Singular Learning Theory XII : Density of states

The state density function $v(t)$ of [W, §4.2] is one of the most important mathematical objects in singular learning theory. Its asymptotic behaviour as $t \to 0$ is controlled by the singularities of the set of true parameters, and in particular by the RLCT λ, and it is via this dependence that λ can be seen to control the Bayesian posterior [W, p. 33]. Thus, in order to understand singular learning theory we must understand the state density function.

1. Introduction

In physics, particularly in solid state physics and condensed matter physics, the density of states (DOS) is a fundamental mathematical object that reflects the basic structure of the physical system. For a system with N discrete configurations (e.g. a quantum particle in a box) and volume V the density of states is

$$D(E) = \frac{1}{V} \sum_{i=1}^{N} \delta(E - E_i) \qquad (1.1)$$

where E_i is the energy of the ith state, and $\delta(E - E_i) = 1$ if $E_i = E$ and zero otherwise. In this case the DOS is simply a histogram counting the number of states (per unit volume) with a given energy. For a continuous system

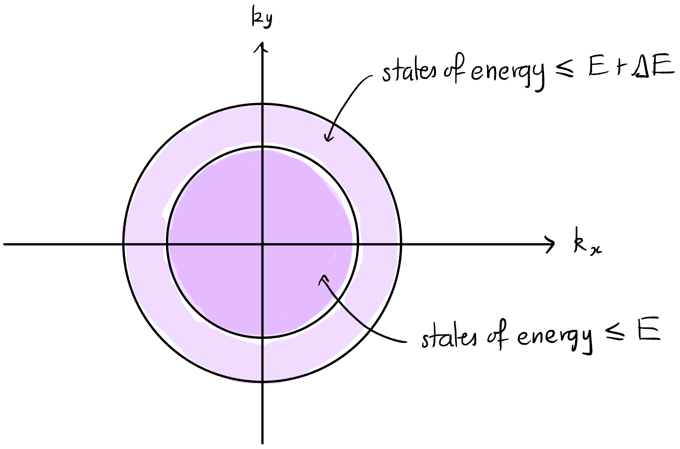
$$D(E) = \frac{1}{V} \cdot \frac{N(E+\Delta E) - N(E)}{\Delta E} \qquad (1.2)$$

where $N(E)$ is the number of states with energy $\leq E$. That is, $D(E)$ is the number of new states (per unit volume and energy) accessible to the system when the allowed maximum energy is increased from E to $E + \Delta E$. Of course, to make a rigorous definition "number" must be replaced by "measure" and some other subtleties dealt with. Before doing so, however, it will be instructive to consider a simple example where $D(E)$ is easily understood.

Example For a free electron gas in d dimensions [K, Ch.6, (6.20)] states are parametrised by wavevectors $k \in \mathbb{R}^d$ (frequencies for waves in three spatial directions) and the energy is $E(k) = \frac{\hbar^2}{2m} k^2$ where k^2 is the dot product of k with itself, i.e. $k_x^2 + k_y^2 + k_z^2$, if $d = 3$. The total number of states of energy $\leq E$ is therefore proportional to the volume of a sphere in k-space of radius $E^{1/2}$, so $N(E) \propto E^{d/2}$ and hence

$$D(E) \propto E^{d/2 - 1} \qquad (2.1)$$

This formula is used in the (rough, first) treatment of semiconductors in [K, Ch. 8] see [K, p.218, p.219], which exhibits the basic role the DOS plays in solid state physics. However if the energy E depends on the parameters (k in this case) in a less trivial way then the function $D(E)$ may be much more complicated; see the discussion of van Hove singularities in [K, p.129] and [YIF].


k_y
states of energy $\leq E + \Delta E$
k_x
states of energy $\leq E$

In mathematics the density of states is intimately related to the generalisation of Dirac distributions to higher-dimensional submanifolds (Dirac distributions being the "measurement" of functions at a point, that is, a zero-dimensional submanifold). The generalisation is in the following sense: any distribution concentrated on a point x is a linear combination of the Dirac distribution $\delta(x)$ and its derivatives [GS2, Ch.2 §4.5] and for a submanifold Z cut out by smooth equations $P_1 = \cdots = P_k = 0$ an analogous role is played by the "generalised Dirac distribution" $\delta(P_1, \ldots, P_k)$ and its derivatives [GS1, p.209]. The density of states "is" the parametrised family $\{\delta(t - E)\}_t$ of such distributions.

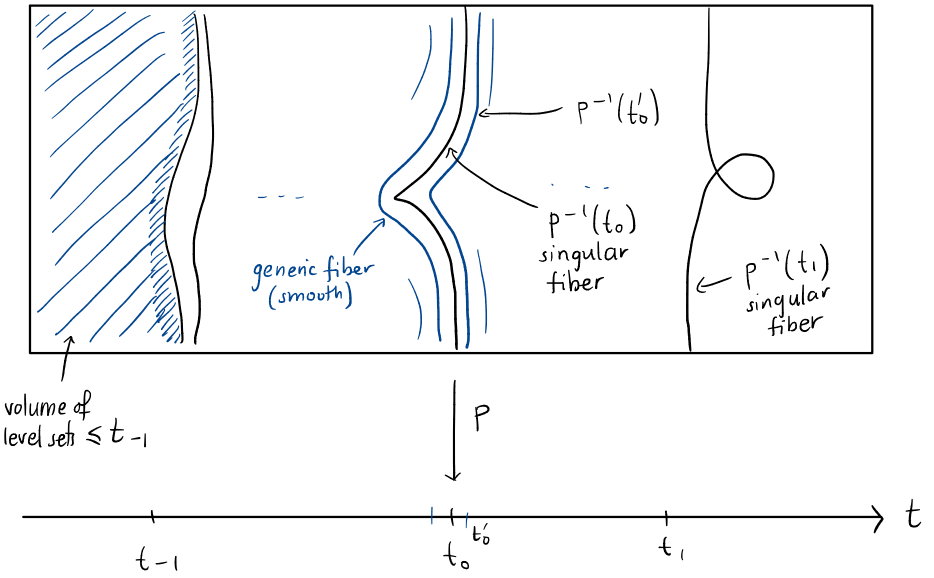
Remark I'm actually not sure of a precise reference for the general statement; it isn't in Gelfand-Shilov. I think it is equivalent to saying $\delta(P_1, \dots, P_k)$ generates a certain D-module and this is related to K. Vilonen "Intersection homology D-module on local complete intersections with isolated singularities" Invent. Math. 1985.

Remark One might expect derivatives $\delta^{(i)}(t-E)$ to play some role in singular learning theory, but I don't know what.

2. Volume between level sets

We fix $U \subseteq \mathbb{R}^n$ open and $P : U \longrightarrow \mathbb{R}$ smooth, following the notation of [AS1, Ch. 3]. In the physics context P would be the energy E, while in singular learning theory it would be the KL divergence K between the model and the truth.

Consider the partition $\{P^{-1}(t)\}_{t \in \mathbb{R}}$ of U into fibers of P (i.e. level sets)



$P^{-1}(t_0')$
$P^{-1}(t_0)$ singular fiber
generic fiber (smooth)
$P^{-1}(t_1)$ singular fiber
volume of level sets $\leq t_{-1}$
P
t
t_{-1}
t_0
t_0'
t_1

(3.1)

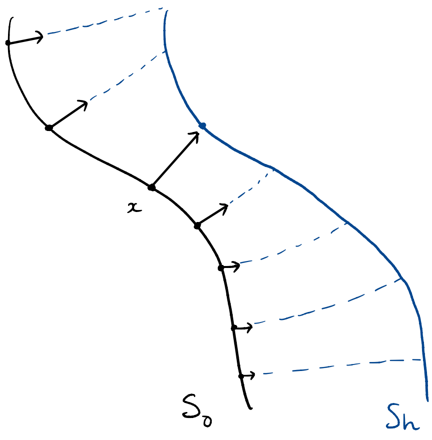
Two basic questions come to mind: which level sets are submanifolds, and as we "flood the graph" by considering $\{x \in U \mid P(x) \leq t\}$ for increasing values t, how quickly does the flooded area increase as a function of t? This is of course nothing else than the density of states.

For a generic t, the preimage $P^{-1}(t) \subseteq U$ will be a submanifold of dimension $n-1$. More precisely, recall [B, Theorem 5.8] that if ∇P is nonzero (that is, some partial derivative is nonzero) at every point $x \in U$ with $P(x) = t$ (here t is fixed) then P has rank 1 in an open neighbourhood of the closed set $P^{-1}(t) \subseteq U$ and hence $P^{-1}(t)$ is a regular submanifold. Now by Sard's theorem [GG, Theorem 1.12] the set of critical values of P (that is, those $t \in \mathbb{R}$ for which there exists $x \in P^{-1}(t)$ with $\nabla P(x) = 0$) has measure zero in $\mathbb{R}$. If t is *not* a critical value then $P^{-1}(t)$ is a submanifold, as we have just shown. Another way of saying this is that *a generic level set is a smooth submanifold*.

Def^n We set $\mathrm{Crit}(P) = \{ x \in U \mid \nabla P(x) = 0 \}$ and call $t \in \mathbb{R}$ a *regular value* if $P^{-1}(t) \cap \mathrm{Crit}(P) = \phi$ and a *critical value* otherwise. If t is a regular value we call $P^{-1}(t)$ a *regular fiber* otherwise it is a *singular fiber*.

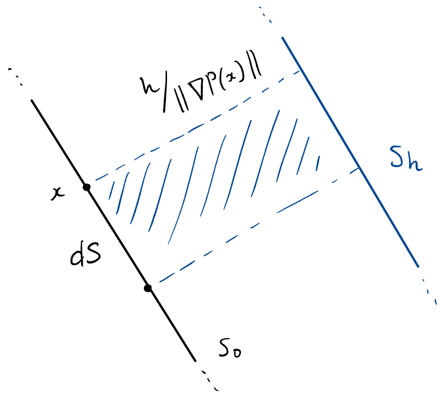
Remark While the singular fibers are "rare" they are in many situations "more important" than the regular fibers. For instance if $P = E$ is the energy then a fiber which contains a local minima $x \in U$ of the energy is a singular fiber. In particular if $P \geqslant 0$ (as is the case in singular learning theory) and $P^{-1}(0)$ is nonempty (i.e. the true distribution is realisable) then this fiber is singular.

Now let $t \in \mathbb{R}$ be a regular value. There is an open neighbourhood of t in $\mathbb{R}$ consisting of regular values, and for small h we consider the regular fibers $S_h = P^{-1}(t+h)$ and how they vary with h, as shown below, where the arrows show ∇P. Note that to a first approximation we have



$$P(x + h\nabla P(x)) \approx P(x) + \sum_i h \left(\frac{\partial P}{\partial x_i} \right)^2 = P(x) + h \| \nabla P \|^2 \qquad (4.1)$$

so that for h small, moving at x a distance $h / \| \nabla P(x) \|$ along the direction of ∇P gets you to the level set S_h.



(5.1)

Hence if we were to estimate the volume contained between the level sets S_0 and S_h, we could divide it into volume elements whose "base" is a volume element dS on S_0 and whose "height" is $h/\|\nabla P\|$, leading to the "definition" (φ being some function with compact support on $\mathbb{R}^n$ which we introduce so that the integrals are finite; you could think of this as the characteristic function of the compact set of allowed states in the physics setting)

$$\mathrm{vol}(t, t+h) = \int_{S_0} \frac{h\,\varphi(x)}{\|\nabla P\|}\, dS \qquad (5.2)$$

Going by (1.2) it would be reasonable therefore to identify the <u>density of states</u> at energy t_0 (viewing P as the energy) to be (where $V = \int_{\mathbb{R}^n} \varphi\, dx$)

$$D(t) = \frac{1}{V}\frac{1}{h}\mathrm{vol}(t, t+h) = \frac{1}{V}\int_{S_0} \frac{\varphi}{\|\nabla P\|}\, dS \qquad (5.3)$$

Indeed up to a prefactor this is the definition arrived at in [K, p. 128]. Note that the density of states is non-negative, and large if $\|\nabla P\|$ is small on the level set S_0, that is, if the graph of P is <u>relatively flat above S_0</u> (since t is a regular value it is never zero). Indeed $D(t)$ is itself a reasonable <u>measure of flatness</u> above $S_0 = \{x \mid P(x) = t\}$ on the graph. This measure only applies for <u>regular</u> values t, but it stands to reason we could try to study it at <u>singular</u> values t by considering the behaviour of $D(s)$ as $s \longrightarrow t$.

Example $P = \sum_{i=1}^n x_i^2$ then $\nabla P = 2(x_1, \dots, x_n)$ and the regular values of t are $\mathbb{R}\setminus\{0\}$. For $t \neq 0$ the level set $S_0 = P^{-1}(t)$ is a sphere of radius $t^{1/2}$ and

$$\int_{S_0} \frac{1}{\|\nabla P\|} dS = \int_{S_0} \frac{1}{2\|x\|} dS = \frac{1}{2\sqrt{t}} \cdot \int_{S_0} dS \propto t^{-1/2} \cdot \left(t^{1/2}\right)^{n-1} = t^{n/2 - 1} \qquad (6.1)$$

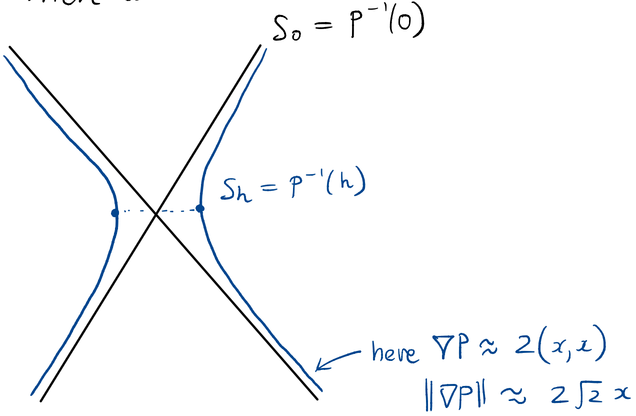
which is of course in agreement with (2.1). Note that $t=0$ is a singular value, with $P^{-1}(0) = \{0\}$, and $\lim_{t \to 0} D(t) = 0$ for $n > 2$.

Note that in this example the limit of $D(s)$ exists as s approaches a singular value. More generally if the only critical points $x \in P^{-1}(t)$ are local minima and the Hessian of P is nondegenerate at each such critical point, then $\lim_{s \to t} D(s)$ can be treated as a limit of a sum of contributions like (5.4) and so this limit exists and is zero. In singular learning theory these conditions hold for the zero level set of <u>regular models</u>. In general we expect that the integrand $\frac{1}{\|\nabla P\|}$ in (5.3) will be large and positive at points $x \in S_h$ which "converge" as $h \to 0$ to critical points of S_0 (see the next example), and that such "near critical points" will make <u>large positive</u> contributions to $D(s)$. However as the previous example shows, we cannot a priori rule out that also as $s \to t$ the <u>region</u> over which $\frac{1}{\|\nabla P\|}$ is large is simultaneously shrinking, so that the integral $D(s)$ ends up converging nonetheless as $h \to 0$. The next example shows a case where this "conspiracy" does <u>not</u> occur.

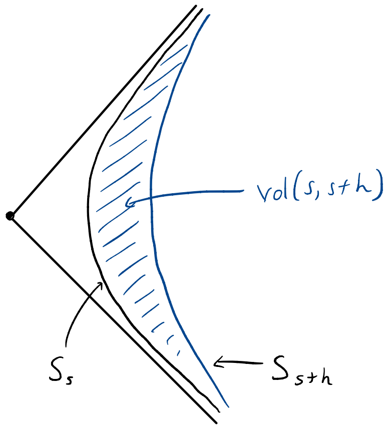
Example Set $P = x^2 - y^2$ and $t = 0$ which is a singular value. For small h, $S_h = P^{-1}(h)$ is a hyperbola and $(\pm\sqrt{h}, 0) \in S_h$ have $\|\nabla P\| = \sqrt{2h}$ so $\frac{1}{\|\nabla P\|} \longrightarrow \infty$ at these points as $h \to 0$. We set $\varphi(x,y) = e^{-(x^2+y^2)}$ then as $h \to 0$

$$\int_{S_h} \frac{\varphi}{\|\nabla P\|} dS \longrightarrow 4 \int_0^\infty \frac{e^{-x^2}}{2\sqrt{2}\,x} dx$$

which diverges to $+\infty$. This sketches an argument that $D(s) \longrightarrow \infty$ as $s \longrightarrow 0$.


$S_0 = P^{-1}(0)$
$S_h = P^{-1}(h)$
here $\nabla P \approx 2(x,x)$
$\|\nabla P\| \approx 2\sqrt{2}\,x$

The divergence $D(s) \longrightarrow \infty$ should not be confused with some volume between level sets going to infinity (as any such volume is bounded above by V). It means that for sufficiently small h and s sufficiently close to zero the ratio $\frac{1}{h}\,\mathrm{vol}(s, s+h)$ may be made arbitrarily large



(7.1)

Clearly then the asymptotic behaviour of the density of states $D(t)$ as t approaches a critical value is not easily understood for a general function P. It may not be immediately obvious, but the coefficients and exponents in this asymptotic expansion determine important physical properties in the case where P is an energy [YIF] and important observables such as the Bayes generalisation error when P is the KL divergence of a (p, q, φ) triple in singular learning theory. Indeed by taking the resolution [W, p. 32] shows that

$$D(s) \sim a\, s^{\lambda-1} \quad \text{as} \quad s \longrightarrow 0 \qquad (7.2)$$

where λ is the global RLCT. In particular this diverges to ∞ if $\lambda < 1$ and converges to zero if $\lambda > 1$ (e.g. in the regular case, which puts us in the setting of the example on p. 6)

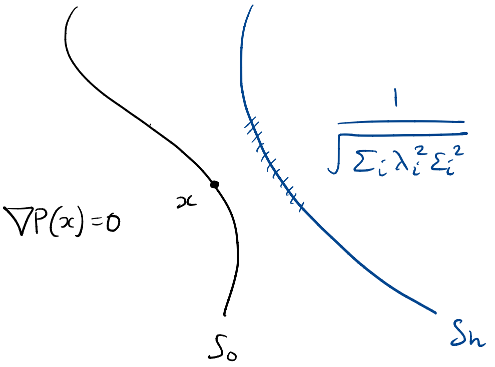
3. Density of states and curvature

We return briefly to the comments on flatness from p. ⑤. Let x be a critical point of P. Then for a vector $\varepsilon \in \mathbb{R}^n$ of small real numbers

$$\nabla P(x+\varepsilon) \approx \nabla P(x) + \left(\sum_{j=1}^{n} \frac{\partial}{\partial x_j} \frac{\partial P}{\partial x_i}(x)\, \varepsilon_j \right)_{i=1}^{n} = H(x)\,\varepsilon \qquad (8.1)$$

where $H(x) = \left(\frac{\partial^2 P}{\partial x_i \partial x_j}(x) \right)_{1 \le i,j \le n}$ is the Hessian matrix of P at x. If the matrix $H(x)$ is invertible then by the Morse Lemma we may choose local coordinates such that, locally, $P = \sum_{i=1}^{n} \lambda_i x_i$ and so $H(x) = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$ and $\| \nabla P(x+\varepsilon) \|^2 = \sum_{i=1}^{n} \lambda_i^2 \varepsilon_i^2$.

Recall that the eigenvalues of the Hessian measure curvature (to refresh your memory consider $f = a_1 x_1^2 + \cdots + a_n x_n^2$ and that the osculating circle in the x_i direction at the origin has radius $R_i = \frac{1}{2a_i}$ so the curvature in the x_i direction is $2a_i$) so that $\sum_i \lambda_i^2$ is a measure of total curvature.



(8.2)

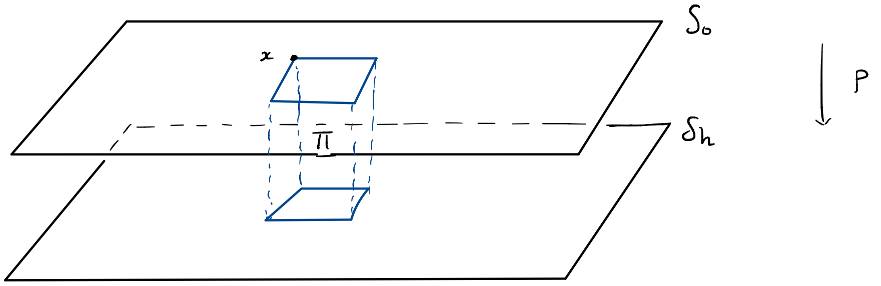
If t is a critical value of P, and $S_0 = P^{-1}(t)$, $S_h = P^{-1}(t+h)$ as above, then for h small the integral $\int_{S_h} \frac{\varphi}{\|\nabla P\|} dS$ is dominated by parts of S_h close to critical points of P on S_0. If the critical point is Morse (i.e. H is nondegenerate) then $\frac{1}{\|\nabla P\|} \sim 1/\sqrt{\sum_i \lambda_i^2 \varepsilon_i^2}$ on these regions, hence the suggestion that $\int_{S_h} \frac{\varphi}{\|\nabla P\|} dS$ for small h is a measure of flatness.

Note The λ_i do not affect the exponent in (6.1) with $P = \sum_i \lambda_i x_i^2$ but they affect the coefficient. So in the asymptotic expansion (7.2) we care about both the exponent and coefficient.

4. The definition of density of states

Before proceeding we have to face up to some flaws in this definition. Firstly, it appears to depend on a choice of coordinates: firstly to define ∇P and secondly to choose a volume form dS on S And of course at the moment the justification hinges on (4.1), an approximation. Following [GS1, Ch. 3 §1.2] let us now give the "correct" derivation.

We let P be as above and $t \in \mathbb{R}$ a regular value so that $S_h = P^{-1}(t+h)$ is a regular submanifold for h small. Let us revisit the diagram (5.1), this time depicting S_0, S_h as surfaces. Consider the small volume Π depicted above, with vertex $x \in S_0$



(9.1)

Suppose chosen some local coordinates $u_2, \ldots, u_n$ on S_0 at x so that $P, u_2, \ldots, u_n$ give local coordinates for x in $\mathbb{R}^n$. Assume the coordinates are chosen so that the "rectangle" $[0,1]^{n-1}$ of sidelength 1 with corner x fits in the coordinate chart. Then Π has volume h. In more sophisticated language the volume element is the n form $dP du_2 \cdots du_n$. We can relate this to a standard volume form $dv = dx_1 \cdots dx_n$ on $\mathbb{R}^n$ via (with $u_n = P$)

$$dv = D\left(\begin{smallmatrix} x \\ u \end{smallmatrix}\right) dP du_2 \cdots du_n \qquad (9.2)$$

where $D\left(\begin{smallmatrix} x \\ u \end{smallmatrix}\right)$ is the Jacobian. This shows that the $n-1$ form $D\left(\begin{smallmatrix} x \\ u \end{smallmatrix}\right) du_2 \cdots du_n$ is (up to a term whose product with dP vanishes) <u>independent of the choice of coordinates</u> on S_0. One shows (see [GS1]) that this ambiguity vanishes on restriction to S_0, so in fact

$$\omega = \left[D\left(\begin{smallmatrix} u \\ x \end{smallmatrix}\right)\right]^{-1} du_2 \cdots du_n \qquad (9.3)$$

Is an intrinsic $n-1$ form on S_0 depending only on a choice of volume form dv for $\mathbb{R}^n$ and P itself. This form ω, which we think of as "dv/dP" is the rate of change of the volume Π with h (the u_i are local coordinates, but since ω is independent of this choice it is defined globally on S_0). The functional which takes a compactly supported smooth function φ, restricts it to S_0 and integrates it against ω (i.e. the formula (5.3)) is a distribution.

Our references for distributions are [FJ], [GS1]. Given an open set $U \subseteq \mathbb{R}^n$ we write $C_c^\infty(U)$ for the topological $\mathbb{R}$-vectorspace of "test functions" that is, smooth functions $f : U \longrightarrow \mathbb{R}$ with compact support. A distribution (or generalised function) on U is a continuous linear map $\eta : C_c^\infty(U) \longrightarrow \mathbb{R}$. Note that $C_c^\infty(U)$ is denoted K in [GS1].

Def^n Let t be a regular value of a smooth function P. The distribution $\delta(t-P)$ is given by restricting φ to S_0 and integrating against the $(n-1)$ form ω

$$\varphi \longmapsto \int_{S_0} \varphi\,\omega \qquad\qquad \varphi \in C_c^\infty(U),\ U \subseteq \mathbb{R}^n \text{ open} \qquad (10.1)$$

Up to normalisation by the volume this means that the density of states $D(t)$ with function φ in (5.3) is the value of the Schwartz distribution $\delta(t-P)$ on φ.

For standard reasons (see e.g. [W, Remark 4.1 p. 110]) it suffices to define $\delta(t-P)$ locally, e.g. on a coordinate patch where (9.3) applies. If we assume $\frac{\partial P}{\partial x_1}$ is nonzero at $x \in S_0$ and that P, $u_2 = x_2, \dots, u_n = x_n$ give local coordinates at x (as in [W, p. 111]) then

$$\omega = \left[D\left(\tfrac{u}{x}\right)\right]^{-1} du_2 \cdots du_n = \frac{dx_2 \cdots dx_n}{|\partial f/\partial x_1|} \qquad (10.2)$$

and hence $\delta(t-P)$ sends (cf. the definition in [W, §4.2])

$$\varphi \longmapsto \int \frac{\varphi(t, x_2, \dots, x_n)}{|\partial f/\partial x_1(t, x_2, \dots, x_n)|}\, dx_2 \cdots dx_n \qquad (10.3)$$

References

[W] S. Watanabe "Algebraic geometry and statistical learning theory" 2009.

[K] C. Kittel "Introduction to Solid State Physics : Seventh ed. " John Wiley & Sons 1996.

[GS1] I. M. Gel'fand, G. E. Shilov "Generalized functions Volume I : Properties and operations" Academic Press 1964.

[YIF] N. F. Q. Yuan, H. Isobe, L. Fu "Magic of high-order van Hove singularity" Nature Communications 2019.

[GS2] I. M. Gel'fand, G. E. Shilov "Generalized functions Volume II : Spaces of fundamental and generalised functions" Academic Press 1968.

[FJ] F. G. Friedlander, M. Joshi "Introduction to the theory of distributions" Cambridge University Press 1998.

[B] W. M. Boothby "An introduction to differentiable manifolds and Riemannian geometry" Revised Second Edition, Academic Press 2003.

[GG] M. Golubitsky, V. Guillemin "Stable mappings and their singularities" Springer-Verlag, 1973.