Singular learning theory IV : local RLCT

Throughout [W] means Watanabe's book. We work in the setting of [W, §6.1] and assume Fundamental Conditions (I), (II) with $s=2$. (a somewhat informal, but more readable, account is given in [W, §1.4]). In particular $q(x)$ is realisable, and W is as in [W, Def^n 6.3]. We aim to clarify (a) the role of local RLCTs (to be defined) and (b) the role of the prior in RLCTs.

Setup Recall the partition of parameter space of [W, Theorem 6.5]. For some $\epsilon > 0$ we replace W by

$$W_\epsilon = \{ w \in W \mid K(w) \leq \epsilon \}$$

which is legitimate since we are only concerned with the RLCT, in this note. Then $\mathcal{M} := g^{-1}(W_\epsilon)$ is covered by a finite set

$$\mathcal{M} = \cup_\alpha M_\alpha$$

where the M_α are a very particular kind of set (not open) given in Remark 2.14, and constructed as follows: for each $p \in \mathcal{M}$ choose an open subset of a coordinate chart centered on p of the form

$$O_p(b) := (-b,b)^d = \{ u = (u_1, \ldots, u_d) \mid |u_i| < b \quad 1 \leq i \leq d \} \qquad (1.1)$$

where the dimension of W (hence also $\mathcal{M}$) is d. The construction of Remark 2.14 covers $\mathcal{M}$ by $\{ O_p(b) \}_{p \in \mathcal{M}}$ uses compactness to choose a finite subcover. For each p chosen, it then writes (using again local coordinates)

$$O_p(b) = \left[(-b,0] \cup [0,b) \right]^d$$
$$= (-b,0] \times (-b,0] \times \cdots \times (-b,0]$$
$$\cup (-b,0] \times \cdots \times [0,b) \cup \cdots$$
$$= \bigcup_{i=1}^{2^d} M_i \quad \text{where } M_i \cong [0,b)^d.$$

As p ranges over the finite list of chosen points, and i over indices for these sequences of "positive/negative" choices for local coordinates u_i around p, we enumerate some finite set of subsets $M_\alpha \subseteq M$ such that $M = \cup_\alpha M_\alpha$. There is an associated family of functions $\beta_\alpha(u)$ with $\operatorname{supp} \beta_\alpha \subseteq M_\alpha$ ($\beta_\alpha(u)$ being C^ω by the usual partition of unity tricks using the prescribed $\beta_\alpha^{(0)}$ rather than $\equiv 1$ check) such that for any integrable H,

$$\int_W H(w)\,dw = \int_M H(g(u))\,|g'(u)|\,du \tag{2.1}$$
$$= \sum_\alpha \int_{M_\alpha} H(g(u))\,\beta_\alpha(u)\,|g'(u)|\,du$$

In [W, Theorem 6.5], it continues to say that in each M_α (keeping in mind $0 \le u_i \le b$) there is a C^∞ function $\phi(u)$ such that

$$K(g(u)) = u^{2k} = u^{2k_1} \cdots u^{2k_d} \qquad k_i \in \mathbb{N}$$

$$\gamma(g(u))\,|g'(u)| = \phi(u) u^h = \phi(u) u^{h_1} \cdots u^{h_d} \qquad h_i \in \mathbb{N}$$

(2.2)

Hence (2.1) may be further refined to

$$\int_W H(w)\gamma(w)\,dw = \sum_\alpha \int_{M_\alpha} H(g(u))\,\phi_\alpha^*(u)\,u^h\,du$$

$$\text{where } \phi_\alpha^*(u) = \beta_\alpha(u)\phi(u). \tag{2.3}$$

Calculating RLCTs consider the zeta function of the triple $(P, q, \mathcal{J}, W)$

$$\zeta(z) = \int_W K(w)^z \mathcal{J}(w)\, dw \qquad (\mathrm{Re}(z) > 0) \qquad (3.1)$$

which can be analytically continued [W, Thm 6.6] to a unique meromorphic function on the entire complex plane whose poles are all real, negative and rational. The learning coefficient λ is by defⁿ such that $-\lambda$ is the largest pole of $\zeta(z)$. Here $K(w)$ is defined by

$$K(w) = \int q(x) \log \frac{q(x)}{p(x|w)}\, dx \qquad (3.2)$$

Note that the learning coefficient depends, in principle, on the prior $\mathcal{J}$, whereas the Real Log Canonical Threhold (RLCT), defined in [W, Defⁿ 2.7] for the pair (W, K) is associated to $\{w \mid K(w) = 0\}$ independently of the prior. As we will see, and as stated in [W, Remark 6.7(1)] provided that the prior is positive on $\{w \mid K(w) = 0\}$ (actually a weaker condition suffices) the learning coefficient is equal to the RLCT, and in particular is independent of the prior.

The first remark (proof of [W, Theorem 6.6]) is that for $\mathrm{Re}(z) > 0$

$$\zeta(z) = \int_{K(w)<\epsilon} K(w)^z \mathcal{J}(w) dw \longleftarrow \zeta_1(z) \qquad (3.3)$$

$$+ \int_{K(w)\geq\epsilon} K(w)^z \mathcal{J}(w) dw \longleftarrow \zeta_2(z)$$

and the second summand extends to a holomorphic function on the entire complex plane, so in analysing λ we may restrict to integrals over $W\epsilon$ as above.

Then using (2.3)

$$\zeta_1(z) = \sum_\alpha \int_{M_\alpha} u^{2kz} u^h \phi_\alpha^*(u)\, du \qquad (4.1)$$

Note that $b(u)$ in the statement of [W, Thm 2.3] is real analytic, so $\phi(u)$ may be taken real analytic provided $\mathcal{S}$ is (and we assume so). In this case, by shrinking the M_α if necessary, we may arrange $\phi_\alpha^*(u)$ to actually have an expansion as

$$\phi_\alpha^*(u) = \sum_{|j| \le N} a_j^{(\alpha)} u^j + R_N(u). \qquad (4.2)$$

Therefore in the region $\mathrm{Re}(z) > 0$, taking N_α sufficiently large,

$$\zeta_1(z) = \sum_\alpha \sum_{|j| \le N_\alpha} \int_{M_\alpha} a_j^{(\alpha)} u^{2kz} u^h u^j\, du \qquad (4.3)$$

+ terms with holomorphic extension to all of $\mathbb{C}$.

Since $M_\alpha = [0,b)^d$ this is

$$= \sum_\alpha \sum_{|j| \le N_\alpha} a_j^{(\alpha)} \int_{[0,b]^d} u^{2kz+h+j}\, du \qquad (4.4)$$

$$= \sum_\alpha \sum_{|j| \le N_\alpha} a_j^{(\alpha)} \prod_{p=1}^{d} \frac{b^{2k_p z + h_p + j_p + 1}}{(2k_p z + h_p + j_p + 1)}$$

Hence the function $\zeta_1(z)$ can be analytically continued as claimed, with poles at the rational numbers (where h, k vary with α and $1 \le p \le d$)

$$z = -\frac{h_p + j_p + 1}{2k_p} \qquad (4.5)$$

Then clearly the learning coefficient is

$$\lambda = \min_{\alpha} \min_{1 \le p \le d} \left(\frac{h_p + 1}{2k_p} \right). \qquad (5.1)$$

and its order (the power of $z+\lambda$ appearing in (4.4)), is

$$m = \max_{\alpha} \#\{ p \mid \lambda = \frac{h_p+1}{2k_p} \}. \qquad (5.2)$$

From this we can derive (see [W, p.32, p.33] for the short version) e.g.

$$F_n^\circ \cong \lambda \log n - (m-1) \log\log n + F^R(\xi). \qquad (5.3)$$

Remark Assuming $\varphi > 0$ on $\{w \mid K(w) = 0\}$ we may in the above apply [W, Thm 2.3] so that h depends only on (W, K), and then multiply by φ and still satisfy the conditions of [W, Thm 6.5 (3)]. That is, we may arrange on each α for $h \in \mathbb{N}^d$ to be independent of the prior. Since k is clearly independent, we have that $\lambda = \lambda(p, q, W)$ is independent of φ and agrees with the RLCT of [W, Def^n 2.7]. Observe that the RLCT is defined as

$$\text{RLCT} = \inf_{w \in W} \min_{1 \le p \le d} \left(\frac{h_p + 1}{\tilde{k}_p} \right).$$

[h is from the Jacobian det only, see e.g. p. 62 for regular case]

Actually we only really need $\varphi > 0$ on those M_α for which the minimization in (5.1) is obtained. (check)

Remark The prior does affect the Schwartz distributions $D_{km}(u)$ on [W, p. 32] and hence the random variable ξ in (5.3).

Def^n We say $w \in W$ is essential for the tuple (P, q, W) if the local resolution of singularities around $P = w$ in [W, Def^ 2.7] produces

$$\frac{h_p + 1}{\widetilde{k_p}} = \lambda \qquad \text{(the learning coefficient)}$$

Necessarily $w \in W_0 = \{w \mid K(w) = 0\}$ is a singular point of W_0, and it is an instance of the "worst" singularity type.

Local lambda

In the above we fixed a tuple $(P, q, \mathcal{S}, W)$. By Fundamental condition (II), [W, Def^ 6.3] we have assumed that W is a compact set given by

$$W = \{ w \in W^{(R)} \mid \pi_1(w) \geq 0, \pi_2(w) \geq 0, \ldots, \pi_k(w) \geq 0 \}$$

where $\pi_1, \ldots, \pi_k : W^{(R)} \longrightarrow \mathbb{R}$ are real analytic and $W^{(R)} \subseteq \mathbb{R}^d$ is open. This hypothesis is used in a subtle but crucial way in [W, Remark 2.14], i.e. the construction of the M_α, since we assume the resolution $g : U \to W$ makes $\pi_1(g(u)), \ldots, \pi_k(g(u))$ normal crossing, i.e. (see [W, p. 70]) for any $u_0 \in U$, there is a local coordinate u s.t.

$$\pi_i(g(u)) = a_i(u) u^{k_{i1}} \cdots u^{k_{id}}$$

where $a_i(u)$ is everywhere nonzero. This means we may shrink the $O_p(b)$ above if necessary, so that on each M_α the function $\pi_i(g(u))$ takes a consistent sign.

Now let $C \subseteq W$ be compact, defined by

$$C = \{ w \in W \mid f_1(w) \geq 0, \ldots, f_\ell(w) \geq 0 \}$$

where f_j are real-analytic. Then adding the f_j to the list of functions to be "resolved" (i.e. made normal crossings) we may assume that the local coordinates u and $M_\alpha \subseteq \mathcal{M}$ used above are adapted to C in the sense that the boundary of $g^{-1}(C)$ is contained in $u_1 \cdots u_d = 0$ in every such coordinate chart [W, Remark 2.12]. We assume ← i.e. realisability for C $C_0 = \{ w \in K \mid K(w) = 0 \}$ is nonempty, and $\varphi' > 0$ on C_0.

Lemma Let λ denote the learning coefficient of the tuple (p, q, φ, W) and λ_C the learning coefficient of (p, q, φ', C) for any prior φ' as above. Then $\lambda_C \geq \lambda$.

Proof We have

$$g^{-1}(C) = \bigcup_\alpha g^{-1}(C) \cap M_\alpha .$$

In the local coordinates u appropriate to M_α, $g^{-1}(C) \cap M_\alpha$ is given by $f_1(g(u)) \geq 0, \ldots, f_\ell(g(u)) \geq 0$ and we may assume the balls $O_p(b)$ are chosen small enough so that for each index α in the relevant local coordinate

$$f_j(g(u)) = c_j(u) u^{q_{j1}} \cdots u^{q_{jd}}$$

where we may assume $c_j(u)$ takes a fixed sign on M_α.
By construction M_α is a product of $(-b, 0]$ and $[0, b)$ intervals, and let m_α be the number of occurences of $(-b, 0]$.

We may assume these are the coordinates $u_1, \dots, u_{m_\alpha}$

Then $f_1(g(u)) \geq 0, \dots, f_\ell(g(u)) \geq 0$ on M_α

$$\iff c_j(u) u^{q_{j1}} \cdots u^{q_{jd}} \geq 0 \text{ on } M_\alpha \quad \forall\, 1 \leq j \leq d$$

$$\iff (-1)^{\gamma_j} u^{q_j} \geq 0 \text{ on } M_\alpha \quad \forall\, 1 \leq j \leq d \qquad (\gamma_j = \text{sign of } c_j)$$
$$(q_j := q_{j1} + \cdots + q_{jd})$$

$$\iff (-1)^{\gamma_j} u_1^{(q_j)_1} \cdots u_{m_\alpha}^{(q_j)_{m_\alpha}} \geq 0 \text{ on } M_\alpha \quad \forall\, 1 \leq j \leq d$$

$$\iff \gamma_j + (q_j)_1 + \cdots + (q_j)_{m_\alpha} \text{ is even}, \quad \forall\, 1 \leq j \leq d.$$

So either $\gamma_j + (q_j)_1 + \cdots + (q_j)_{m_\alpha}$ is even for all $1 \leq j \leq d$, in which case $g^{-1}(C) \cap M_\alpha = M_\alpha$, or this sum is odd for some $1 \leq j \leq d$ in which case

$$g^{-1}(C) \cap M_\alpha = \{ u \in M_\alpha \mid u_1 = \cdots = u_{m_\alpha} = 0 \}$$

Let Λ denote the set of α indices, $\Lambda' \subseteq \Lambda$ the set of indices for which $\gamma_j + (q_j)_1 + \cdots + (q_j)_{m_\alpha}$ is even.

By def^N λ_C is the largest pole of

$$\zeta_C(z) = \int_C K(\omega)^z \varphi'(\omega)\, d\omega$$

By restricting to W_ϵ, C_ϵ if necessary we may assume $\varphi > 0$ on W and $\varphi' > 0$ on C. Then in the partition of [W, Thm 6.5] we can assume $|g'(u)| = \phi^{\text{pre}}(u) u^h$ and $\varphi(g(u)) |g'(u)| = \varphi(g(u)) \phi^{\text{pre}}(u) u^h$

so that $\phi(u) = f(g(u))\,\phi^{pre}(u)$, hence in (2.3)

$$\phi^*_\alpha(u) = \delta_\alpha(u)\,\phi(u) = \delta_\alpha(u)\, f(g(u))\,\phi^{pre}(u).$$

We now compute for $\mathrm{Re}(z) > 0$ using (2.3)

$$\zeta_c(z) = \int_c K(w)^z f'(w)\,dw \qquad (k,h \text{ depending on } \alpha)$$

$$= \sum_\alpha \int_{g^{-1}(c)\cap M_\alpha} K(g(u))^z f'(g(u))\,\phi^{pre}(u)\,u^h\,du$$

$$= \sum_\alpha \int_{g^{-1}(c)\cap M_\alpha} u^{2kz}\,u^h\,\left[f'(g(u))\,\phi^{pre}(u)\right]du$$

By the earlier argument (e.g. in (4.2)) we may ignore, for the purpose of computing λ_c, the term in brackets involving f'. So we must compute the largest pole of

$$\sum_\alpha \int_{g^{-1}(c)\cap M_\alpha} u^{2kz+h}\,du$$

$$= \sum_{\alpha\in\Lambda'} \int_{M_\alpha} u^{2kz+h}\,du + \sum_{\alpha\notin\Lambda'} \int_{\{u_1=\cdots=u_{m_\alpha}=0\}} u^{2kz+h}\,du$$

$$= \sum_{\alpha\in\Lambda'} \prod_{p=1}^{d} \int_0^b u_p^{2k_p z + h_p}\,du_p$$

$$= \sum_{\alpha\in\Lambda'} \prod_{p=1}^{d} \frac{b^{2k_p z + h_p}}{2k_p z + h_p + 1} \qquad (9.1)$$

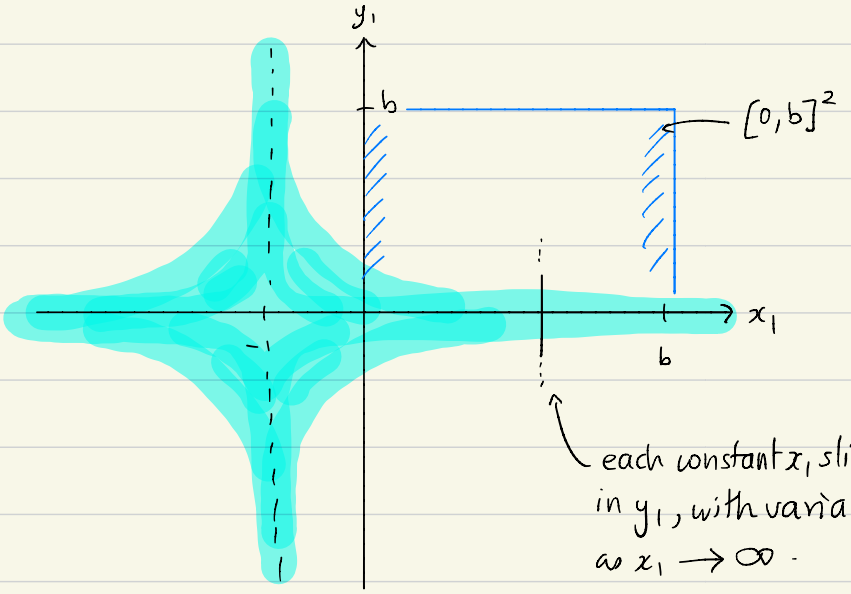
Note that these tuples $k, h \in \mathbb{N}^d$ are the same as used to define λ, so that

$$\lambda_c = \min_{\alpha \in \Lambda'} \min_{1 \le p \le d} \left(\frac{h_p + 1}{2 k_p} \right) \tag{10.1}$$

$$\geqslant \min_{\alpha} \min_{1 \le p \le d} \left(\frac{h_p + 1}{2 k_p} \right) = \lambda . \quad \square$$

Resolutions and the posterior

Consider the function $e^{-(1+x_1^2)y_1^2}$, viewed as a density:



(10.2)

If we set $x = x_1 y_1$ and $y = y_1$ this becomes a standard Gaussian on $x, y \geqslant 0$

$$e^{-(1+\frac{x^2}{y^2})y^2} = e^{-x^2-y^2} \tag{10.3}$$

Put differently, if $g : [0,b]^2 \longrightarrow \mathbb{R}^2$ is $g(x_1, y_1) = (x_1 y_1, y_1)$ and $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$ is $f(x,y) = \exp(-x^2 - y^2)$ then $f(g(x_1, y_1)) = \exp(-(1+x_1^2)y_1^2)$.
This is (one coordinate patch of) the resolution of $x^2 + y^2$ [W, Example 2.7].

The resolution procedure [W, §1.4.2] expresses the measure $\exp(-n\beta K_n(w))\,\varphi(w)\,dw$, whose normalisation is the Bayesian posterior, as asymptotically approximated by a sum

$$\sum_{\alpha} \exp\left(-n\beta u^{2k} + \sqrt{n}\beta u^{k} \xi_n(u)\right) \phi^*(u) |u^h| du \qquad (11.1)$$

For the sake of building intuition let us ignore the stochastic process $\xi_n(u)$, and the $\phi^*(u)$ (which is made up of the partition of unity and prior). The $|u^h|$ is a Jacobian factor, which corresponds to pushing forward the measure

$$\exp\left(-n\beta u^{2k}\right) du = \exp\left(-n\beta u_1^{2k_1} \cdots u_d^{2k_d}\right) \qquad (11.2)$$

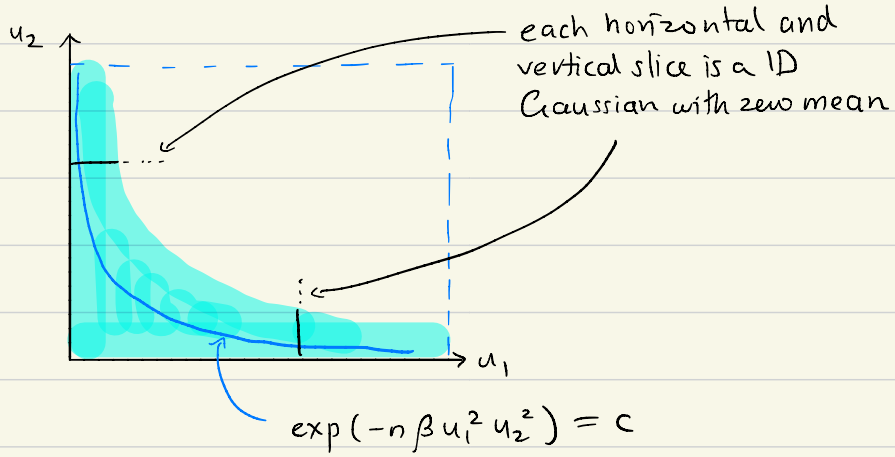
along $g : U \longrightarrow W$. Recall that the index α is associated with a product $[0,b]^d$ in some local coordinates [W, Remark 2.14] and (11.1) is to be interpreted as each of these local patches $M_\alpha \cong [0,b]^d$ contributing additively to approximate the posterior.

So what kind of function is (11.2) on $M_\alpha = [0,b]^d$?

$\boxed{d=2}$ $k=(1,1)$

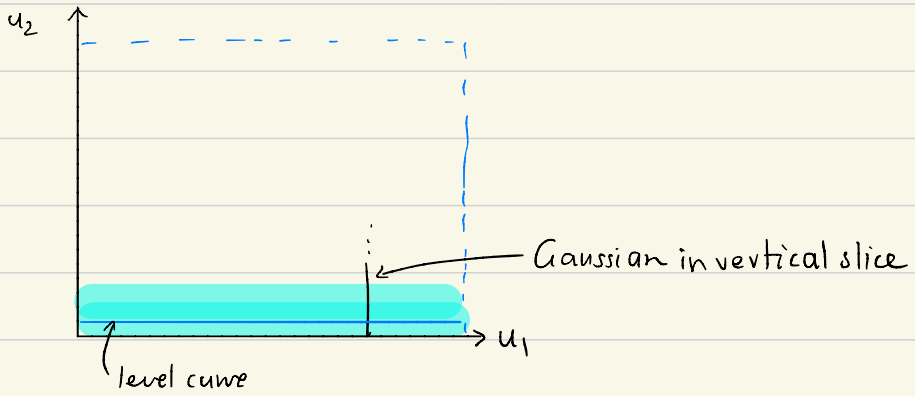
level sets $u_1^2 u_2^2 = -\log c$

$0 \le c \le 1$ $u_1 u_2 = \sqrt{-\log c}$



d=2 $k=(0,1)$

level sets $u_2^2 = -\log c$
$0 \le c \le 1$ $u_2 = \sqrt{-\log c}$

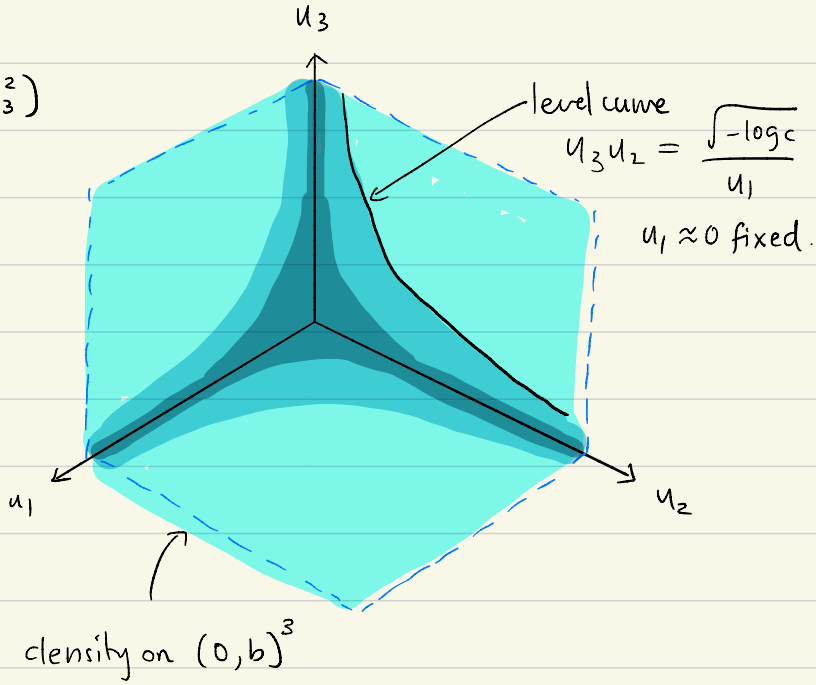


Similarly for $k=(1,0)$. The other cases are similar, but the approach to the coordinate hyperplanes is more gradual than Gaussian, e.g. in $k=(0,2)$ we have $\exp(-n\beta u_2^4)$ so the measure is more spread out along the u_2 axis.

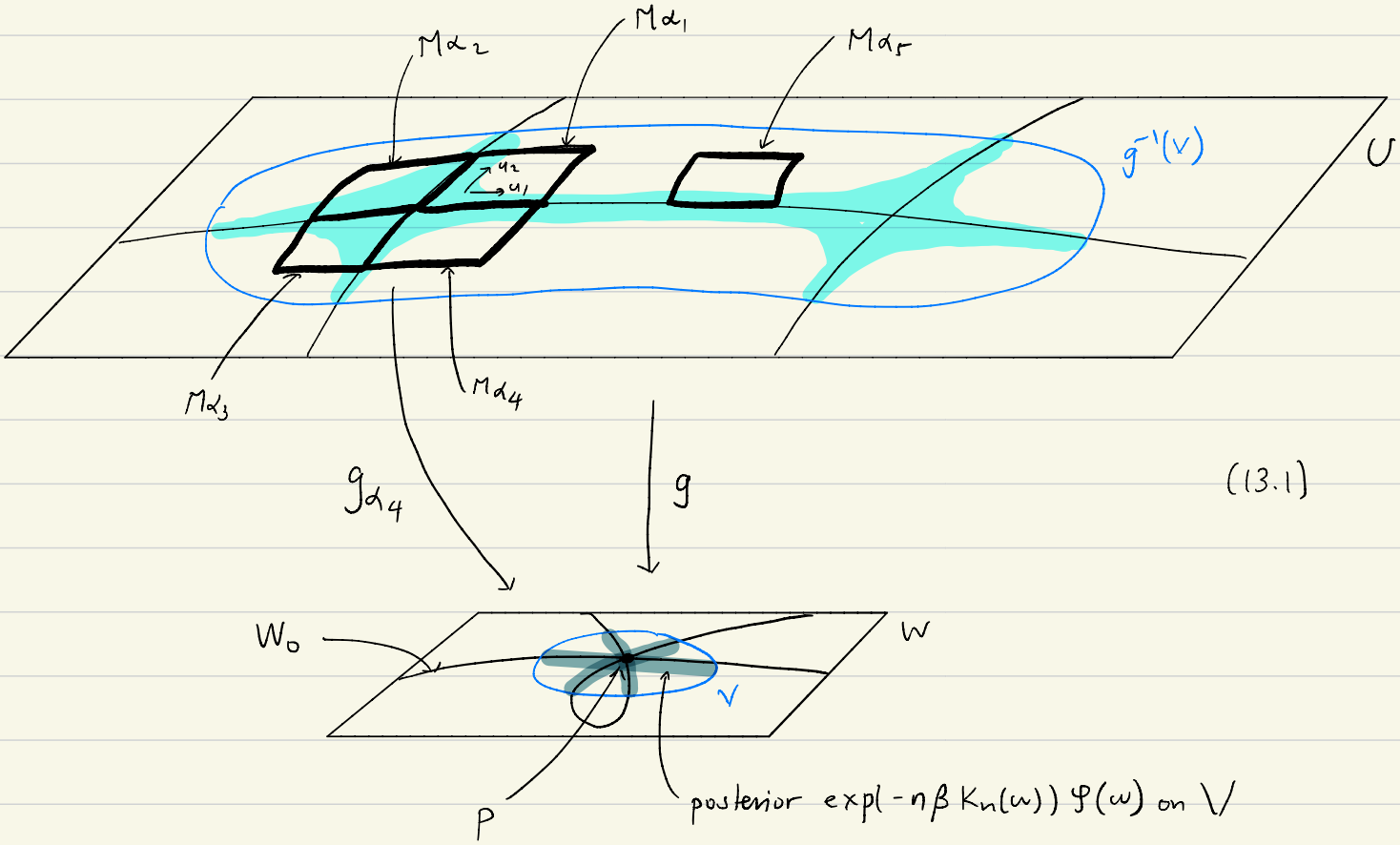
In higher dimensions, e.g. $d=3$, the picture is similar. We pick some subset of the coordinate hyperplanes (those with the corresponding coordinate of k nonzero), place a large measure along those hyperplanes and "round off the corners" near $u=0$. Higher indices k_i correspond to flatter approaches along that coordinate.

d=3 $k=(1,1,1)$ $\exp(-u_1^2 u_2^2 u_3^2)$

level set $-u_1^2 u_2^2 u_3^2 = \log c$
$0 \le c \le 1$ $u_1 u_2 u_3 = \sqrt{-\log c}$



We now imagine a collection of such measures on sets M_α, each with their own map $g_\alpha : M_\alpha \longrightarrow V$ to some neighborhood V of W, such that a weighted sum over α of the pushforward of these measures along each g_α approximates the posterior on V. To borrow from [W, Fig 2.5, Fig 2.10]:



(13.1)

This suggests a potential generalisation of implicit variational inference, which we term <u>microscopic variational inference</u> (MVI). We fix an integer S and let $G_{\theta_1}, \ldots, G_{\theta_S}$ be feedforward networks with weight vectors θ_i, whose outputs are in $\mathbb{R}^d$ and whose inputs are in $[0,1]^d$. For each $1 \leq i \leq S$ we choose a vector $k_i = (k_{i1}, \ldots, k_{id}) \in \mathbb{N}^d$ and we introduce additional weights $a_1, \ldots, a_S$ which are logits for weighting each network.

We want to sample, for each i, from a normalisation of $\exp(-t_i u^{k_i})$ where t_i is some (inverse) temperature, pass that sample through G_{θ_i} and sum these outputs with weights from a softmax of $(a_1, \ldots, a_S)$.

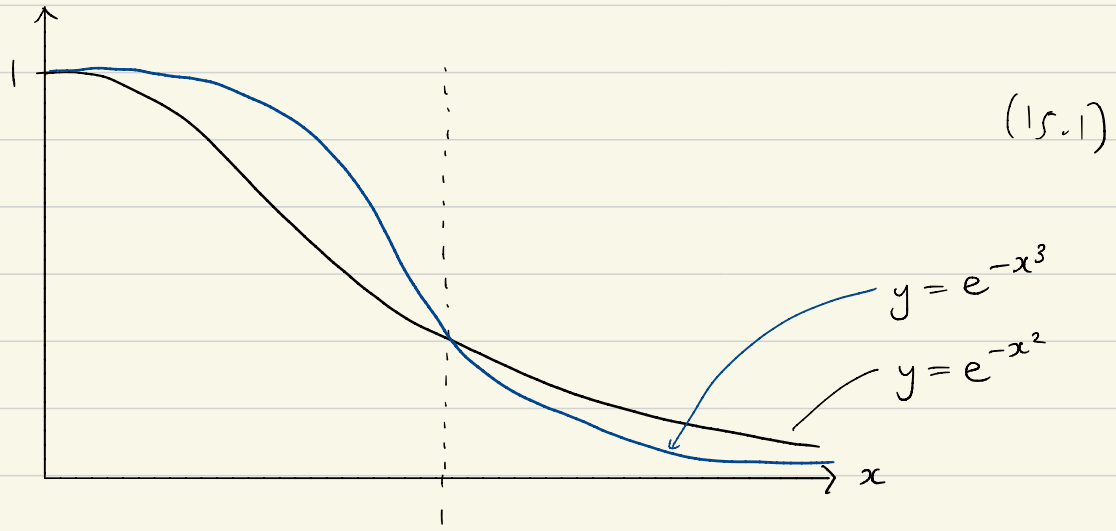
Set $Z_i := \int_{[0,1]^d} \exp(-t_i u^{k_i})\, du$. We assume some target weight vector $P \in W$ is specified (as in (13.1)) and we wish to approximate the posterior in a small neighborhood of P, e.g. to estimate the local RLCT. So we use the output of each $G_{\mathcal{O}_i}$ as a delta from P, so that our probability distribution associated to the parameters $(\mathcal{O}_1, \dots, \mathcal{O}_S, a_1, \dots, a_S)$ is

$$\sum_{i=1}^{S} \frac{\exp(a_i)}{\sum_j \exp(a_j)} (G_{\mathcal{O}_i})_* \left(\frac{1}{Z_i} \exp(-t_i u^{k_i}) \right) + P \qquad (14.1)$$

where $(G_{\mathcal{O}_i})_*$ means pushforward along the function computed by the network $G_{\mathcal{O}_i}$.

The idea behind MVI is that in the singular case, it is not enough to think about modes and Gaussians; you need to think in terms of the distributions $\exp(-n\beta u^k)$ on p. (11), (12) more generally, and how to sum transformations of these. By universality and the resolution theorem (eliding the difference between partition of unity and the softmax) for some weights $\mathcal{O}_1, \dots, \mathcal{O}_S$ you can approximate the true posterior this way (asymptotically, as in [W, p. 33], but this enough to compute local RLCTs) as $S \to \infty$, the k_i's range over all tuples and the depths of the $G_{\mathcal{O}_i}$ go to infinity.

Remark Consider the following graphs



(15.1)

For $x \approx 0$ we have

$$\frac{d}{dx}\exp(-x^2) = -2x\exp(-x^2) \approx -2x + O(x^3)$$

$$\frac{d}{dx}\exp(-x^3) = -3x^2\exp(-x^2) \approx -3x^2 + O(x^4) \qquad (15.2)$$

So the larger k_i is, the flatter $\exp(-n\beta u^k)$ is as we approach the ith coordinate hyperplane in $[0,b]^d \cong M_d$. But it is actually the transformed measure (11.1) we are interested in. It is not simple to explain the interaction of k, h in this formula directly, so to understand how h, k combine to determine the local behaviour of the posterior, we switch to talking about volumes.

Effective number of parameters and the RLCT

Consider a regular model in a neighborhood of the true parameter, which we may assume is $w_0 = 0$, so that near w_0 (possibly after changing coordinates)

$$K(w) \approx \sum_{i=1}^{d} w_i^2 \qquad (16.1)$$

Consider the volume of almost correct parameters

$$V(t) = \int_{K(w)<t} \varphi(w)\, dw. \qquad (16.2)$$

We know the prior is (almost) irrelevant for the purposes of understanding the RLCT, so for simplicity we take it to be uniform so that $V(t)$ is the volume of all d-ball of radius $t^{1/2}$

$$V(t) \propto t^{d/2} \qquad (16.3)$$

Hence for $a > 0$, $a \neq 1$ we have

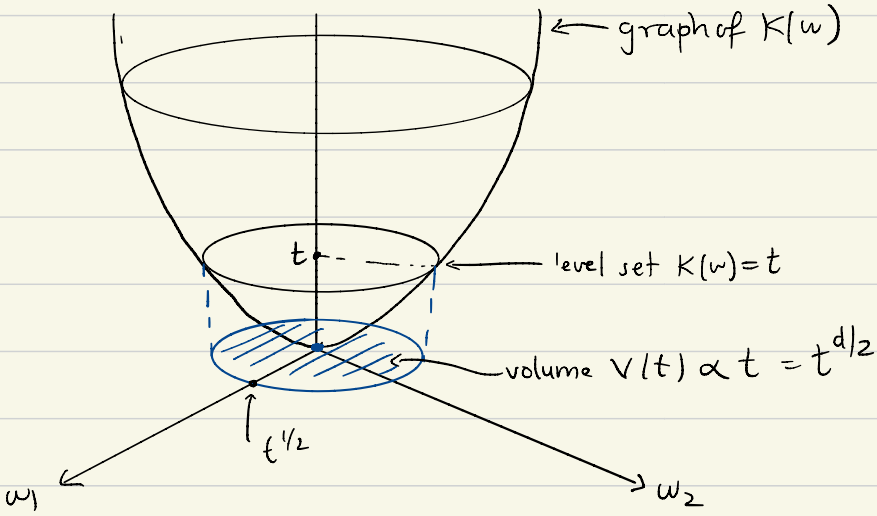
$$\log\{ V(at)/V(t) \} = \log\{ a^{d/2} t^{d/2} / t^{d/2} \} \qquad (16.4)$$

$$= \log\{ a^{d/2} \} = d/2 \log a.$$

Hence [W, Theorem 7.1] gives $\lambda = d/2$, as the RLCT.

In the regular case 2λ is the number of parameters. We now proceed to justify the claim that 2λ is an appropriate definition of the effective number of parameters in general, starting with a mildly singular case.

Remark In the case $d=2$ the relevant diagram is



(17.1)

Mildly singular case suppose now that near the true w_0 we have

$$K(w) \approx \sum_{i=1}^{d'} w_i^2 \tag{17.2}$$

where $d' < d$. Then this model is singular, and $K(w) < t$ imposes no constraint on the variables $w_{d'+1}, \dots, w_d$ so that (with a nice choice of prior and local neighborhood, but the details don't matter)

$$\{w \mid K(w) < t\} = B^{d'}_{t^{1/2}}(w_0) \times [0,1]^{d-d'} \tag{17.3}$$

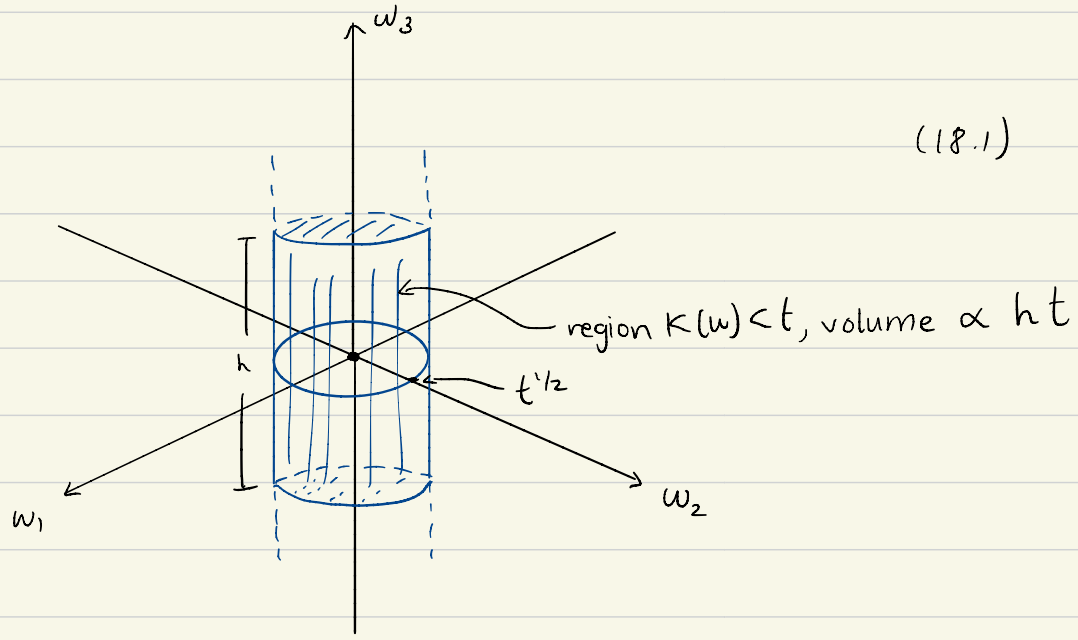
where $B^{d'}_{t^{1/2}}(w_0)$ is a d'-ball of radius $t^{1/2}$ around w_0. The volume is

$$V(t) \propto t^{d'/2} \tag{17.4}$$

and hence the RLCT is $d'/2$. The RLCT (doubled) is the dimension of the normal directions to $W_0 = \{w \in W \mid K(w) = 0\}$ near w_0.

Note that, at least near w_0, $\{w \mid K(w)=0\}$ has a free variable w_i for $i > d'$ and hence dimension $d-d'$, so the normal bundle is d' dimensional. Hence $2\lambda = d'$ is the number of directions we can vary w_0 which change the KL divergence. That is, 2λ is the effective number of parameters in the model near w_0.

In the case $d=3$, $d'=2$ the relevant diagram is (supposing $K(w) \approx w_1^2 + w_2^2$)



(18.1)

General singular case in general while $W_0 = \{w \mid K(w)=0\}$ is a real analytic set it is not a submanifold, and near a singular point there is no naive notion of a "normal direction". So we cannot simply count the number of normal directions (for the same reason the tubular neighborhood theorem fails in algebraic geometry) as we have done in (17.1) (where every direction was normal, as W_0 was a point) and (18.1) (where W_0 was a line and there are two normal directions).

Remark If $w_0 \in W_0$ is a nondegenerate critical point of W_0 (noting that the degenerate ones have measure zero, so this accounts for "most" true parameters) then the local λ near w_0 will simply be $d'/2$ where d' is the number of normal directions to W_0 near w_0 (as W_0 looks like a submanifold near w_0).

However the volume ratios, which give the dimension of the normal bundle in the regular and "mildly singular" cases, are still well-defined and by [W, Thm 7.1]

$$\lambda = \lim_{t \to 0} \frac{\log\{V(at)/V(t)\}}{\log a} \qquad (a>0, a \neq 1)$$

Here is how to think about this: suppose someone gave you the loss surface $\{(w, K(w)) \mid w \in W\}$ and a point $w_0 \in W$. You start computing volumes $V(0.99t), V(t)$ for some small values of t. This involves cutoffs and priors but the ratio $V(0.99t)/V(t)$ is basically independent of these. And that someone says to you: <u>if this were a mildly singular model, what would the dimension of the normal bundle be?</u>

That is, what dimensionality would these volumes have, if they were d' balls for some d'? The answer to this question is what 2λ computes, hence why Watanabe refers to this as <u>volume dimension</u> in [W, Remark 7.1].

Of course 2λ need not be an integer, but this corresponds to the fact that there is no "true" normal bundle, so there cannot really be an integral "effective" number of parameters in a singular model at a degenerate critical point.

<u>Remark</u> The RLCT is not a measure of "flatness" in the same sense as e.g. the determinant of the Hessian: note that in the regular case it is $d/2$ regardless of the eigenvalues. Instead the RLCT is a kind of "count" of the <u>number of flat directions</u>, at least indirectly, since

$$\text{small } \lambda \implies \text{small normal bundle} \implies \text{many flat directions.}$$

Learning coefficient vs. generalisation error

Suppose as above we have (p, q, W, φ) and $C \subseteq W$ compact, with $q(x)$ realisable by some parameter in C, let λ_C be the learning coefficient for C. Then with (p, q, C, φ') as our input to [W, §6] we have (at $\beta = 1$)

$$\mathbb{E}[B_g^C] = \frac{\lambda_C}{n} + o\left(\frac{1}{n}\right) \qquad (20.1)$$

where B_g^C is the Bayes generalisation error [W, (6.1)]. By [W, Defn 6.5]

$$B_g^C := \mathbb{E}_x\left[\log \frac{q(x)}{\mathbb{E}_w[p(x|w)]}\right]$$

where $\mathbb{E}_x$ means $\int q(x)(-)\,dx$ and

$$\mathbb{E}_w[p(x|w)] = \int p(w|D_n)\, p(x|w)\, dw$$

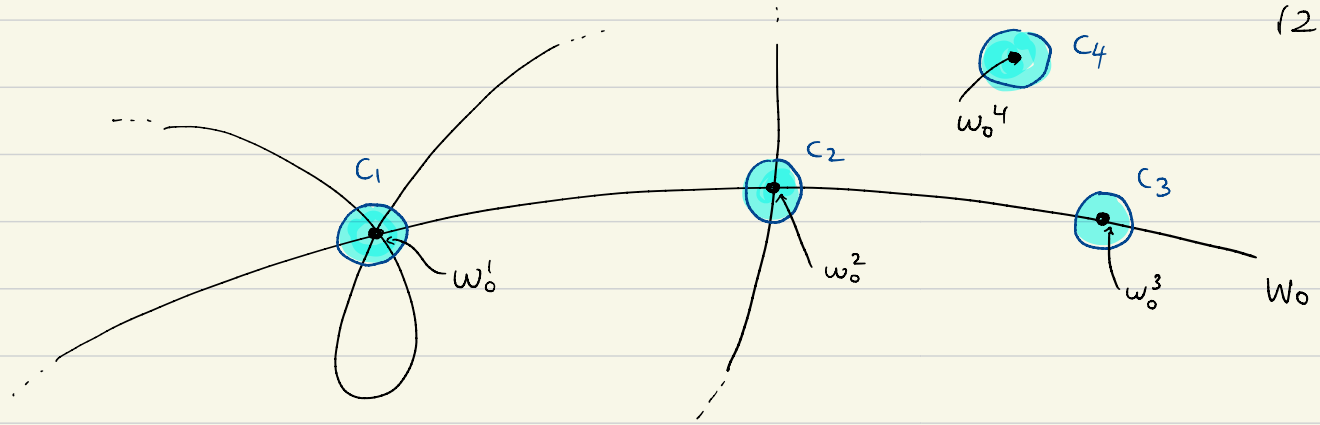
$$= \frac{1}{Z_n^\circ} \int p(x|w) \exp(-n\beta K_n(w))\, dw$$

is the C-restricted predictive distribution. So B_g^C is the KL divergence between the true distribution and the C-restricted predictive distribution, and the stochasticity involved in sampling D_n makes B_g^C a random variable. The expectation in (20.1) integrates out this stochasticity.

Conclusion: the big picture

So the picture is the following: supposing q is realisable by (p, W), the set $W_0 = \{w \in W \mid K(w) = 0\}$ is nonempty, and in general a complex real analytic set with multiple components and singularities:

(21.1)



Shown are three points $w_0^1, w_0^2, w_0^3, w_0^4 \in W_0$ and compact $C_i \subseteq W$ containing them. The point w_0^4 is an isolated zero of $K(w)$ and while it may nonetheless be a degenerate critical point, let us imagine it to be nondegenerate so $\lambda_{C_4} = d/2$. The point w_0^3 is a smooth point of W_0 and looks perhaps like the "mild" case above, i.e. locally $K(w) \approx \sum_{i=1}^{d'} w_i^2$, with $\lambda_{C_3} = d'/2 < d/2$, and locally d' honest normal directions to the set of true parameters.

The point w_0^2 is a more complex singularity, and w_0^1 more complex still, so that we have

$$\lambda_{C_1} < \lambda_{C_2} < \lambda_{C_3} < \lambda_{C_4} = d/2$$

Hence at a given dataset size, we have the same relationship between the Bayesian generalisation errors. The models near w_0^1 have the least number of effective parameters, or degrees of freedom, and generalise best.