{m=1}^{\mu_i} \frac{c_{i,m}}{(z+\lambda_i)^m} + H_R(z), \tag{26}$$

with H_R holomorphic on $\operatorname{Re} z > -R$. The rightmost pole, at $z = -\lambda$ with order μ, has a positive leading coefficient.

Here $\lambda = \lambda_1$ and $\mu = \mu_1$. Only finitely many poles lie to the right of any vertical line, so each sum in (26) is finite. The remainder H_R is essential, and we read (26) like Taylor's theorem with remainder: one statement for each R.

Proof. First recall the definition of the zeta function, $\zeta(z) = \int_{B(w^*)} K(w)^z \, \mathrm{d}w$. Since K is bounded on $B(w^*)$, this integral converges and is holomorphic for $\operatorname{Re} z > 0$. Near each pole, the terms in (26) are the principal part of the Laurent expansion of ζ, and that much follows from meromorphy alone. The real work is to show that ζ can be continued to a meromorphic function, to find the poles and their orders, and to show that the remainder H_R is holomorphic on the whole half-plane $\operatorname{Re} z > -R$; the last of these comes from the Taylor argument in Step 5. The plan is to rewrite the zeta function in coordinates where it can be computed explicitly, and then read off the poles.

Step 1: resolve the singularities. We use the resolution of singularities from Lecture 2. Since K is analytic, there is a real analytic manifold $\widetilde{\mathcal{W}}$ and a proper analytic map $g\colon \widetilde{\mathcal{W}} \to \mathcal{W}$, which is an isomorphism away from $\{K = 0\}$, such that, around each point of $g^{-1}(B(w^*))$ where $K \circ g$ vanishes, there are local coordinates $u = (u_1, \dots, u_d)$ in which

$$K(g(u)) = u_1^{2k_1} \cdots u_d^{2k_d}, \qquad \left|\det Dg(u)\right| = b(u) \left|u_1^{h_1} \cdots u_d^{h_d}\right|, \tag{27}$$

for non-negative integers k_p, h_p and an analytic function $b(u) > 0$. At least one k_p is positive, since $K \circ g$ vanishes at the centre of the chart. Since $\{K = 0\}$ has measure zero, we can substitute $w = g(u)$:

$$\zeta(z) = \int_{g^{-1}(B(w^*))} K(g(u))^z \left|\det Dg(u)\right| \, \mathrm{d}u.$$

This is an integral over the resolution manifold, so we compute it chart by chart using a partition of unity.

Step 2: choose charts. Since g is proper, $g^{-1}(B(w^*))$ is compact, so we can cover it by finitely many charts. We use two types:

- *Type 1:* charts on which $K \circ g \geq \delta > 0$ on the support of the chart weight.
- *Type 2:* charts that intersect $\{K \circ g = 0\}$. By the resolution theorem, (27) holds on these.

Type 1 charts do not contribute any poles to ζ since if $K \circ g \geq \delta > 0$ then $(K \circ g)^z = e^{z \log(K \circ g)}$ is perfectly well behaved for *every* $z \in \mathbb{C}$. The contribution of such a chart to ζ is an entire function of z. All of the poles come from the type 2 charts.

Step 3: split the integral into pieces. Cover the type 2 charts by finitely many small boxes $(-\delta,\delta)^d$ and rescale each to $(-1,1)^d$. Keep the boxes that intersect $\{K \circ g = 0\}$. The rest have $K \circ g > 0$, so (after shrinking them slightly) they are of type 1. Take a partition of unity $\{\rho_\alpha\}$ for the boxes and type 1 charts, chosen so that each ρ_α is strictly positive on its box or chart. Then split each remaining box into its 2^d orthants (regions where each coordinate has a fixed sign) and reflect the negative coordinates $u_p \mapsto -u_p$, so that every piece becomes $[0,1)^d$. Reflection preserves the normal crossing form (27), since the exponents $2k_p$ are even and the Jacobian only involves $|u_p|$. We can close $[0,1)^d$ to $[0,1]^d$ without changing the integral. From now on α labels these pieces.

Step 4: the zeta function in the new coordinates. Using (27), we get

$$\zeta(z) = \sum_\alpha \zeta_\alpha(z) + \zeta_+(z),$$

where ζ_+ is the entire function coming from the type 1 charts, and

$$\begin{aligned} \zeta_\alpha(z) &= \int_{[0,1]^d} u_1^{2k_1 z} \cdots u_d^{2k_d z}\, u_1^{h_1} \cdots u_d^{h_d}\, b_\alpha(u)\, \mathrm{d}u \\ &= \int_{[0,1]^d} u_1^{2k_1 z + h_1} \cdots u_d^{2k_d z + h_d}\, b_\alpha(u)\, \mathrm{d}u. \end{aligned} \tag{28}$$

Here b_α absorbs b, ρ_α and the constant Jacobian of the rescaling. It is smooth and nonnegative, with $b_\alpha(0) > 0$, because the centre of the box lies inside the chart, where $\rho_\alpha > 0$. hThe exponents k_p, h_p depend on α, but we suppress this to keep the notation light.

Step 5: one variable. To see where the poles come from, first take $d = 1$, so that $\zeta_\alpha(z) = \int_0^1 u^{2kz+h}\, b(u)\, \mathrm{d}u$ with $k \geq 1$. Taylor expand $b(u) = \sum_{j=0}^{n-1} a_j u^j + u^n r_n(u)$, where r_n is smooth, and integrate term by term:

$$\begin{aligned} \zeta_\alpha(z) &= \int_0^1 u^{2kz+h} \Big(\sum_{j=0}^{n-1} a_j u^j + u^n r_n(u) \Big)\, \mathrm{d}u \\ &= \sum_{j=0}^{n-1} \frac{a_j}{2kz + h + j + 1} \;+\; \int_0^1 u^{2kz+h+n}\, r_n(u)\, \mathrm{d}u. \end{aligned} \tag{29}$$

The sum has possible simple poles at $z = -\frac{h+j+1}{2k}$, when $a_j \neq 0$, and each of its terms is otherwise defined on all of $\mathbb{C}$. The remainder has n extra powers of u, so it converges and is holomorphic for $\operatorname{Re} z > -\frac{h+n+1}{2k}$. Taking n larger moves this half-plane as far left as we like: each extra order of the Taylor expansion reveals one more possible pole, further to the left than all the previous ones. Thus, for every $R > 0$, ζ_α is a finite sum of principal parts plus a holomorphic remainder on $\operatorname{Re} z > -R$.

Step 6: several variables. Take $d = 2$ with $k_1, k_2 > 0$. Do the u_1 integral first, treating u_2 as a parameter:

$$\zeta_\alpha(z) = \int_0^1 u_2^{2k_2 z + h_2} \left[\int_0^1 u_1^{2k_1 z + h_1}\, b(u_1, u_2)\, \mathrm{d}u_1 \right] \mathrm{d}u_2.$$

Taylor expand in u_1 only: $b(u_1,u_2)=\sum_{j=0}^{n-1} a_j(u_2)\,u_1^j + u_1^n\,r_n(u_1,u_2)$. The inner integral is then

$$\int_0^1 u_1^{2k_1z+h_1}\,b(u_1,u_2)\;\mathrm{d}u_1 = \sum_{j=0}^{n-1}\frac{a_j(u_2)}{2k_1z+h_1+j+1} + H_n(z,u_2),$$

with H_n holomorphic for $\operatorname{Re} z > -\frac{h_1+n+1}{2k_1}$, and smooth in u_2. Putting this back into the u_2 integral,

$$\begin{aligned}\zeta_\alpha(z) = &\sum_{j=0}^{n-1}\frac{1}{2k_1z+h_1+j+1}\int_0^1 u_2^{2k_2z+h_2}\,a_j(u_2)\;\mathrm{d}u_2\\ &+\int_0^1 u_2^{2k_2z+h_2}\,H_n(z,u_2)\;\mathrm{d}u_2.\end{aligned}$$

Each u_2 integral is the one-variable case again, with possible simple poles at $z=-\frac{h_2+l+1}{2k_2}$ and a holomorphic remainder on a right half-plane. (In the last integral the weight $H_n(z,u_2)$ also depends holomorphically on z, but the same argument applies.) So the poles of ζ_α lie on two *ladders*,

$$z=-\frac{h_1+j+1}{2k_1}\qquad\text{and}\qquad z=-\frac{h_2+l+1}{2k_2}.$$

Where they meet, a u_1 factor times a u_2 factor can give a double pole.

In d variables we expand one coordinate at a time. Each coordinate with $k_p>0$ contributes its own ladder of possible poles, and coordinates with $k_p=0$ contribute none. Hence ζ_α extends to a meromorphic function on all of $\mathbb{C}$ which, for every $R>0$, is a finite sum of principal parts plus a holomorphic remainder on $\operatorname{Re} z > -R$, and whose poles lie among the negative rational numbers

$$z=-\frac{h_p+j+1}{2k_p},\qquad k_p>0,\quad j=0,1,2,\dots \tag{30}$$

The order of each of these poles is the number of coordinates whose ladders pass through that point, so poles can be up to order d.

Step 7: read off the rightmost pole. In chart α, each coordinate with $k_p>0$ has a ladder of possible poles

$$z=-\frac{h_p+1+j_p}{2k_p},\qquad j_p=0,1,2,\dots$$

The rightmost pole on this ladder is $j_p=0$. So the rightmost pole of ζ_α is at $z=-\lambda_\alpha$, with

$$\lambda_\alpha=\min_{p\,:\,k_p>0}\frac{h_p+1}{2k_p},\qquad \mu_\alpha=\#\{p \text{ attaining the minimum}\}.$$

At this point only the zeroth Taylor term ($j_p=0$) of each coordinate attaining the minimum matters, and it contributes $\frac{1}{2k_pz+h_p+1}=\frac{1}{2k_p(z+\lambda_\alpha)}$. Therefore,

$$\zeta_\alpha(z)=\frac{C_\alpha}{(z+\lambda_\alpha)^{\mu_\alpha}}+\big(\text{poles at } -\lambda_\alpha \text{ of lower order}\big)+\big(\text{holomorphic near } -\lambda_\alpha\big),$$

where $C_\alpha>0$ since $b_\alpha\geq 0$ with $b_\alpha(0)>0$.

Step 8: combine the charts. The rightmost pole of ζ is the rightmost over all the finitely many charts:

$$\lambda = \min_{\alpha} \lambda_\alpha = \min_{\alpha} \min_{p\,:\,k_p>0} \frac{h_p+1}{2k_p}, \qquad \mu = \max\{\mu_\alpha \mid \lambda_\alpha = \lambda\}. \tag{31}$$

So λ is the smallest λ_α over all charts and μ is the largest multiplicity among the charts achieving λ. In algebraic geometry, λ is the **real log canonical threshold**. Now $\zeta = \sum_\alpha \zeta_\alpha + \zeta_+$ has no poles to the right of $z = -\lambda$. At $z = -\lambda$ the coefficient of $(z+\lambda)^{-\mu}$ is the sum of the C_α over the charts with $\lambda_\alpha = \lambda$ and $\mu_\alpha = \mu$. Since every C_α is positive these cannot cancel, so the pole at $z = -\lambda$ has order exactly μ and positive leading coefficient. This is the first pole, $\lambda_1 = \lambda$ and $\mu_1 = \mu$.

The remaining poles come from the ladders (30) of the finitely many charts, so they are negative rational numbers, since the exponents h_p, k_p are integers. Given $R > 0$, only finitely many rungs of these ladders lie to the right of $-R$. Adding up the expansions of the ζ_α from Step 6 over the finitely many charts, together with the entire function ζ_+, gives (26). Here H_R collects the holomorphic remainders and the terms whose poles lie at or to the left of $-R$, so it is holomorphic on $\operatorname{Re} z > -R$. □

Remark 2.8. Resolution of singularities is not the only way to prove that ζ has a meromorphic continuation to $\mathbb{C}$ (Bernstein and Gelfand, 1969; Atiyah, 1970). However alternatives do not give a practical recipe for finding the location and order of the poles. The advantage of the resolution is that the poles and their orders can be read off from the normal crossing exponents k_p and h_p. In principle these exponents can be computed by a sequence of blow-ups, but carrying this out for the loss function of a neural network is hard and has only been done for a few architectures.

2.5 Inverting the Mellin transform

Now that we have expanded the zeta function in terms of its poles, the next step is to apply the inverse Mellin transform.

Exercise 2.1 (An inverse Mellin transform)**.** Let $\alpha > 0$ and let m be a positive integer. Define

$$f_{\alpha,m}(t) = \begin{cases} \dfrac{1}{(m-1)!}\, t^{\alpha-1}(-\log t)^{m-1} & 0 < t < 1, \\ 0 & t \geq 1. \end{cases}$$

(a) Show that $\mathcal{M}[f_{\alpha,m}](z) = \dfrac{1}{(z+\alpha)^m}$ for $\operatorname{Re} z > -\alpha$.

[*Hint:* Substitute $t = e^{-s}$, and use $\int_0^\infty s^{m-1}e^{-as}\,\mathrm{d}s = (m-1)!/a^m$ for $\operatorname{Re} a > 0$.]

(b) Show that the integral defining $\mathcal{M}[f_{\alpha,m}](z)$ diverges for real $z \leq -\alpha$, so that the pole sits exactly where the Mellin integral stops converging.

This gives a dictionary between the pole structure of ζ and the behaviour of v near $t = 0$:

$\zeta(z)$	$v(t)$ as $t \to 0$
pole at $z = -\alpha$	$t^{\alpha-1}$
order m	$(-\log t)^{m-1}$
poles further left	smaller as $t \to 0$

By the linearity of the Mellin transform we invert (26) term by term using Exercise 2.1: each term $\frac{1}{(z+\lambda_i)^m}$ becomes $\frac{1}{(m-1)!}\, t^{\lambda_i - 1}(-\log t)^{m-1}$, giving

$$v(t) = \sum_{i=1}^{\infty} \sum_{m=1}^{\mu_i} \frac{c_{i,m}}{(m-1)!}\, t^{\lambda_i - 1}(-\log t)^{m-1} \tag{32}$$

for small t. Strictly speaking, inverting an infinite sum term by term needs justification, and (32) should be understood as an asymptotic expansion as $t \to 0$ rather than a convergent series. We take this on trust (see Watanabe, 2009, Section 4.5).

The term with the smallest power of t dominates the assymptotics and between terms with the same power, the largest log power dominates. Therefore,

$$v(t) = c' t^{\lambda - 1}(-\log t)^{\mu - 1} + (\text{lower order terms}) \qquad \text{as } t \to 0 \tag{33}$$

for some constant $c' > 0$.

2.6 From the density of states to the volume

It remains to integrate. By Proposition 2.2(b) and (33),

$$\begin{aligned} V(\epsilon) = \int_0^{\epsilon} v(t)\, \mathrm{d}t &= c' \int_0^{\epsilon} t^{\lambda-1}(-\log t)^{\mu-1}\, \mathrm{d}t + (\text{lower order terms}) \\ &= \frac{c'}{\lambda}\, \epsilon^{\lambda}(-\log \epsilon)^{\mu-1} + (\text{lower order terms}), \end{aligned} \tag{34}$$

where the last step is Exercise 2.2. The same computation shows that lower order terms in v integrate to lower order terms in V. This is the asymptotic form (4) of Theorem 1.1, with $c = c'/\lambda > 0$, which completes the sketch.

Exercise 2.2 (From the density of states to the volume)**.** Let $\lambda > 0$ and let μ be a positive integer. For $0 < \epsilon < 1$, let $I_\mu(\epsilon) = \int_0^{\epsilon} t^{\lambda-1}(-\log t)^{\mu-1}\, \mathrm{d}t$.

(a) Integrate by parts to show that, for $\mu \geq 2$,

$$I_\mu(\epsilon) = \frac{1}{\lambda}\, \epsilon^{\lambda}(-\log \epsilon)^{\mu-1} + \frac{\mu - 1}{\lambda}\, I_{\mu-1}(\epsilon).$$

(b) Deduce by induction that, for $\mu \geq 2$, $I_\mu(\epsilon) = \frac{1}{\lambda}\, \epsilon^{\lambda}(-\log \epsilon)^{\mu-1} + O\big(\epsilon^{\lambda}(-\log \epsilon)^{\mu-2}\big)$ as $\epsilon \to 0$.

Further reading. Watanabe (2009, Section 1.4 and Chapter 4) gives the full argument using Schwartz distributions.

Exercise 2.3 (Zeta functions of monomials)**.**

(a) Let $K(w) = w_1^{2k_1} \cdots w_d^{2k_d}$ on $B = [-1,1]^d$, where the $k_j \geq 0$ are integers, not all zero. Compute $\zeta(z)$ and deduce that

$$\lambda = \min_{j\,:\,k_j>0} \frac{1}{2k_j}, \qquad \mu = \#\{j \mid k_j = \max_i k_i\}.$$

(b) Now let $K(x,y) = x^2y^2$ on $[-1,1]^2$. Find λ and μ, and compare with $K(x,y) = x^2$ on the same square.

(c) In part (b), ζ consists of a single pole term, so the inverse Mellin transform can be computed exactly. Use Exercise 2.1 to find $v(t)$, and then integrate to find $V(\epsilon)$ for $0 < \epsilon < 1$. Check that the result has the form (4).

(d) Verify your answer to (c) by computing $V(\epsilon)$ directly.

[*Hint:* By symmetry, $V(\epsilon) = 4\,\mathrm{Vol}\{(x,y) \in [0,1]^2 \mid xy < \sqrt{\epsilon}\}$.]

References and further reading

Miki Aoyagi and Sumio Watanabe. Stochastic complexities of reduced rank regression in Bayesian estimation. *Neural Networks*, 18(7):924–933, 2005. Access via Crossref. Cited on page 6.

Michael F. Atiyah. Resolution of singularities and division of distributions. *Communications on Pure and Applied Mathematics*, 23(2):145–150, 1970. Access via Crossref. Cited on page 15.

Joseph N. Bernstein and Sergei I. Gelfand. Meromorphic property of the functions P^λ. *Functional Analysis and Its Applications*, 3(1):68–69, 1969. Access via Crossref. Cited on page 15.

Heisuke Hironaka. Resolution of singularities of an algebraic variety over a field of characteristic zero: I, II. *Annals of Mathematics*, 79(1–2):109–203, 205–326, 1964. Access via Crossref. Cited on page 11.

Edmund Lau, Zach Furman, George Wang, Daniel Murfet, and Susan Wei. The local learning coefficient: A singularity-aware complexity measure. In *The 28th International Conference on Artificial Intelligence and Statistics*. 2025. Access via OpenReview. Cited on pages 3 and 5.

Simon Pepin Lehalleur, Jesse Hoogland, Matthew Farrugia-Roberts, Susan Wei, Alexander Gietelink Oldenziel, George Wang, Liam Carroll, and Daniel Murfet. You are what you eat–AI alignment requires understanding how data shapes structure and generalisation. 2025. Preprint arXiv:2502.05475 [cs.LG].

Sumio Watanabe. *Algebraic Geometry and Statistical Learning Theory*. Cambridge University Press, 2009. Cited on pages 3, 8, 10, 11, and 16.

Susan Wei, Daniel Murfet, Mingming Gong, Hui Li, Jesse Gell-Redman, and Thomas Quella. Deep learning is singular, and that's good. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12):10473–10486, 2023.

A Solutions

Solution to Exercise 1.1 (Volume scaling in the regular case) 20
Solution to Exercise 1.2 (Quadratic Valley Volume Scaling) 20
Solution to Exercise 1.3 (Two equations for λ) 21
Solution to Exercise 1.4 (Calculating the LLC) 22
Solution to Exercise 1.5 (Upper bound on the LLC) 22
Solution to Exercise 1.6 (LLC vs Hessian Rank) 23
Solution to Exercise 1.7 (The cubically-parameterised loss) 23
Solution to Exercise 1.8 (Hölder exponent and learning coefficient) 25
Solution to Exercise 1.9 (Learning coefficients of two-layer DLNs) 25

Solution to Exercise 2.1 (An inverse Mellin transform) 27
Solution to Exercise 2.2 (From the density of states to the volume) 27
Solution to Exercise 2.3 (Zeta functions of monomials) 27

A.1 Solutions to exercises from Section 1

Solution to Exercise 1.1 (Volume scaling in the regular case)**.** Because w^* is a local minimum, the gradient $\nabla L(w^*)$ is zero. By Taylor's theorem, we can write the loss function near w^* exactly as:

$$L(w) - L(w^*) = \frac{1}{2}(w - w^*)^\top H(w - w^*) + o(\|w - w^*\|^2)$$

where $H = \nabla^2 L(w^*)$ is the strictly positive definite Hessian matrix.

The sublevel set is defined by the condition $L(w) - L(w^*) < \epsilon$. As $\epsilon \to 0$, the neighborhood of parameters that satisfy this condition shrinks toward w^*. In this limit, the higher-order error term $o(\|w - w^*\|^2)$ becomes negligible, and the boundary of the sublevel set is governed entirely by the quadratic form:

$$\frac{1}{2}(w - w^*)^\top H(w - w^*) < \epsilon.$$

We can rewrite this inequality into the standard form of an ellipsoid, $x^\top A x < 1$, by dividing both sides by ϵ:

$$(w - w^*)^\top \left(\frac{1}{2\epsilon} H\right)(w - w^*) < 1.$$

Using the hint, the volume of a d-dimensional ellipsoid defined by a matrix A is proportional to $(\det A)^{-1/2}$. Here, our matrix is $A = \frac{1}{2\epsilon} H$.

Therefore,

$$\det\left(\frac{1}{2\epsilon} H\right)^{-1/2} = \left(\left(\frac{1}{2\epsilon}\right)^d \det H\right)^{-1/2} = \epsilon^{d/2} \underbrace{\left(\frac{\det H}{2^d}\right)^{-1/2}}_{:=c}$$

Because the true loss function deviates from this pure quadratic form only by an error of $o(\|w - w^*\|^2)$, the volume of the true sublevel set deviates from the ellipsoid's volume only by a correspondingly higher-order error term. Thus, we conclude:

$$V(\epsilon) = c\epsilon^{d/2} + o(\epsilon^{d/2}) \quad \text{as } \epsilon \to 0.$$

□

Solution to Exercise 1.2 (Quadratic Valley Volume Scaling)**.**

(a) The loss is $L(x_1, x_2) = x_1^2$, which is a parabolic trough: a parabola in the x_1 direction and constant along x_2. The zero set is the line $x_1 = 0$.

(b) $\mathcal{W}_0 = \{0\}^{d'} \times [-R, R]^{d-d'}$, which is a $(d - d')$-dimensional affine subspace of $\mathcal{W}$. Hence $\dim(\mathcal{W}_0) = d - d'$.

(c) Since $L(w) = \sum_{i=1}^{d'} x_i^2$, the partial derivatives are

$$\frac{\partial L}{\partial x_i} = \begin{cases} 2x_i & i \le d', \\ 0 & i > d', \end{cases} \qquad \frac{\partial^2 L}{\partial x_i \partial x_j} = \begin{cases} 2 & i = j \le d', \\ 0 & \text{otherwise.} \end{cases}$$

So the Hessian is the block diagonal matrix:

$$\nabla^2 L = \begin{pmatrix} 2I_{d'} & 0 \\ 0 & 0_{d-d'} \end{pmatrix},$$

which has rank d'. In particular it is positive semi-definite but not positive definite whenever $d' < d$.

(d) The minimum of L is 0, achieved on $\mathcal{W}_0$. For any $w^* \in \mathcal{W}_0$ and $\epsilon > 0$, the sublevel set is

$$B(w^*, \epsilon) = \left\{ (x_1, \ldots, x_d) \in [-R, R]^d \;\middle|\; \sum_{i=1}^{d'} x_i^2 < \epsilon \right\} = \left\{ x \in \mathbb{R}^{d'} : \sum_{i=1}^{d'} x_i^2 < \epsilon \right\} \times [-R, R]^{d-d'}$$

for R sufficiently large (specifically $\sqrt{\epsilon} < R$). The condition $\sum_{i=1}^{d'} x_i^2 < \epsilon$ can be rewritten as $\mathbf{x}^\top A \mathbf{x} < 1$ where $A = \epsilon^{-1} I_{d'}$. By the ellipsoid volume formula from Exercise 1.1, the volume of this region is proportional to $(\det A)^{-1/2} = (\epsilon^{-d'})^{-1/2} = \epsilon^{d'/2}$. Multiplying by the volume $(2R)^{d-d'}$ of the remaining coordinates gives $V(\epsilon) \propto \epsilon^{d'/2}$. □

Solution to Exercise 1.3 (Two equations for λ)**.**

(a) From Theorem 1.1, as $\epsilon \to 0$,

$$V(\epsilon) = c\,\epsilon^{\lambda} (-\log \epsilon)^{m-1} \big(1 + o(1)\big).$$

Substituting $\epsilon \mapsto a\epsilon$,

$$V(a\epsilon) = c\,(a\epsilon)^{\lambda} (-\log(a\epsilon))^{m-1} \big(1 + o(1)\big).$$

Taking the ratio,

$$\frac{V(a\epsilon)}{V(\epsilon)} = a^{\lambda} \left(\frac{-\log(a\epsilon)}{-\log \epsilon} \right)^{m-1} \big(1 + o(1)\big).$$

As $\epsilon \to 0$, $-\log \epsilon \to \infty$, and

$$\frac{-\log(a\epsilon)}{-\log \epsilon} = \frac{-\log a - \log \epsilon}{-\log \epsilon} = 1 + \frac{\log a}{-\log \epsilon} \xrightarrow{\epsilon \to 0} 1.$$

Therefore $V(a\epsilon)/V(\epsilon) \to a^{\lambda}$. Taking logarithms and dividing by $\log a > 0$ gives the result.

(b) From Theorem 1.1, $V(\epsilon) = c\,\epsilon^{\lambda(w^*)} (-\log \epsilon)^{\mu(w^*)-1} (1 + o(1))$, so

$$\log V(\epsilon) = \lambda(w^*) \log \epsilon + (\mu(w^*) - 1) \log(-\log \epsilon) + \log c + o(1).$$

Dividing by $\log \epsilon$

$$\frac{\log V(\epsilon)}{\log \epsilon} = \lambda(w^*) + \frac{(\mu(w^*) - 1) \log(-\log \epsilon)}{\log \epsilon} + \frac{\log c + o(1)}{\log \epsilon}.$$

It remains to prove that the final two terms in this sum vanish as $\epsilon \to 0$. The final term tends to 0 since $\log \epsilon \to -\infty$ and the numerator is bounded as $\epsilon \to 0$. Now we look at the middle term which is proportional to

$$\lim_{\epsilon \to 0} \frac{\log(-\log \epsilon)}{\log \epsilon} = \lim_{x \to \infty} \frac{\log(x)}{-x} = 0,$$

completing the proof. □

Solution to Exercise 1.4 (Calculating the LLC)**.**

(a) $V(\epsilon) = \mathrm{Vol}(\{x : x^2 < \epsilon\}) = 2\sqrt{\epsilon}$, so $V(\epsilon) \propto \epsilon^{1/2}$ and $\lambda = 1/2$. The Hessian is $L''(0) = 2$, which has rank 1 (full rank).

(b) $V(\epsilon) = \mathrm{Vol}(\{x : x^{2m} < \epsilon\}) = 2\epsilon^{1/2m}$, so $\lambda = 1/2m$. For $m \geq 2$, the first $2m-1$ derivatives vanish at $x = 0$, so the Hessian is $L''(0) = 0$, which has rank 0. Since $\lambda = 1/2m < 1/2 = d/2$, the singularity makes the model effectively simpler than its parameter count suggests.

(c) The sublevel set is $\{(x, y) : x^4 + y^4 < \epsilon\}$. Substituting $x = \epsilon^{1/4}s$, $y = \epsilon^{1/4}t$:

$$V(\epsilon) = \epsilon^{1/2} \iint_{\{s^4+t^4<1\}} ds\, dt = c\,\epsilon^{1/2},$$

so $\lambda = 1/2$. The Hessian at the origin is $\nabla^2 L(0,0) = \mathrm{diag}(12x^2, 12y^2)\big|_{(0,0)} = 0$, which has rank 0.

(d) The sublevel set is $\{(x, y) : x^2 + y^4 < \epsilon\}$. For each fixed y with $|y| < \epsilon^{1/4}$, x ranges over $|x| < \sqrt{\epsilon - y^4}$. Substituting $y = \epsilon^{1/4}t$:

$$V(\epsilon) = \int_{-\epsilon^{1/4}}^{\epsilon^{1/4}} 2\sqrt{\epsilon - y^4}\, dy = 2\epsilon^{1/4} \int_{-1}^{1} \sqrt{\epsilon - \epsilon t^4}\, dt = 2\epsilon^{3/4} \int_{-1}^{1} \sqrt{1 - t^4}\, dt = c\,\epsilon^{3/4},$$

so $\lambda = 3/4$. The Hessian is $\nabla^2 L(0,0) = \mathrm{diag}(2, 12y^2)\big|_{(0,0)} = \mathrm{diag}(2, 0)$, which has rank 1.

□

Solution to Exercise 1.5 (Upper bound on the LLC)**.**

(a) Since w^* is a local minimum of L, we have $\nabla L(w^*) = 0$. By Taylor's theorem with Lagrange remainder, for any $w \in B(w^*)$ there exists $\tilde{w}$ on the line segment between w^* and w such that

$$\begin{aligned} L(w) - L(w^*) &= \nabla L(w^*)^\top (w - w^*) + \frac{1}{2}(\tilde{w} - w^*)^\top \nabla^2 L(\tilde{w})(\tilde{w} - w^*) \\ &= \frac{1}{2}(\tilde{w} - w^*)^\top \nabla^2 L(\tilde{w})(\tilde{w} - w^*). \end{aligned}$$

Since $\tilde{w} \in B(w^*)$, the eigenvalues of $\nabla^2 L(\tilde{w})$ are bounded above by Λ, so

$$(\tilde{w} - w^*)^\top \nabla^2 L(\tilde{w})(\tilde{w} - w^*) \leq \Lambda \|\tilde{w} - w^*\|^2 \leq \Lambda \|w - w^*\|^2.$$

Combining, $L(w) - L(w^*) \leq \frac{1}{2}\Lambda \|w - w^*\|^2$.

(b) Set $r = \sqrt{2\epsilon/\Lambda}$. Set ϵ sufficiently small that $B_r(w^*) \subseteq B(w^*)$. If $w \in B_r(w^*)$, then $\|w - w^*\| < r$, so by part (a),

$$L(w) - L(w^*) \leq \frac{1}{2}\Lambda \|w - w^*\|^2 < \frac{1}{2}\Lambda \cdot \frac{2\epsilon}{\Lambda} = \epsilon.$$

Hence $w \in B(w^*, \epsilon)$, and therefore $B_r(w^*) \subseteq B(w^*, \epsilon)$.

(c) Since $B_r(w^*) \subseteq B(w^*, \epsilon)$,

$$V(\epsilon) = \mathrm{Vol}(B(w^*, \epsilon)) \geq \mathrm{Vol}(B_r(w^*)) = c\, r^d = c \left(\frac{2\epsilon}{\Lambda}\right)^{d/2} = C\, \epsilon^{d/2},$$

for some $c > 0$ and $C = c(2/\Lambda)^{d/2} > 0$. This holds for all sufficiently small $\epsilon > 0$ (i.e. small enough that $B_r(w^*) \subseteq B(w^*)$).

(d) From part (c), $V(\epsilon) \geq C\epsilon^{d/2}$ for sufficiently small ϵ (w.l.o.g. assume $\epsilon < 1$). Taking logarithms,

$$\log V(\epsilon) \geq \log C + \frac{d}{2} \log \epsilon.$$

Dividing by $\log \epsilon < 0$ (since $\epsilon < 1$) reverses the inequality:

$$\frac{\log V(\epsilon)}{\log \epsilon} \leq \frac{\log C}{\log \epsilon} + \frac{d}{2}.$$

Since $\log C / \log \epsilon \to 0$ as $\epsilon \to 0^+$, taking the limit and using Equation (6) gives

$$\lambda(w^*) = \lim_{\epsilon \to 0^+} \frac{\log V(\epsilon)}{\log \epsilon} \leq \frac{d}{2}. \qquad \square$$

Solution to Exercise 1.6 (LLC vs Hessian Rank)**.**

Let $L_1(x, y) = x^2 + y^4$ with $w_1 = (0, 0)$. Then $\nabla^2 L_1(0, 0) = \mathrm{diag}(2, 0)$, so $\operatorname{rank} \nabla^2 L_1(0, 0) = 1$. By Exercise 1.4, $\lambda(w_1) = \frac{3}{4}$.

Let $L_2(x, y) = x^2$ with $w_2 = (0, 0)$. Then $\nabla^2 L_2(0, 0) = \mathrm{diag}(2, 0)$, so $\operatorname{rank} \nabla^2 L_2(0, 0) = 1$. The minimiser locus is $\{x = 0\}$, which is the setting of Exercise 1.2 with $d' = 1$, giving $\lambda(w_2) = \frac{1}{2}$.

Both Hessians have rank 1, but $\lambda(w_1) = \frac{3}{4} \neq \frac{1}{2} = \lambda(w_2)$. The difference is that the flat direction in L_2 is *exactly* flat (the minimisers form a line), while in L_1 it is only flat to second order but quartic beyond that. The LLC detects this distinction; the Hessian rank does not. $\square$

Solution to Exercise 1.7 (The cubically-parameterised loss)**.**

(a) The cubically-parameterised loss $L(\mu) = \frac{1}{2}(\mu^3 - \mu_0^3)^2$ vanishes if and only if $\mu^3 = \mu_0^3$, i.e. $\mu = \mu_0$ (since $t \mapsto t^3$ is a bijection on $\mathbb{R}$). The ordinary quadratic loss $\ell(\theta) = \frac{1}{2}(\theta - \theta_0)^2$ vanishes if and only if $\theta = \theta_0$. Since for every $\theta_0 \in \mathbb{R}$ we can write $\theta_0 = \mu_0^3$ for a unique $\mu_0 = \theta_0^{1/3}$, both losses achieve their minimum value of zero at exactly the same set of "targets" $\theta_0 \in \mathbb{R}$, so the set of achievable minima is the same.

(b) Computing derivatives:

$$L'(\mu) = 3\mu^2(\mu^3 - \mu_0^3),$$
$$L''(\mu) = 6\mu(\mu^3 - \mu_0^3) + 9\mu^4.$$

At the minimum $\mu = \mu_0$: $L''(\mu_0) = 6\mu_0 \cdot 0 + 9\mu_0^4 = 9\mu_0^4$.

- If $\mu_0 = 0$: $L''(0) = 0$, so the Hessian vanishes and the loss is degenerate at its minimum.

- If $\mu_0 \neq 0$: $L''(\mu_0) = 9\mu_0^4 > 0$, so the loss is non-degenerate.

(c) We compute $V(\epsilon) = \mathrm{vol}(\{\mu \in \mathbb{R} : L(\mu) < \epsilon\})$ in each case.

Case $\mu_0 = 0$. When $\mu_0 = 0$ the loss simplifies to $L(\mu) = \frac{1}{2}\mu^6$, so

$$V(\epsilon) = \int_{\frac{1}{2}\mu^6 < \epsilon} d\mu = \int_{|\mu| < (2\epsilon)^{1/6}} d\mu = 2(2\epsilon)^{1/6} = 2^{7/6}\,\epsilon^{1/6}.$$

Case $\mu_0 \neq 0$. The sublevel set condition $L(\mu) < \epsilon$ is $\frac{1}{2}(\mu^3 - \mu_0^3)^2 < \epsilon$, i.e.

$$\mu_0^3 - \sqrt{2\epsilon} \;<\; \mu^3 \;<\; \mu_0^3 + \sqrt{2\epsilon}.$$

Since the cube root is monotone on $\mathbb{R}$, this is equivalent to

$$\left(\mu_0^3 - \sqrt{2\epsilon}\right)^{1/3} \;<\; \mu \;<\; \left(\mu_0^3 + \sqrt{2\epsilon}\right)^{1/3},$$

and therefore

$$V(\epsilon) = \left(\mu_0^3 + \sqrt{2\epsilon}\right)^{1/3} - \left(\mu_0^3 - \sqrt{2\epsilon}\right)^{1/3}.$$

(d) **Case** $\mu_0 = 0$**:** $V(\epsilon) \propto \epsilon^{1/6}$, so $\lambda = \frac{1}{6}$.

Case $\mu_0 \neq 0$**:** For small ϵ we expand by factoring out μ_0:

$$V(\epsilon) = \mu_0 \left[\left(1 + \frac{\sqrt{2\epsilon}}{\mu_0^3}\right)^{1/3} - \left(1 - \frac{\sqrt{2\epsilon}}{\mu_0^3}\right)^{1/3} \right].$$

Applying the binomial expansion $(1+x)^{1/3} = 1 + \frac{1}{3}x + o(x)$ to each term with $x = \pm\frac{\sqrt{2\epsilon}}{\mu_0^3}$, the constant terms cancel and the linear terms add:

$$V(\epsilon) = \mu_0 \left[\frac{1}{3}\frac{\sqrt{2\epsilon}}{\mu_0^3} + \frac{1}{3}\frac{\sqrt{2\epsilon}}{\mu_0^3} + o(\epsilon^{1/2}) \right] = \frac{2\sqrt{2}}{3\mu_0^2}\,\epsilon^{1/2} + o(\epsilon^{1/2}).$$

In the non-degenerate case $\mu_0 \neq 0$, we recover the regular value $\lambda = d/2 = 1/2$ as expected. In the degenerate case $\mu_0 = 0$, $\lambda = 1/6 < 1/2$: the singularity at $\mu = 0$ (where the loss vanishes to sixth order rather than second order) makes the model effectively simpler than its parameter count suggests.

(e) From part (c), for fixed $\epsilon > 0$ we plot

$$V(\epsilon;\, \mu_0) = \left(\mu_0^3 + \sqrt{2\epsilon}\right)^{1/3} - \left(\mu_0^3 - \sqrt{2\epsilon}\right)^{1/3}$$

The function is a smooth, symmetric bump centred at $\mu_0 = 0$ with no discontinuity. Making ϵ smaller narrows the bump and lowers the peak; making ϵ larger widens it. Despite the learning coefficient jumping discontinuously from $1/2$ to $1/6$ at $\mu_0 = 0$ exactly, the finite-ϵ volume transitions smoothly, with the singularity's influence confined to a neighbourhood that shrinks as $\epsilon \to 0$.

□

Solution to Exercise 1.8 (Hölder exponent and learning coefficient)**.** Under the loss pseudo-metric $\rho(w,u) = |L(w) - L(u)|$, the ϵ-ball centred at w^* (within $B(w^*)$) is

$$B_\epsilon(w^*) = \{w \in B(w^*) : |L(w) - L(w^*)| < \epsilon\}.$$

Since w^* is a local minimum and $L(w) \geq L(w^*)$ on $B(w^*)$, this simplifies to $B_\epsilon(w^*) = \{w \in B(w^*) : L(w) - L(w^*) < \epsilon\} = B(w^*, \epsilon)$, the sublevel set from (1). In particular, $\nu(B_\epsilon(w^*)) = V(\epsilon)$ where ν is Lebesgue measure.

The Hölder exponent is therefore

$$\alpha(w^*) = \lim_{\epsilon \to 0} \frac{\log \nu(B_\epsilon(w^*))}{\log \epsilon} = \lim_{\epsilon \to 0} \frac{\log V(\epsilon)}{\log \epsilon} = \lambda$$

by Equation (6)

□

Solution to Exercise 1.9 (Learning coefficients of two-layer DLNs)**.**

(a) We first check which case of Theorem 1.5 applies. Setting $M = N = H = m$, the three conditions of Case 1 each reduce to $m+r \leq 2m$, i.e. $r \leq m$, which always holds. So we are always in Case 1. The parity condition is on $M + H + N + r = 3m + r$, which has the same parity as $m + r$.

Case 1(a): $m + r$ even. Setting $M = N = H = m$ in the formula:

$$\begin{aligned}
\lambda &= \frac{-(m+r)^2 - m^2 - m^2 + 2(m+r)m + 2(m+r)m + 2m^2}{8} \\
&= \frac{-(m+r)^2 + 4m(m+r)}{8} \\
&= \frac{(m+r)\big(4m - (m+r)\big)}{8} \\
&= \frac{(m+r)(3m-r)}{8},
\end{aligned}$$

with $\mu = 1$.

Case 1(b): $m + r$ odd. The same calculation gives

$$\lambda = \frac{(m+r)(3m-r)+1}{8},$$

with $\mu = 2$.

(b) Differentiating with respect to r:

$$\frac{d\lambda}{dr} = \frac{1}{8}\frac{d}{dr}[(m+r)(3m-r)] = \frac{1}{8}\left[(3m-r) - (m+r)\right] = \frac{1}{4}(m-r) \geq 0,$$

so λ is increasing in r therefore **lower rank** $\implies$ **lower** λ. This is consistent with Exercise 4.9(d) of Lecture 1 where we saw that at lower-rank parameters, more directions in parameter space are degenerate.

(c) The parameter space is $\mathcal{W} = \mathbb{R}^{2m^2}$, so the regular-model baseline is $d/2 = m^2$. At full rank $r = m$:

$$\lambda = \frac{(2m)(2m)}{8} = \frac{m^2}{2} = \frac{d}{4},$$

which is already half the regular-model value. At rank $r = 0$:

$$\lambda = \frac{m \cdot 3m}{8} = \frac{3m^2}{8} = \frac{3d}{16},$$

even smaller. In fact $\lambda < d/2$ for every rank $0 \leq r \leq m$.

This tells us that the two-layer DLN is singular at every point in its loss landscape, even at full-rank parameters.

□

A.2 Solutions to exercises from Section 2

Solution to Exercise 2.1 (An inverse Mellin transform)**.**

(a) By definition,

$$\mathcal{M}[f_{\alpha,m}](z) = \frac{1}{(m-1)!}\int_0^1 t^{z+\alpha-1}(-\log t)^{m-1}\,\mathrm{d}t.$$

Substitute $t = e^{-s}$, so that $-\log t = s$, $\mathrm{d}t = -e^{-s}\,\mathrm{d}s$, and the limits $t = 0, 1$ become $s = \infty, 0$. Then $t^{z+\alpha-1}\,\mathrm{d}t = -e^{-(z+\alpha)s}\,\mathrm{d}s$ and

$$\mathcal{M}[f_{\alpha,m}](z) = \frac{1}{(m-1)!}\int_0^\infty s^{m-1}e^{-(z+\alpha)s}\,\mathrm{d}s = \frac{1}{(m-1)!}\cdot\frac{(m-1)!}{(z+\alpha)^m} = \frac{1}{(z+\alpha)^m},$$

using the hint with $a = z+\alpha$, which has $\operatorname{Re} a > 0$.

(b) For real $z \le -\alpha$, set $a = z+\alpha \le 0$. Then $e^{-as} \ge 1$ for $s \ge 0$, so the integrand $s^{m-1}e^{-as} \ge s^{m-1}$, which is not integrable on $(0,\infty)$. Hence the Mellin integral diverges. □

Solution to Exercise 2.2 (From the density of states to the volume)**.**

(a) Integrate by parts, differentiating $(-\log t)^{\mu-1}$ and integrating $t^{\lambda-1}$:

$$\begin{aligned} I_\mu(\epsilon) &= \Big[\frac{t^\lambda}{\lambda}(-\log t)^{\mu-1}\Big]_0^\epsilon - \int_0^\epsilon \frac{t^\lambda}{\lambda}\,(\mu-1)(-\log t)^{\mu-2}\Big(-\frac{1}{t}\Big)\,\mathrm{d}t \\ &= \frac{1}{\lambda}\,\epsilon^\lambda(-\log\epsilon)^{\mu-1} + \frac{\mu-1}{\lambda}\,I_{\mu-1}(\epsilon). \end{aligned}$$

The boundary term at $t = 0$ vanishes because $t^\lambda(-\log t)^{\mu-1} \to 0$ as $t \to 0$ when $\lambda > 0$.

(b) The base case is $I_1(\epsilon) = \epsilon^\lambda/\lambda$. For $\mu \ge 2$, assume that $I_{\mu-1}(\epsilon) = O\big(\epsilon^\lambda(-\log\epsilon)^{\mu-2}\big)$. This holds for $\mu = 2$ by the base case, and for larger μ by the inductive hypothesis, since its leading term is of this order. Part (a) then gives

$$I_\mu(\epsilon) = \frac{1}{\lambda}\,\epsilon^\lambda(-\log\epsilon)^{\mu-1} + O\big(\epsilon^\lambda(-\log\epsilon)^{\mu-2}\big).$$

Since $(-\log\epsilon)^{-1} \to 0$, the error term is $o\big(\epsilon^\lambda(-\log\epsilon)^{\mu-1}\big)$, so the first term dominates. □

Solution to Exercise 2.3 (Zeta functions of monomials)**.**

(a) For $\operatorname{Re} z > 0$ the integral factorises:

$$\zeta(z) = \prod_{j=1}^d \int_{-1}^1 |w_j|^{2k_j z}\,\mathrm{d}w_j = \prod_{j=1}^d \frac{2}{2k_j z + 1}.$$

The right-hand side is a meromorphic function on $\mathbb{C}$. Each factor with $k_j = 0$ equals the constant 2, and each factor with $k_j > 0$ has a simple pole at $z = -\frac{1}{2k_j}$. The rightmost pole is therefore at $-\lambda$ with $\lambda = \min_{k_j>0}\frac{1}{2k_j}$. Its order μ is the number of factors with a pole there, i.e. the number of j with $k_j = \max_i k_i$. This agrees with (31): K is already in the normal crossing form (27), so the identity map is a resolution, with a single chart, $b = 1$ and all $h_j = 0$.

(b) Here $k=(1,1)$, so $\zeta(z)=\frac{4}{(2z+1)^2}=\frac{1}{(z+\frac12)^2}$, giving $\lambda=\frac12$ and $\mu=2$. For $K(x,y)=x^2$ on $[-1,1]^2$ we have $k=(1,0)$ and $\zeta(z)=\frac{2\cdot2}{2z+1}=\frac{2}{z+\frac12}$, giving $\lambda=\frac12$ and $\mu=1$. The two losses have the same LLC but different multiplicities. The zero set of x^2y^2 consists of two lines crossing at the origin, and the crossing produces an extra logarithm.

(c) By Exercise 2.1 with $\alpha=\frac12$ and $m=2$, the function $t^{-1/2}(-\log t)$ on $(0,1)$ has Mellin transform $\frac{1}{(z+\frac12)^2}=\zeta(z)$. Since $K\le1$ on the square, v vanishes for $t>1$, and so

$$v(t)=t^{-1/2}(-\log t),\qquad 0<t<1.$$

Integrating (check by differentiating the right-hand side),

$$V(\epsilon)=\int_0^\epsilon t^{-1/2}(-\log t)\,\mathrm{d}t=2\sqrt{\epsilon}\,(-\log\epsilon)+4\sqrt{\epsilon}.$$

The first term dominates as $\epsilon\to0$, so $V(\epsilon)=2\,\epsilon^{1/2}(-\log\epsilon)^{2-1}+o\big(\epsilon^{1/2}(-\log\epsilon)\big)$, which is (4) with $c=2$, $\lambda=\frac12$ and $\mu=2$.

(d) For fixed $x\in[0,1]$, the condition $xy<\sqrt{\epsilon}$ with $y\in[0,1]$ means $y<\min(1,\sqrt{\epsilon}/x)$. Hence

$$\begin{aligned}V(\epsilon)=4\int_0^1\min\Big(1,\frac{\sqrt{\epsilon}}{x}\Big)\,\mathrm{d}x&=4\Big[\sqrt{\epsilon}+\int_{\sqrt{\epsilon}}^1\frac{\sqrt{\epsilon}}{x}\,\mathrm{d}x\Big]\\&=4\sqrt{\epsilon}+4\sqrt{\epsilon}\,\log\frac{1}{\sqrt{\epsilon}}=4\sqrt{\epsilon}+2\sqrt{\epsilon}\,(-\log\epsilon),\end{aligned}$$

in exact agreement with (c). □