BULLETIN (New Series) OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 33, Number 1, January 1996

Dirichlet forms & symmetric Markov processes, by M. Fukushima, Y. Oshima, and M. Takeda, deGruyter Studies in Math., vol. 19, Walter de Gruyter, Berlin and Hawthorne, NY, 1994, viii + 392 pp., \$79.95, ISBN 3-11-011626-X

Dirichlet forms, by Zhi-Ming Ma and Michael Rockner, Universitext, Springer-Verlag, Berlin and New York, 1992, vi + 209 pp., \$39.00, ISBN 3-540-558480-9

Sometimes the value of a theory is that it allows progess even in the absence of understanding, and, in many of its applications, that is precisely what the theory of Dirichlet forms does allow. Here is a relatively easy example.

Consider the standard Laplace operator $\Delta = \partial^2_{x_1} + \partial^2_{x_2}$ on the spaces $C_c^\infty(\mathbb{D})$ of smooth, compactly supported, $\mathbb{R}$-valued functions on the open unit disk $\mathbb{D} = \{x \in \mathbb{R}^2 : |x| < 1\}$. After a simple integration by parts, one sees that

$$(1) \qquad -\int_{\mathbb{D}} f \Delta g \, dx = \int_{\mathbb{D}} \nabla f \cdot \nabla g \, dx, \quad f, g \in C_c^\infty(\mathbb{D}).$$

In particular, $-\Delta$ on $C_c^\infty(\mathbb{D})$ is a symmetric, non-negative definite operator in the (real) Hilbert space $L^2(\mathbb{D})$. On the other hand, Δ on $C_c^\infty(\mathbb{D})$ is certainly not self-adjoint: $f \in L^2(\mathbb{D})$ is in the domain of its adjoint Δ^* if and only if Δf, computed in the sense of distributions, admits a square integrable representative on $\mathbb{D}$. Hence, before one can apply the spectral theorem to the analysis of Δ on $C_c^\infty(\mathbb{D})$, one has to produce a self-adjoint extension $\overline{\Delta}$, which will lie somewhere between Δ (which is too small) and Δ^* (which is too large). In fact, one has to be careful about *which* self-adjoint extension one wants: there are lots of them and the differences between them are dramatic. To wit, there is the *Dirichlet extension* $\overline{\Delta}_0$ whose domain is described intuitively as the set of $f \in \mathrm{Dom}(\Delta^*)$ *which vanish at the boundary of* $\mathbb{D}$. Alternatively, one can look at the *Neumann extension* $\overline{\Delta}_1$ whose domain may be described intuitively as those $f \in \mathrm{Dom}(\Delta^*)$ *with vanishing normal derivative at the boundary*. And there are plenty of others.

Now suppose that one were to get ambitious and try to classify all the possible self-adjoint extensions of Δ. (So far, we have imprecise descriptions of two candidates.) One would probably begin by mistakenly thinking that the extensions can be linearly ordered via their domains. However, they cannot be. Indeed, it is an easy exercise to show that one self-adjoint operator can extend a second only if the extension is trivial (i.e., the two operators coincide). For example, in the present situation, the constant function $\mathbf{1}$ is in the domain of $\overline{\Delta}_1$ but not of $\overline{\Delta}_0$, whereas

$$x_1^2 + x_2^2 - 1 \in \mathrm{Dom}(\overline{\Delta}_0) \cap \mathrm{Dom}(\overline{\Delta}_1)\complement.$$

Nonetheless, some sort of linear ordering ought to exist: one cannot help harboring the suspicion that $\overline{\Delta}_1$ is, in fact, *larger* than $\overline{\Delta}_0$. But to rationalize this suspicion, one must adopt a more subtle test of size. For instance, one comes closer to success when one bases an ordering on the supply of harmonic functions (i.e., $f \in \mathrm{Dom}(\Delta^*)$ with $\Delta^* f = \mathbf{0}$) in the domains of the various self-adjoint extensions. To wit, $\mathrm{Dom}(\overline{\Delta}_0)$ contains no harmonic functions other than $\mathbf{0}$, whereas $\mathbf{1}$ is a non-zero element of $\mathrm{Dom}(\overline{\Delta}_1)$; and, more generally, one can linearly order all the self-adjoint extensions of Δ by this procedure. But obviously, whatever route one adopts, rigorous statements are going to rely on a precise description of the domains involved, and that is precisely what our discussion has not yielded as yet.

It seems that Kurt Friedrichs was the first to realize that the shortest distance to the heart of a symmetric, non-negative definite operator is through its quadratic form. For example, instead of attempting a frontal attack on the self-adjoint extensions of Δ, it is smarter to concentrate on possible extensions of the associated quadratic form as it appears on the right-hand side of (1). That is, given a self-adjoint extension $\overline{\Delta}$ of Δ, note that $-\overline{\Delta}$ must be non-negative definite, and apply the spectral theorem to find a resolution of the identity $\{E_\lambda^{(\overline{\Delta})} : \lambda \in [0,\infty)\}$ by orthogonal projections so that

$$-\overline{\Delta} = \int_{[0,\infty)} \lambda \, dE_\lambda^{(\overline{\Delta})}.$$

Then the quadratic form

$$(f,g) \in \mathrm{Dom}(\overline{\Delta})^2 \to -(f,\overline{\Delta}g)_{L^2(\mathbb{D})} = \int_{[0,\infty)} \lambda d(f, E_\lambda^{(\overline{\Delta})} g)_{L^2(\mathbb{D})}$$

is an extension of

$$(f,g) \in C_c^\infty(\mathbb{D})^2 \mapsto \mathcal{E}^{(\Delta)}(f,g) \equiv \int_{\mathbb{D}} \nabla f \cdot \nabla g \, dx.$$

Moreover, if one sets

$$\mathcal{E}^{(\overline{\Delta})}(f,f) = \int_{[0,\infty)} \lambda d(f, E_\lambda^{(\overline{\Delta})} f)_{L^2(\mathbb{D})} \in [0,\infty] \text{ for any } f \in L^2(\mathbb{D}), \tag{2}$$

then

$$\mathrm{Dom}(\mathcal{E}^{(\overline{\Delta})}) \equiv \{f \in L^2(\mathbb{D}) : \mathcal{E}^{(\overline{\Delta})}(f,f) < \infty\}$$

is precisely the domain of

$$\sqrt{-\overline{\Delta}} = \int_{[0,\infty)} \lambda^{\frac{1}{2}} dE_\lambda^{(\overline{\Delta})},$$

the square root of $-\overline{\Delta}$. Finally, because $\mathcal{E}^{(\overline{\Delta})}(f,f) \geq 0$, it is easy to see that

$$\begin{aligned} (f,g) \in \mathrm{Dom}(\mathcal{E}^{(\overline{\Delta})})^2 \mapsto \mathcal{E}^{(\overline{\Delta})}(f,g) \\ \equiv \frac{1}{4}[\mathcal{E}^{(\overline{\Delta})}(f+g,f+g) - \mathcal{E}^{(\overline{\Delta})}(f-g,f-g)] \in \mathbb{R} \end{aligned}$$

is well defined and satisfies

$$\mathcal{E}^{(\overline{\Delta})}(f,g) \leq \sqrt{\mathcal{E}^{(\overline{\Delta})}(f,f)\mathcal{E}^{(\overline{\Delta})}(g,g)}.$$

Next, suppose that one runs the preceding line of reasoning in the opposite direction. That is, following Friedrichs, think about the possibility of getting at $\overline{\Delta}$ by way of $\mathcal{E}^{(\overline{\Delta})}$. In other words, look at extensions $\mathcal{E}$ of $\mathcal{E}^{(\Delta)}$ which are the $\mathcal{E}^{(\overline{\Delta})}$ for some self-adjoint extension $\overline{\Delta}$. To this end, one must recognize that the essential feature of such an extension $\mathcal{E}$ will be the fact it is *closed* in the sense that if $\{f_n\}_1^\infty$ is a sequence from

$$\mathrm{Dom}(\mathcal{E}) \equiv \{f \in L^2(\mathbb{D}) : \mathcal{E}(f,f) < \infty\}$$

which converges in $L^2(\mu)$ and if $\mathcal{E}(f_n, f_n)$ is bounded, then its limit f must be an element of $\mathrm{Dom}(\mathcal{E})$ and

$$\mathcal{E}(f,f) \leq \varliminf_{n\to\infty} \mathcal{E}(f_n, f_n).$$

Hence, it is reasonable to see what happens if one *closes* $\mathcal{E}^{(\Delta)}$ itself, and this is precisely what Friedrichs did. Abstractly, the steps are as follows: First, one completes $\mathrm{Dom}(\mathcal{E}^{(\Delta)})$ (alias $C_c^\infty(\mathbb{D})$) with respect to the metric determined by the Hilbert norm

$$\sqrt{\|f\|^2_{L^2(\mathbb{D})} + \mathcal{E}^{(\Delta)}(f,f)}.$$

In general, a completion procedure leads to abstract nonsense. However (and this is Friedrichs's key observation), here one can identify the completion as a linear subspace $\mathrm{Dom}(\mathcal{E}_0^{(\Delta)})$ of $L^2(\mathbb{D})$. Next, one sets $\mathrm{Dom}(\Delta_0) = \mathrm{Dom}(\mathcal{E}_0^{(\Delta)}) \cap \mathrm{Dom}(\Delta^*)$ and defines $\Delta_0 = \Delta^* \upharpoonright \mathrm{Dom}(\Delta_0)$. Finally, one verifies that Δ_0 is a self-adjoint extension of Δ and that $\mathcal{E}_0^{(\Delta)} = \mathcal{E}^{(\Delta_0)}$. In fact, this is exactly the Dirichlet extension alluded to earlier.

The scheme just outlined is called *Friedrichs's extension procedure*, and the resulting extension is called the *Friedrichs extension*. It works in a completely abstract setting: all one needs is a symmetric, non-negative definite operator on a Hilbert space. (Obviously, it suffices to start with a semi-bounded operator, since one can then reduce to the non-negative definite case by trivial manipulation.) In addition, M. G. Kreĭn developed a beautiful theory in which he showed that all the self-adjoint extensions of the operator can be linearly ordered via their associated quadratic forms, on which scale Friedrichs's extension is *minimal* and there is a *maximal* extension, called the *Kreĭn extension*. (Kreĭn's extension is obtained by appending as many *harmonic* functions as possible to the domain of the quadratic form.) In fact, one does not really need to start with an operator at all. What is required is a densely defined, non-negative, quadratic form which is *closable* in the sense that the completion process outlined above leads to a set which can be identified as a subspace of the Hilbert space. (One can interpret Friedrich's key observation as saying that the quadratic form coming from a symmetric, non-negative definite operator is always closable.)

As the preceding advertisement about the Friedrichs-Kreĭn theory intimates, their theory is too general to capture all the properties of the Laplace operator, as opposed, say, to Δ^2. Indeed, as long as the discussion stays at the pure Hilbert space level, the distinction between $-\Delta$ and Δ^2 is not all that great. Nonetheless, when the discussion takes into account properties which are peculiar to $L^2(\mathbb{D})$, as opposed to a general Hilbert space, one may discover that there is a profound distinction: the *minimum principal* holds for Δ, but it does not for Δ^2. In other words, if $f \in C_c^\infty(\mathbb{D})$ is non-negative and vanishes at some $x \in \mathbb{D}$, then $\Delta f(x) \geq 0$, but the sign of $\Delta^2 f(x)$ is ambiguous. The profundity of this apparently arcane piece of information may not be obvious to the casual observer. However, probabilists make their living on it and therefore attach to it considerable weight. What it says is that the semi-group of operators

$$e^{t\overline{\Delta}_0} \equiv \int_{[0,\infty)} e^{-\lambda t}\, dE_\lambda^{(\overline{\Delta}_0)}, t \in [0,\infty)$$

preserves the space of non-negative functions. An operator P which preserves non-negativity and leaves $\mathbf{1}$ invariant is called a *Markov operator*; and when invariance of $\mathbf{1}$ is replaced by $P\mathbf{1} \leq \mathbf{1}$, P is said to be *sub-Markov.* In particular, one can hope that (with luck) it is possible to represent a (sub-)Markov P in terms of a (sub-)probability measure valued kernel:

$$Pf = \int f(y)P(\cdot, dy)$$

where, for each $x \in \mathbb{D}$, $P(x, \cdot)$ is a non-negative measure with total mass (less than or) equal to 1. The kernel $P(x, \cdot)$ is called the transition (sub-)probability associated with P and provides the initial strand in the intricate web which probabilists weave around operators which satisfy the minimum principle.

To get a feeling for the connection between the minimum principle and non-negativity preservation, suppose that A generates the semigroup $\{e^{tA} : t > 0\}$ of operators on the space $C_b(X)$ of bounded, continuous functions on some X. If the semigroup preserves non-negativity and if f is a non-negative function which vanishes at x, then

$$Af(x) = \lim_{t \searrow 0} t^{-1}([e^{tA}f](x) - f(x)) \geq 0.$$

That is, A satisfies the minimum principle. The argument in the other direction is a little less transparent, but the idea is clear. Namely, suppose that f is non-negative and that $e^{tA}f$ fails to remain non-negative forever. Let $t_0 = \inf\{t > 0 : e^{tA}f \not\geq 0\}$. Then, presumably, $e^{t_0A}f$ is non-negative and there is an x_0 such that $0 = [e^{t_0A}f](x_0) > [e^{tA}f](x_0)$ for $t > t_0$ which are arbitrarily close to t_0. Hence, if A satisfies the minimum principle, one can hope to extract a contradiction from

$$\lim_{t \searrow t_0} (t - t_0)^{-1}[e^{tA}f - e^{t_0A}f](x_0) = [Ae^{t_0A}f](x_0) \geq 0$$
$$\text{but } [e^{tA}f - e^{t_0A}f](x_0) < 0$$

for a sequence of t's which decrease to t_0. Of course, this last argument is riddled with holes, but its does contain a grain of truth.

Returning to the earlier discussion about the advantages of dealing with operators via their quadratic forms, one now should ask how the minimum principle is manifested in the quadratic form of an operator for which it holds. To appreciate the answer, it may be helpful to know that an operator A which satisfies the minimum principle can be *no more than second order* (in the sense of differential or, more precisely, pseudo-differential operators). Thus, the square root $\sqrt{-A}$ of such an operator can be no more than first order. In particular, because the quadratic form is really $\|\sqrt{-A}f\|^2$, one should hope to see the minimum principle reflected in some first order type property of the quadratic form. Actually, it is not very hard to make all this more precise. To this end, suppose that A is a non-positive definite, self-adjoint operator on $L^2(\mu)$ which generates a semigroup $\{e^{tA} : t > 0\}$ of self-adjoint, Markov operators. Further, assume that, for each $t > 0$, there exists a transition probability $P(t, x, \cdot)$ for which

$$[e^{tA}f](x) = \int f(y)P(t, x, dy). \tag{3}$$

Because e^{tA} is self-adjoint, the measure $P(t,x,dy)\mu(dx)$ is symmetric, from which it is an easy step to

$$\begin{aligned}\int_{[0,\infty}(1-e^{-\lambda t})d(f,E_\lambda^{(A)}f)_{L^2(\mu)} &= (f,f-e^{tA}f)_{L^2(\mu)} \\ &= \frac{1}{2}\int\int(f(y)-f(x))^2P(t,x,dy)\mu(dx).\end{aligned} \tag{4}$$

Hence, after dividing through by $t>0$ and letting $t\searrow 0$, one finds that

$$\begin{aligned}\mathcal{E}^{(A)}(f,f) &\equiv \int_{[0,\infty)}\lambda d(f,E_\lambda^{(A)}f)_{L^2(\mu)} \\ &= \lim_{t\searrow 0}\frac{1}{2t}\int\int(f(y)-f(x))^2P(t,x,dy)\mu(dx).\end{aligned} \tag{5}$$

In the sub-Markov case, the right-hand side of (4) has to be replaced by

$$\frac{1}{2}\int\int(f(y)-f(x))^2P(t,x,dy)\mu(dx)+(\mathbf{1}-e^{tA}\mathbf{1})f(x)^2$$

and so

$$\begin{aligned}\mathcal{E}^{(A)}(f,f) &\equiv \int_{[0,\infty)}\lambda d(f,E_\lambda^{(A)}f)_{L^2(\mu)} \\ &= \lim_{t\searrow 0}\frac{1}{2t}\int\int(f(y)-f(x))^2P(t,x,dy)\mu(dx)+V(x)f(x)^2\end{aligned}$$

where

$$V\equiv\lim_{t\searrow 0}t^{-1}(\mathbf{1}-e^{tA}\mathbf{1}).$$

The message in (4) and (5) should be clear: the quadratic form $\mathcal{E}=\mathcal{E}^{(A)}(f,f)$ is the norm squared of some sort of generalized *gradient*. In particular, one has that

$$\begin{aligned}&g\in\mathrm{Dom}(\mathcal{E}), f\in L^2(\mu), \text{ and } |f(y)-f(x)|\le|g(y)-g(x)| \\ &\Rightarrow f\in\mathrm{Dom}(\mathcal{E}) \text{ and } \mathcal{E}(f,f)\le\mathcal{E}(g,g).\end{aligned} \tag{6}$$

A closable, densely defined quadratic form $\mathcal{E}$ on $L^2(\mu)$ whose closure satisfies (6) is said to be a *Dirichlet form*; and a fundamental observation, due to Beurling and Deny, is that the semigroup generated by the Friedrichs extension $\overline{A}$ of a symmetric, non-positive definite A on $L^2(\mu)$ is Markov precisely when the quadratic form

$$f\in\mathrm{Dom}(A)\mapsto\mathcal{E}^{(A)}(f,f)\equiv-(f,Af)_{L^2(\mu)}$$

is a Dirichlet form.

Because the semigroup associated with a Dirichlet form enjoys the best that both Hilbert space and probability theory have to offer, it is not surprising that the operators in such semigroups are more tractable than operators which are either self-adjoint but not Markov or Markov but not self-adjoint. In particular, these semigroups are surprisingly amenable to comparisons based on their Dirichlet forms. For instance, although they did not do so themselves, one can view the magnificent theory of DeGiorgi, Moser, and Nash as an example of the sort of results to which such comparisons lead. (See [D] for an exposition of the D-M-N theory from this standpoint.) However, neither of the books under review emphasizes such applications. Instead, both the book by Fukushima, Oshima, and Takeda and the one by Ma and Röckner are written from the probabilistic perspective. Thus, for these authors, the semi-group associated with a Dirichlet form is only step number one in a program which leads ultimately to the construction of a Markov process for which the semigroup is the transition mechanism. As anyone familiar with the tradition of abstract Markov process theory would suspect, the details can be a little tedious: how much should the average mortal care whether a construction leads to a **Hunt process** or a **right process**?[1] However, as distinguished from the usual presentation of the abstract theory, where one makes the entirely unjustified assumption that one starts on intimate terms with the resolvent operators, the development here is founded on the much more reasonable assumption that what one starts with is information about the Dirichlet form. The advantage is that, in order to apply this theory, you do not need to have basically solved all the hard problems ahead of time. Of course, one pays for this in the end, because the detail in which one can work is limited by the resolving power of the Dirichlet form: that is, everything must be done modulo a set of *capacity* zero. If you want greater detail, then you have to work harder.

Finally, how do these books differ from one another? The one by Fukushima et al. is an updating and expansion of Fukushima's 1980 book [F]. The theory is developed here with an eye to finite-dimensional applications, of which there is a host of beautiful examples provided. The book by Ma and Röckner aims at generality and is a tour de force in the expository tradition of abstract functional analysis. As a consequence, it makes easier reading. In addition, it has the advantage that it prepares its readers for applications in the infinite dimensional realm. It has the disadvantage that its examples are not as rich.

References

[D] E. B. Davies, *Heat Kernels and Spectral Theory*, Cambridge Tract #92, Cambridge Univ. Press, Cambridge, UK and NY, 1989. MR **92a:**35035

[F] M. Fukushima, *Dirichlet Forms and Markov Processes*, North Holland, 1980. MR **81f:**60105

Daniel W. Stroock
Massachusetts Institute of Technology
E-mail address: dws@math.mit.edu

[1] The adjective *Hunt* refers to Gilbert Hunt, who wrote the papers which brought the abstract theory of Markov processes into the modern era. The term *right* is more ambiguous. My impression is that it is applied to certain processes in much the same spirit that it used to be applied to certain whales.