



# Spectral Pollution and How to Avoid It

(With Applications to Dirac and Periodic Schrödinger Operators)

Mathieu LEWIN

*CNRS and Laboratoire de Mathématiques (CNRS UMR 8088), Université de Cergy-Pontoise, 2, avenue Adolphe Chauvin, 95 302 Cergy-Pontoise Cedex - France.  
Email: Mathieu.Lewin@math.cnrs.fr*

Éric SÉRÉ

*Ceremade (CNRS UMR 7534), Université Paris-Dauphine, Place du Maréchal de Lattre de Tassigny, 75775 Paris Cedex 16 - France.  
Email: sere@ceremade.dauphine.fr*

December 11, 2008

This paper, devoted to the study of spectral pollution, contains both abstract results and applications to some self-adjoint operators with a gap in their essential spectrum occurring in Quantum Mechanics.

First we consider Galerkin basis which respect the decomposition of the ambient Hilbert space into a direct sum  $\mathfrak{H} = P\mathfrak{H} \oplus (1 - P)\mathfrak{H}$ , given by a fixed orthogonal projector  $P$ , and we localize the polluted spectrum exactly. This is followed by applications to periodic Schrödinger operators (pollution is absent in a Wannier-type basis), and to Dirac operator (several natural decompositions are considered).

In the second part, we add the constraint that within the Galerkin basis there is a certain relation between vectors in  $P\mathfrak{H}$  and vectors in  $(1 - P)\mathfrak{H}$ . Abstract results are proved and applied to several practical methods like the famous *kinetic balance* of relativistic Quantum Mechanics.

© 2008 by the authors. This paper may be reproduced, in its entirety, for non-commercial purposes.

## Contents

<b>Introduction</b>	<b>2</b>
<b>1 Spectral pollution</b>	<b>4</b>
<b>2 Pollution associated with a splitting of <math>\mathfrak{H}</math></b>	<b>7</b>
2.1 A general result . . . . .	8
2.2 A simple criterion of no pollution . . . . .	12
2.3 Applications . . . . .	14
2.3.1 Periodic Schrödinger operators in Wannier basis . . . . .	14
2.3.2 Dirac operators in upper/lower spinor basis . . . . .	16
2.3.3 Dirac operators in dual basis . . . . .	19
2.3.4 Dirac operators in free basis . . . . .	21
<b>3 Balanced basis</b>	<b>25</b>
3.1 General results . . . . .	25
3.1.1 Sufficient conditions . . . . .	25

3.1.2	Necessary conditions . . . . .	27
3.2	Application to Dirac operator . . . . .	29
3.2.1	Kinetic Balance . . . . .	30
3.2.2	Atomic Balance . . . . .	33
3.2.3	Dual Kinetic Balance . . . . .	37

## Introduction

This paper is devoted to the study of *spectral pollution*. This phenomenon of high interest occurs when one approximates the spectrum of a (bounded or unbounded) self-adjoint operator  $A$  on an infinite-dimensional Hilbert space  $\mathfrak{H}$ , using a sequence of finite-dimensional spaces. Consider for instance a sequence  $\{V_n\}$  of subspaces of the domain  $D(A)$  of  $A$  such that  $V_n \subset V_{n+1}$  and  $P_{V_n} \rightarrow 1$  strongly (we denote by  $P_{V_n}$  the orthogonal projector on  $V_n$ ). Define the  $n \times n$  matrices  $A_n := P_{V_n} A P_{V_n}$ . It is well-known that such a Galerkin method may in general lead to *spurious eigenvalues*, i.e. numbers  $\lambda \in \mathbb{R}$  which are limiting points of eigenvalues of  $A_n$  but do not belong to  $\sigma(A)$ . This phenomenon is known to occur in gaps of the essential spectrum of  $A$  only.

Spectral pollution is an important issue which arises in many different practical situations. It is encountered when approximating the spectrum of perturbations of periodic Schrödinger operators [4] or Strum-Liouville operators [35, 36, 1]. It is a very well reported difficulty in Quantum Chemistry and Physics in particular regarding relativistic computations [13, 18, 22, 34, 14, 27, 32]. It also appears in elasticity, electromagnetism and hydrodynamics; see, e.g. the references in [2]. Eventually, it has raised as well a huge interest in the mathematical community, see, e.g., [23, 9, 4, 21, 10, 28, 29].

In this article we will study spectral pollution from a rather new perspective. Although many works focus on how to determine if an approximate eigenvalue is spurious or not (see, e.g., the rather successful second-order projection method [23, 4]), we will on the contrary concentrate on finding conditions on the sequence  $\{V_n\}$  which ensure that there will not be any pollution at all, in a given interval of the real line.

Our work contains two rather different aspects. On the one hand we will establish some theoretical results for abstract self-adjoint operators: we characterize exactly (or partially) the polluted spectrum under some specific assumptions on the approximation scheme as will be explained below. On the other hand we apply these results to two important cases of Quantum Physics: perturbations of periodic Schrödinger operators and Dirac operators. For Dirac operators, we will show in particular that some very well-known methods used by Chemists or Physicists indeed allow to partially avoid spurious eigenvalues in certain situations, or at the contrary that they are theoretically of no effect in other cases.

Let us now summarize our results with some more details.

Our approach consists in adding some assumptions on the approximating scheme. We start by considering in Section 2 a fixed orthogonal projector  $P$  acting on the ambient Hilbert space  $\mathfrak{H}$  and we define  $P$ -spurious eigenvalues  $\lambda$  as limiting points obtained by a Galerkin-type procedure, in a basis which respects the decomposition associated with  $P$ . This means  $\lambda = \lim_{n \rightarrow \infty} \lambda_n$  with  $\lambda \notin \sigma(A)$  and  $\lambda_n \in \sigma(P_{V_n} A P_{V_n})$ , where  $V_n = V_n^+ \oplus V_n^-$  for some  $V_n^+ \subset \mathfrak{H}^+ := P\mathfrak{H}$  and  $V_n^- \subset \mathfrak{H}^- := (1 - P)\mathfrak{H}$ . We show that, contrarily to the general case and depending on  $P$ , there might exist an interval in  $\mathbb{R}$  in which there is never any pollution occurring. More precisely, we exactly determine the location of the polluted

Imposed splitting of Hilbert space	External potential $V$	Spurious spectrum in the gap $(-1, 1)$
<i>none</i>	any	$(-1, 1)$
<i>upper/lower spinors</i> $\begin{pmatrix} \varphi_n \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ \chi_n \end{pmatrix}$	$V = 0$	$\emptyset$
	$V$ bounded	$(-1, -1 + \sup(V)] \cup [1 + \inf(V), 1)$
	unbounded (ex: Coulomb)	$(-1, 1)$
<i>dual decomposition [32]</i> $\begin{pmatrix} \varphi_n \\ \epsilon \sigma \cdot p \varphi_n \end{pmatrix}, \begin{pmatrix} -\epsilon \sigma \cdot p \chi_n \\ \chi_n \end{pmatrix}$ $0 < \epsilon \leq 1$	$V = 0$	$\emptyset$
	$V$ bounded	$(-1, -2/\epsilon + 1 + \sup(V)] \cup [2/\epsilon - 1 + \inf(V), 1)$
	unbounded (ex: Coulomb)	$(-1, 1)$
<i>free decomposition</i> $P_+^0 \Psi_n, P_-^0 \Psi'_n$	any	$\emptyset$

Table 1. Summary of our results from Section 2.3 for the Dirac operator  $D^0 + V$ , when a splitting is imposed on the Hilbert space  $L^2(\mathbb{R}^3, \mathbb{C}^4)$ .

spectrum in Section 2.1 and we use this in Section 2.2 to derive a simple criterion on  $P$ , allowing to completely avoid the appearance of spurious eigenvalues in a gap of the essential spectrum of  $A$ .

Then we apply our general result to several practical situations in Section 2.3. We in particular show that the usual decomposition into upper and lower spinors *a priori* always leads to pollution for Dirac operators. We also study another decomposition of the ambient Hilbert space which was proposed by Shabaev et al [32] and we prove that the set which is free from spectral pollution is larger than the one obtained from the simple decomposition into upper and lower spinors. Eventually, we prove that choosing the decomposition given by the spectral projectors of the free Dirac operator is completely free of pollution. For the convenience of the reader, we have summarized all these results in Table 1.

As another application we consider in Section 2.3.1 the case of a periodic Schrödinger operator which is perturbed by a potential which vanishes at infinity. We prove again that choosing a decomposition associated with the unperturbed (periodic) Hamiltonian allows to avoid spectral pollution, as was already demonstrated numerically in [6] using Wannier functions.

In Section 3, we come back to the theory of a general operator  $A$  and we study another method inspired by the ones used in quantum Physics and Chemistry. Namely, additionally to a splitting as explained before, we add the requirement that there is a specific relation (named *balance condition*) between the vectors of  $\mathfrak{H}^-$  and that of  $\mathfrak{H}^+$ . This amounts to choosing a fixed operator  $L : \mathfrak{H}^+ \rightarrow \mathfrak{H}^-$  and taking as approximation spaces  $V_n = V_n^+ \oplus LV_n^+$ . We do not completely characterize theoretically the possible spurious eigenvalues for this kind of methods but we give necessary and sufficient conditions which are enough to fully understand the case of the Dirac operator. In Quantum Chemistry and Physics the main method is the so-called *kinetic balance* which consists in choosing  $L = \sigma(-i\nabla)$  and the decomposition into upper and lower spinors. We show in Section 3.2.1 that this method allows to avoid spectral pollution in the upper part of the spectrum only for bounded potentials and that it does not help for unbounded functions like the Coulomb potential. We prove in Section 3.2.2 that the so-called (more complicated) *atomic balance* indeed allows to solve this problem also for

Balance condition	External potential $V$	Spurious spectrum in the gap $(-1, 1)$
<i>kinetic balance</i> $\begin{pmatrix} \varphi_n \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ \sigma \cdot p \varphi_n \end{pmatrix}$	$V$ bounded with $-1 + \sup(V) < 1 + \inf(V)$	$(-1, -1 + \sup(V))$
	$V(x) = -\frac{\kappa}{ x },$ $0 < \kappa < \sqrt{3}/2$	$(-1, 1)$
<i>atomic balance</i> $\begin{pmatrix} \varphi_n \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ \frac{1}{2-V} \sigma \cdot p \varphi_n \end{pmatrix}$	$V$ such that $-\frac{\kappa}{ x } \leq V(x)$ where $0 \leq \kappa < \sqrt{3}/2,$ and $\sup(V) < 2$	$(-1, -1 + \sup(V))$
<i>dual kinetic balance</i> [32] $\begin{pmatrix} \varphi_n \\ \epsilon \sigma \cdot p \varphi_n \end{pmatrix}, \begin{pmatrix} -\epsilon \sigma \cdot p \varphi_n \\ \varphi_n \end{pmatrix}$	$V$ bounded	$(-1, -2/\epsilon + 1 + \sup(V))$ $\cup [2/\epsilon - 1 + \inf(V), 1)$
	unbounded (ex: Coulomb)	$(-1, 1)$

Table 2. Summary of our results for the Dirac operator  $D^0 + V$  when a balance is imposed between vectors of the basis.

Coulomb potentials, as was already suspected in the literature. Eventually, we show that the *dual kinetic balance* method of [32] is not better than the one which is obtained by imposing a splitting without *a priori* adding a balance condition. Our results for balanced methods for Dirac operators are summarized in Table 2.

We have tried to make our results sufficiently general that they could be applied to other situations in which there is a natural way (in the numerical sense) to split the ambient Hilbert space in a direct sum  $\mathfrak{H} = \mathfrak{H}^+ \oplus \mathfrak{H}^-$ . We hope that our results will provide some new insight on the spectral pollution issue.

*Acknowledgements.* The authors would like to thank Lyonell Boulton and Nabile Boussaid for interesting discussions and comments. The authors have been supported by the ANR project *AC-CQuaRel* of the french ministry of research.

## 1. Spectral pollution

In this first section, we recall the definition of *spectral pollution* and give some properties which will be used in the rest of the paper. Most of the material of this section is rather well-known [10, 32, 23, 9].

In the whole paper we consider a self-adjoint operator  $A$  acting on a separable Hilbert space  $\mathfrak{H}$ , with dense domain  $D(A)$ .

**Notation.** For any finite-dimensional subspace  $V \subset D(A)$ , we denote by  $P_V$  the orthogonal projector onto  $V$  and by  $A|_V$  the self-adjoint operator  $V \rightarrow V$  which is just the restriction to  $V$  of  $P_V A P_V$ .

As  $A$  is by assumption a self-adjoint operator, it is closed, i.e. the graph  $G(A) \subset D(A) \times \mathfrak{H}$  is closed. This induces a norm  $\|\cdot\|_{D(A)}$  on  $D(A)$  for which  $D(A)$  is closed. For any  $K \subset D(A)$ , we will use the notation  $\overline{K}^{D(A)}$  to denote the closure of  $K$  for the norm associated with the graph of  $A$ , in  $D(A)$ . On the other hand we simply denote by  $\overline{K}$  the closure for the norm of the ambient space  $\mathfrak{H}$ .

We use like in [23] the notation  $\hat{\sigma}_{\text{ess}}(A)$  to denote the essential spectrum of  $A$  union  $-\infty$  (and/or  $+\infty$ ) if there exists a sequence of  $\sigma(A) \ni \lambda_n \rightarrow -\infty$  (and/or  $+\infty$ ). Finally, we denote by  $\text{Conv}(X)$  the convex hull of any set  $X \subset \mathbb{R}$  and we use the convention that  $[c, d] = \emptyset$  if  $d < c$ .

**Definition 1.1 (Spurious eigenvalues).** We say that  $\lambda \in \mathbb{R}$  is a spurious eigenvalue of the operator  $A$  if there exists a sequence of finite dimensional spaces  $\{V_n\}_{n \geq 1}$  with  $V_n \subset D(A)$  and  $V_n \subset V_{n+1}$  for any  $n$ , such that

- (i)  $\overline{\bigcup_{n \geq 1} V_n}^{D(A)} = D(A)$ ;
- (ii)  $\lim_{n \rightarrow \infty} \text{dist}(\lambda, \sigma(A|_{V_n})) = 0$ ;
- (iii)  $\lambda \notin \sigma(A)$ .

We denote by  $\text{Spu}(A)$  the set of spurious eigenvalues of  $A$ .

If needed, we shall say that  $\lambda$  is a spurious eigenvalue of  $A$  with respect to  $\{V_n\}$  to further indicate a sequence  $\{V_n\}$  for which the above properties hold true. Note that (i) in Definition 1.1 implies in particular that we have  $\overline{\bigcup_{n \geq 1} V_n} = \mathfrak{H}$  since  $D(A)$  is dense in  $\mathfrak{H}$  by assumption.

**Remark 1.1.** As the matrix of  $A$  in a finite-dimensional space only involves the quadratic form associated with  $A$ , it is possible to define spurious eigenvalues by assuming only that  $V_n$  is contained in the form domain of  $A$ . Generalizing our results to quadratic forms formalism is certainly technical, although being actually useful in some cases (Finite Element Methods are usually expressed in this formalism). We shall only consider the simpler case for which  $V_n \subset D(A)$  for convenience.

**Remark 1.2.** If  $\lambda$  is a spurious eigenvalue of  $A$  with respect to  $\{V_n\}$  and if  $B - A$  is compact, then  $\lambda$  is either a spurious eigenvalue of  $B$  in  $\{V_n\}$  or  $\lambda \in \sigma_{\text{disc}}(B)$ . One may think that the same holds when  $B - A$  is only  $A$ -compact, but this is actually not true, as we shall illustrate below in Remark 2.7.

**Remark 1.3.** In this paper we concentrate our efforts on the spectral pollution issue, and we do not study how well the spectrum  $\sigma(A)$  of  $A$  is approximated by the discretized spectra  $\sigma(A|_{V_n})$ . Let us only mention that for every  $\lambda \in \sigma(A)$ , we have  $\text{dist}(\lambda, \sigma(A|_{V_n})) \rightarrow 0$  as  $n \rightarrow \infty$ , provided that  $\overline{\bigcup_{n \geq 1} V_n}^{D(A)} = D(A)$  as required in Definition 1.1.

The following lemma will be very useful in the sequel.

**Lemma 1.1 (Weyl sequences).** Assume that  $\lambda$  is a spurious eigenvalue of  $A$  in  $\{V_n\}$  as above. Then there exists a sequence  $\{x_n\}_{n \geq 1} \subset D(A)$  with  $x_n \in V_n$  for any  $n \geq 1$ , such that

- (1)  $P_{V_n}(A - \lambda)x_n \rightarrow 0$  strongly in  $\mathfrak{H}$ ;
- (2)  $\|x_n\| = 1$  for all  $n \geq 1$ ;
- (3)  $x_n \rightharpoonup 0$  weakly in  $\mathfrak{H}$ .

**Proof.** It is partly contained in [9]. Let  $\lambda \in \text{Spu}(A)$  and consider  $x_n \in V_n \setminus \{0\} \subset D(A)$  such that  $P_{V_n}(A - \lambda)x_n = 0$  with  $\lim_{n \rightarrow \infty} \lambda_n = \lambda$ . Dividing by  $\|x_n\|$  if necessary, we may assume that  $\|x_n\| = 1$  for all  $n$  in which case  $P_{V_n}(A - \lambda)x_n \rightarrow 0$  strongly. As  $\{x_n\}$  is bounded, extracting a subsequence if necessary we may assume that  $x_n \rightharpoonup x$  weakly in  $\mathfrak{H}$ .

What remains to be proven is that  $x = 0$ . Let  $y \in \cup_{m \geq 1} V_m$ . Taking  $n$  large enough we may assume that  $y \in V_n$ . Next we compute the following scalar product

$$0 = \lim_{n \rightarrow \infty} \langle P_{V_n}(A - \lambda)x_n, y \rangle = \lim_{n \rightarrow \infty} \langle x_n, (A - \lambda)y \rangle = \langle x, (A - \lambda)y \rangle.$$

As  $\cup_{m \geq 1} V_m$  is dense in  $D(A)$  for the norm of  $G(A)$ , we deduce that  $\langle x, (A - \lambda)y \rangle = 0$  for all  $y \in D(A)$ . Thus  $x \in D(A^*) = D(A)$  and it satisfies  $Ax = \lambda x$ . Hence  $x = 0$  since  $\lambda$  is not an eigenvalue of  $A$  by assumption.  $\square$

The next lemma will be useful to identify points in  $\text{Spu}(A)$ .

**Lemma 1.2.** *Assume that  $A$  is as above. Let  $(x_n^1, \dots, x_n^K)$  be an orthonormal system of  $K$  vectors in  $D(A)$  such that  $x_n^j \rightarrow 0$  for all  $j = 1..K$ . Denote by  $W_n$  the space spanned by  $x_n^1, \dots, x_n^K$ . If  $\lambda \in \mathbb{R}$  is such that  $\lim_{n \rightarrow \infty} \text{dist}(\lambda, \sigma(A|_{W_n})) = 0$ , then  $\lambda \in \text{Spu}(A) \cup \sigma(A)$ .*

**Proof.** Consider any nondecreasing sequence  $\{V_n\}$  such that  $\overline{\cup_{n \geq 1} V_n}^{D(A)} = D(A)$ . Next we introduce  $V'_1 := V_1$ ,  $m_1 = 0$  and we construct by induction a new sequence  $\{V'_n\}$  and an increasing sequence  $\{m_n\}$  as follows. Assume that  $V'_n$  and  $m_n$  are defined. As  $x_m^k \rightarrow 0$  for all  $k = 1..j$ , we have  $\lim_{m \rightarrow \infty} \langle Ay, x_m^k \rangle = 0$  for all  $y \in V'_n$  and all  $k = 1..K$ . Hence the matrix of  $A$  in  $V'_n + W_m$  becomes diagonal by blocks as  $m \rightarrow \infty$ . Therefore there exists  $m_{n+1} > m_n$  such that the matrix of  $A$  in  $V'_{n+1} := V'_n + W_{m_{n+1}}$  has an eigenvalue which is at a distance  $\leq 1/n$  from  $\lambda$ . As  $V_n \subset V'_n$  for all  $n$ , we have  $\overline{\cup_{n \geq 1} V'_n}^{D(A)} = D(A)$ . By construction we also have  $\lim_{n \rightarrow \infty} \text{dist}(\lambda, \sigma(A|_{V'_n})) = 0$ . Hence either  $\lambda \in \sigma(A)$ , or  $\lambda \in \text{Spu}(A)$ .  $\square$

In the following we shall only be interested in the spurious eigenvalues of  $A$  lying in the convex hull of  $\hat{\sigma}_{\text{ess}}(A)$ . This is justified by the following simple result which tells us that pollution cannot occur below or above the essential spectrum.

**Lemma 1.3.** *Let  $\lambda$  be a spurious eigenvalue of the self-adjoint operator  $A$ . Then one has*

$$\text{Tr}(\chi_{(-\infty, \lambda]}(A)) = \text{Tr}(\chi_{[\lambda, \infty)}(A)) = +\infty. \quad (1.1)$$

*Saying differently,  $\lambda \in \text{Conv}(\hat{\sigma}_{\text{ess}}(A))$ .*

**Proof.** Assume for instance  $P := \chi_{(-\infty, \lambda]}(A)$  is finite-rank. As  $\lambda \notin \sigma(A)$ , we must have  $P = \chi_{(-\infty, \lambda + \epsilon]}(A)$  for some  $\epsilon > 0$ . Let  $\{x_n\}$  be as in Lemma 1.1. As  $P$  is finite rank,  $Px_n \rightarrow 0$  and  $(A - \lambda)Px_n \rightarrow 0$  strongly in  $\mathfrak{H}$ . Therefore  $P_{V_n}(A - \lambda)P^\perp x_n \rightarrow 0$  strongly. Note that  $(A - \lambda)P^\perp \geq \epsilon P^\perp$ , hence  $\langle P_{V_n}(A - \lambda)P^\perp x_n, x_n \rangle = \langle P^\perp(A - \lambda)P^\perp x_n, x_n \rangle \geq \epsilon \|P^\perp x_n\|^2$ . As the left hand side converges to zero, we infer  $\|x_n\| \rightarrow 0$  which contradicts Lemma 1.1.  $\square$

We have seen that pollution can only occur in the convex hull of  $\hat{\sigma}_{\text{ess}}(A)$ . Levitin and Shargorodsky have shown in [23] that (1.1) is indeed necessary and sufficient.

**Theorem 1.1 (Pollution in all spectral gaps [23]).** *Let  $A$  be a self-adjoint operator on  $\mathfrak{H}$  with dense domain  $D(A)$ . Then*

$$\overline{\text{Spu}(A)} \cup \hat{\sigma}_{\text{ess}}(A) = \text{Conv}(\hat{\sigma}_{\text{ess}}(A)).$$

**Remark 1.4.** *As  $J := \text{Conv}(\hat{\sigma}_{\text{ess}}(A)) \setminus \hat{\sigma}_{\text{ess}}(A)$  only contains discrete spectrum by assumption, Theorem 1.1 says that all points but a countable set in  $J$  are potential spurious eigenvalues.*

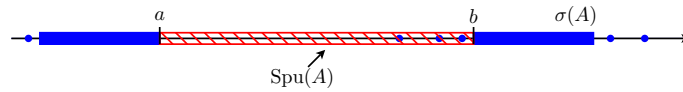


Fig. 1. For an operator  $A$  which has a spectral gap  $[a, b]$  in its essential spectrum, pollution can occur in the whole gap.

**Remark 1.5.** *It is easy to construct a sequence  $V_n$  like in Definition 1.1 such that  $\text{dist}(\lambda, \sigma(A|_{V_n})) \rightarrow 0$  for all  $\lambda \in \text{Conv}(\hat{\sigma}_{\text{ess}}(A))$ , see [23].*

Theorem 1.1 was proved for bounded self-adjoint operators in [28] and generalized to bounded non self-adjoint operators in [10]. For the convenience of the reader, we give a short

**Proof.** Let  $\lambda \in \text{Conv}(\hat{\sigma}_{\text{ess}}(A)) \setminus \hat{\sigma}_{\text{ess}}(A)$  and fix some  $a < \lambda$  and  $b > \lambda$  such that  $a, b \in \hat{\sigma}_{\text{ess}}(A)$  (*a priori* we might have  $b = +\infty$  or  $a = -\infty$ ). Let us consider two sequences  $\{x_n\}, \{y_n\} \subset D(A)$  such that  $(A - a_n)x_n \rightarrow 0$ ,  $(A - b_n)y_n \rightarrow 0$ ,  $\|x_n\| = \|y_n\| = 1$ ,  $x_n \rightarrow 0$ ,  $y_n \rightarrow 0$ ,  $a_n \rightarrow a$  and  $b_n \rightarrow b$ . Extracting subsequences if necessary we may assume that  $\langle x_n, y_n \rangle \rightarrow 0$  as  $n \rightarrow \infty$ . Next we consider the sequence  $z_n(\theta) := \cos \theta x_n + \sin \theta y_n$  which satisfies  $\|z_n(\theta)\| \rightarrow 1$  and  $z_n(\theta) \rightarrow 0$  uniformly in  $\theta$ . We note that  $\langle Az_n(0), z_n(0) \rangle = a_n + o(1)$  and  $\langle Az_n(\pi/2), z_n(\pi/2) \rangle = b_n + o(1)$ . Hence for  $n$  large enough there exists a  $\theta_n \in (0, \pi/2)$  such that  $\langle Az_n(\theta_n), z_n(\theta_n) \rangle = \lambda$ . The rest follows from Lemma 1.2.  $\square$

## 2. Pollution associated with a splitting of $\mathfrak{H}$

As we have recalled in the previous section, the union of the essential spectrum and (the closure of) the polluted spectrum is always an interval: it is simply the convex hull of  $\hat{\sigma}_{\text{ess}}(A)$ . It was also shown in [23] that it is possible to construct one sequence  $\{V_n\}$  such that all possible points in  $\text{Spu}(A)$  are indeed  $\{V_n\}$ -spurious eigenvalues. But of course, not all  $\{V_n\}$  will produce pollution. If for instance  $P_{V_n}$  commutes with  $A$  for all  $n \geq 1$ , then pollution will not occur as is obviously seen from Lemma 1.1. The purpose of this section is to study spectral pollution if we add some assumptions on  $\{V_n\}$ . More precisely we will fix an orthogonal projector  $P$  acting on  $\mathfrak{H}$  and we will add the natural assumption that  $P_{V_n}$  commute with  $P$  for all  $n$ , i.e. that  $V_n$  only contains vectors from  $P\mathfrak{H}$  and  $(1 - P)\mathfrak{H}$ .

As we will see, under this new assumption the polluted spectrum (union  $\hat{\sigma}_{\text{ess}}(A)$ ) will in general be *the union of two intervals*. Saying differently, by adding such an assumption on  $\{V_n\}$ , we can *create a hole in the polluted spectrum*. A typical situation is when our operator  $A$  has a gap in its essential spectrum. Then we will see that it is possible to give very simple conditions<sup>a</sup> on  $P$  which allow to completely avoid pollution in the gap.

Note that our results of this section can easily be generalized to the case of a partition of unity  $\{P_i\}_{i=1}^p$  of commuting projectors such that  $1 = \sum_{i=1}^p P_i$ . Adding the assumption that  $P_{V_n}$  commutes with all  $P_i$ 's, we would create  $p$  holes in the polluted spectrum. This might be useful if one wants to avoid spectral pollution in several gaps at the same time.

<sup>a</sup>Loosely speaking it must not be too far from the spectral projector associated with the part of the spectrum above the gap, as we will see below.

### 2.1. A general result

We start by defining properly  $P$ -spurious eigenvalues.

**Definition 2.1 (Spurious eigenvalues associated with a splitting).** Consider an orthogonal projection  $P : \mathfrak{H} \rightarrow \mathfrak{H}$ . We say that  $\lambda \in \mathbb{R}$  is a  $P$ -spurious eigenvalue of the operator  $A$  if there exist two sequences of finite dimensional spaces  $\{V_n^+\}_{n \geq 1} \subset P\mathfrak{H} \cap D(A)$  and  $\{V_n^-\}_{n \geq 1} \subset (1-P)\mathfrak{H} \cap D(A)$  with  $V_n^\pm \subset V_{n+1}^\pm$  for any  $n$ , such that

- (1)  $\overline{\cup_{n \geq 1} (V_n^- \oplus V_n^+)}^{D(A)} = D(A)$ ;
- (2)  $\lim_{n \rightarrow \infty} \text{dist} \left( \lambda, \sigma \left( A|_{(V_n^+ \oplus V_n^-)} \right) \right) = 0$ ;
- (3)  $\lambda \notin \sigma(A)$ .

We denote by  $\text{Spu}(A, P)$  the set of  $P$ -spurious eigenvalues of the operator  $A$ .

Now we will show as announced that contrarily to  $\overline{\text{Spu}(A)} \cup \hat{\sigma}_{\text{ess}}(A)$  which is always an interval,  $\overline{\text{Spu}(A, P)} \cup \hat{\sigma}_{\text{ess}}(A)$  is the union of two intervals, hence it may have a ‘‘hole’’.

**Theorem 2.1 (Characterization of  $P$ -spurious eigenvalues).** Let  $A$  be a self-adjoint operator with dense domain  $D(A)$ . Let  $P$  be an orthogonal projector on  $\mathfrak{H}$  such that  $PC \subset D(A)$  for some  $C \subset D(A)$  which is a core for  $A$ . We assume that  $PAP$  (resp.  $(1-P)A(1-P)$ ) is essentially self-adjoint on  $PC$  (resp.  $(1-P)C$ ), with closure denoted as  $A|_{P\mathfrak{H}}$  (resp.  $A|_{(1-P)\mathfrak{H}}$ ). We assume also that

$$\inf \hat{\sigma}_{\text{ess}}(A|_{(1-P)\mathfrak{H}}) \leq \inf \hat{\sigma}_{\text{ess}}(A|_{P\mathfrak{H}}). \quad (2.1)$$

Then we have

$$\begin{aligned} \overline{\text{Spu}(A, P)} \cup \hat{\sigma}_{\text{ess}}(A) &= [\inf \hat{\sigma}_{\text{ess}}(A), \sup \hat{\sigma}_{\text{ess}}(A|_{(1-P)\mathfrak{H}})] \\ &\cup [\inf \hat{\sigma}_{\text{ess}}(A|_{P\mathfrak{H}}), \sup \hat{\sigma}_{\text{ess}}(A)]. \end{aligned} \quad (2.2)$$

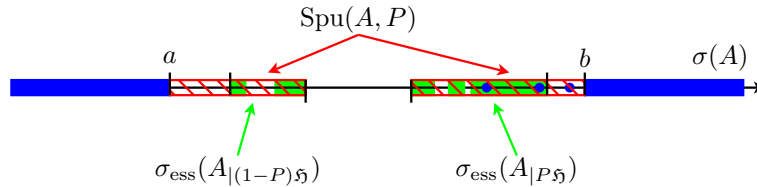


Fig. 2. Illustration of Theorem 2.1: for an operator  $A$  with a gap  $[a, b]$  in its essential spectrum, pollution can occur in the whole gap, except between the convex hulls of  $\hat{\sigma}_{\text{ess}}(A|_{P\mathfrak{H}})$  and  $\hat{\sigma}_{\text{ess}}(A|_{(1-P)\mathfrak{H}})$ .

Let us emphasize that condition (2.1) always holds true, exchanging  $P$  and  $1-P$  if necessary. Usually we will assume for convenience that  $1-P$  is ‘‘associated with the lowest part of the spectrum’’ in the sense of (2.1).

As mentioned before, an interesting example is when  $A$  possesses a gap  $[a, b]$  in its essential spectrum, i.e. such that  $(a, b) \cap \sigma_{\text{ess}}(A) = \emptyset$  and

$$\text{Tr}(\chi_{(-\infty, a]}(A)) = \text{Tr}(\chi_{[b, \infty)}(A)) = +\infty.$$

Then taking  $\Pi = \chi_{[c, \infty)}(A)$  and  $\mathcal{C} = D(A)$  we easily see that  $\text{Spu}(A, \Pi) \cap (a, b) = \emptyset$ . The idea that we shall pursue in the next section is simply that if  $P$  is ‘‘not too far from  $\Pi$ ’’, then we may be able to avoid completely pollution in the gap  $[a, b]$ .

Before writing the proof of Theorem 2.1, we make some remarks.

**Remark 2.1.** *If the symmetric operators  $PAP$  and  $(1-P)A(1-P)$  are both semi-bounded on their respective domains  $PC$  and  $(1-P)C$ , then the inclusion  $\subseteq$  in (2.2) is also true provided that  $A|_{P\mathfrak{H}}$  and  $A|_{(1-P)\mathfrak{H}}$  are defined as the corresponding Friedrichs extensions. The essential self-adjointness is only used to show the converse inclusion  $\supseteq$ .*

**Remark 2.2.** *An interesting consequence of Theorem 2.1 is that the set of spurious eigenvalues varies continuously when the projector  $P$  is changed (in an appropriate norm for which the spectra of  $A|_{P\mathfrak{H}}$  and  $A|_{(1-P)\mathfrak{H}}$  change continuously). This has important practical consequences: even if one knows a projector which does not create pollution, it could in principle be difficult to numerically build a basis respecting the splitting of  $\mathfrak{H}$  induced by  $P$ . However we know that pollution will only appear at the edges of the gap if the elements of the Galerkin basis are only known approximately.*

**Proof.** We will make use of the following result, whose proof will be omitted (it is an obvious adaptation of the proof of Lemma 1.2):

**Lemma 2.1.** *Assume that  $A$  is as above. Let  $(x_n^1, \dots, x_n^K)$  and  $(y_n^1, \dots, y_n^{K'})$  be two orthonormal systems<sup>b</sup> in  $P\mathfrak{H} \cap D(A)$  and  $(1-P)\mathfrak{H} \cap D(A)$  respectively, such that  $x_n^j \rightarrow 0$  and  $y_n^{k'} \rightarrow 0$  for all  $j = 1..K$  and  $k' = 1..K'$ . Denote by  $W_n$  the space spanned by  $x_n^1, \dots, x_n^K, y_n^1, \dots, y_n^{K'}$ . If  $\lambda \in \mathbb{R}$  is such that  $\lim_{n \rightarrow \infty} \text{dist}(\lambda, \sigma(A|_{W_n})) = 0$ , then  $\lambda \in \text{Spu}(A, P) \cup \sigma(A)$ .*

In the rest of the proof, we denote  $[a, b] := \text{Conv}(\hat{\sigma}_{\text{ess}}(A))$ ,  $[c_1, d_1] := \text{Conv}(\hat{\sigma}_{\text{ess}}(A|_{(1-P)\mathfrak{H}}))$  and  $[c_2, d_2] := \text{Conv}(\hat{\sigma}_{\text{ess}}(A|_{P\mathfrak{H}}))$ . For simplicity we also introduce  $c = \min(c_1, c_2) = c_1$ , and  $d = \max(d_1, d_2)$ . Recall that we have assumed  $c_1 \leq c_2$ .

**Step 1.** First we collect some easy facts. The first is to note that  $\text{Spu}(A, P) \subset \text{Spu}(A) \subset [a, b]$ , where we have used Theorem 1.1. Next we claim that

$$[c_1, d_1] \cup [c_2, d_2] \subset \text{Spu}(A, P) \cup \sigma(A) \cap [a, b]. \quad (2.3)$$

This is indeed an obvious consequence of Theorem 1.1 applied to  $A|_{P\mathfrak{H}}$  and  $A|_{(1-P)\mathfrak{H}}$ , and of Lemma 2.1.

**Step 2.** The second step is less obvious, it consists in proving that

$$[a, c] \cup [d, b] \subset \text{Spu}(A, P) \cup \sigma(A) \cap [a, b] \quad (2.4)$$

which then clearly implies

$$[a, d_1] \cup [c_2, b] \subset \text{Spu}(A, P) \cup \sigma(A) \cap [a, b].$$

Let us assume for instance that  $d < b$  and prove the statement for  $[d, b]$  (the proof is the same for  $[a, c]$ ). Note that  $b$  may a priori be equal to  $+\infty$  but of course we always have under this assumption  $d < +\infty$ . In principle we could however have  $d = -\infty$ . In the rest of the proof of (2.4), we fix some finite  $\lambda \in (d, b)$  and prove that  $\lambda \in \text{Spu}(A, P) \cup \sigma(A)$ . We also fix some finite  $d'$  such that  $d < d' < \lambda$ . We will use the following

<sup>b</sup>We will allow  $K = 0$  or  $K' = 0$ .

**Lemma 2.2.** *Assume that  $b \in \hat{\sigma}_{\text{ess}}(A)$ . Then there exists a Weyl sequence  $\{x_n\} \subset \mathcal{C}$  such that  $(A - b_n)x_n \rightarrow 0$ ,  $\|x_n\| = 1$ ,  $x_n \rightharpoonup 0$ ,  $b_n \rightarrow b$  and*

$$\frac{Px_n}{\|Px_n\|} \rightharpoonup 0 \quad \text{and} \quad \frac{(1-P)x_n}{\|(1-P)x_n\|} \rightharpoonup 0 \quad \text{weakly.} \quad (2.5)$$

**Proof.** Let  $b_n \rightarrow b$  and  $\{y_n\} \subset \mathcal{C}$  be a Weyl sequence such that  $(A - b_n)y_n \rightarrow 0$  with  $\|y_n\| = 1$ ,  $y_n \rightharpoonup 0$  (note we may assume  $\{y_n\} \subset \mathcal{C}$  since  $\mathcal{C}$  is a core for  $A$ ). We denote  $y_n = y_n^+ + y_n^-$  where  $y_n^+ \in PC \subset D(A)$  and  $y_n^- \in PC \subset D(A)$ . Extracting a subsequence, we may assume that  $\|y_n^+\|^2 \rightarrow \ell^+$  and that  $\|y_n^-\|^2 \rightarrow \ell^-$ ; note  $\ell^+ + \ell^- = 1$ . It is clear that if  $\ell^\pm > 0$ , then  $y_n^\pm \|y_n^\pm\|^{-1} \rightharpoonup 0$  since  $y_n^\pm \rightharpoonup 0$ . We will assume for instance  $\ell^+ = 0$  and  $\ell^- = 1$ .

Next we fix an orthonormal basis  $\{e_i\} \subset PC$  of  $P\mathfrak{H}$ , we define

$$r_k^+ := \sum_{i=1}^k \langle e_i, y_{n_k} \rangle e_i$$

and note that

$$(A - b_{n_k})r_k^+ = \sum_{i=1}^k \left( \langle e_i, y_{n_k} \rangle Ae_i + \langle e_i, (A - b_{n_k})y_{n_k} \rangle e_i - \langle Ae_i, y_{n_k} \rangle e_i \right).$$

For  $k$  fixed and any  $i = 1..k$ , we have

$$\lim_{n \rightarrow \infty} \langle e_i, y_n \rangle = \lim_{n \rightarrow \infty} \langle e_i, (A - b_n)y_n \rangle = \lim_{n \rightarrow \infty} \langle Ae_i, y_n \rangle = 0.$$

Hence, for a correctly chosen subsequence  $\{y_{n_k}\}$ , we may assume that

$$\text{satisfies} \quad \lim_{k \rightarrow \infty} \|r_k^+\| = \lim_{k \rightarrow \infty} \|(A - b_{n_k})r_k^+\| = 0.$$

Next we define  $x_k := y_{n_k} - r_k^+ = (y_{n_k}^+ - r_k^+) + y_{n_k}^-$  which satisfies  $\|x_k\| = 1 + o(1)$  since  $\|r_k^+\| \rightarrow 0$ . By construction, we have  $x_k^+ = y_{n_k}^+ - r_k^+ \in \text{span}(e_1, \dots, e_k)^\perp$ , hence necessarily  $x_k^+ \|x_k^+\|^{-1} \rightharpoonup 0$ . Eventually, we have  $(A - b_{n_k})x_k \rightarrow 0$  strongly, by construction of  $r_k^+$ .  $\square$

In the rest of the proof we choose a sequence  $\{x_n\}$  like in Lemma 2.2 and denote  $x_n^+ = Px_n$  and  $x_n^- = (1 - P)x_n$ . By the definition of  $d$  and the fact that  $A|_{(1-P)\mathfrak{H}}$  is essentially selfadjoint on  $(1 - P)\mathcal{C}$ , we can choose a Weyl sequence  $\{y_n^-\} \subset (1 - P)\mathcal{C}$  such that  $(1 - P)(A - d_n)y_n^- \rightarrow 0$ ,  $\|y_n^-\| = 1$ ,  $y_n^- \rightharpoonup 0$  weakly and  $d_n \rightarrow d_1 \leq d$ . Extracting a subsequence from  $\{y_n^-\}$  we may also assume that  $y_n^-$  satisfies

$$\lim_{n \rightarrow \infty} \left\langle \frac{x_n^-}{\|x_n^-\|}, y_n^- \right\rangle = \lim_{n \rightarrow \infty} \left\langle \frac{Ax_n^+}{\|x_n^+\|}, y_n^- \right\rangle = \lim_{n \rightarrow \infty} \left\langle \frac{Ax_n^-}{\|x_n^-\|}, y_n^- \right\rangle = 0 \quad (2.6)$$

Let us now introduce the following orthonormal system

$$\left( \frac{x_n^+}{\|x_n^+\|}, v_n(\theta) \right) \quad \text{with} \quad v_n(\theta) := \frac{\cos \theta \frac{x_n^-}{\|x_n^-\|} + \sin \theta y_n^-}{\sqrt{1 + 2\Re \cos \theta \sin \theta \left\langle \frac{x_n^-}{\|x_n^-\|}, y_n^- \right\rangle}} \quad (2.7)$$

and denote by  $A_n(\theta)$  the  $2 \times 2$  matrix of  $A$  in this basis, with eigenvalues  $\lambda_n(\theta) \leq \mu_n(\theta)$ . As  $x_n^+ \|x_n^+\|^{-1} \rightharpoonup 0$  weakly, we have

$$\limsup_{n \rightarrow \infty} \sup_{\theta \in [0, \pi/2]} \lambda_n(\theta) \leq \limsup_{n \rightarrow \infty} \frac{\langle Ax_n^+, x_n^+ \rangle}{\|x_n^+\|^2} \leq d_2 \leq d. \quad (2.8)$$

When  $\theta = 0$ , we know by construction of  $x_n$  that  $A_n(0)$  has an eigenvalue which converges to  $b$  as  $n \rightarrow \infty$ . Since  $b > d$  by assumption, this shows by (2.8) that this eigenvalue must be  $\mu_n(0)$ , hence we have  $\mu_n(0) \rightarrow b$  as  $n \rightarrow \infty$ . On the other hand, the largest eigenvalue of  $A_n(\pi/2)$  satisfies for  $n$  large enough

$$\mu_n(\pi/2) \leq \max \left( \frac{\langle Ax_n^+, x_n^+ \rangle}{\|x_n^+\|^2}, \langle Ay_n^-, y_n^- \rangle \right) + \left| \left\langle \frac{Ax_n^+}{\|x_n^+\|}, y_n^- \right\rangle \right| \leq d',$$

where we have used (2.6),  $x_n^+ \|x_n^+\|^{-1} \rightarrow 0$ ,  $y_n^- \rightarrow 0$ , and the definition of  $d' > d$ .

By continuity of  $\mu_n(\theta)$ , there exists a  $\theta_n \in (0, \pi/2)$  such that  $\mu_n(\theta_n) = \lambda$ . Next we note that the two elements of the basis defined in (2.7) both go weakly to zero by the construction of  $x_n^\pm$  and of  $y_n^-$ . Hence our statement  $\lambda \in \text{Spu}(A, P) \cup \sigma(A)$  follows from Lemma 2.1.

**Step 3.** The last step is to prove that when  $d_1 < c_2$ ,

$$(d_1, c_2) \cap (\text{Spu}(A, P) \cup \sigma_{\text{ess}}(A)) = \emptyset$$

(there is nothing else to prove when  $c_2 \leq d_1$ ). We will prove that  $(d_1, c_2) \cap \text{Spu}(A, P) = \emptyset$ , the proof for  $\sigma_{\text{ess}}(A)$  being similar. Note that under our assumption  $d_1 < c_2$ , we must have  $d_1 < \infty$  and  $c_2 > -\infty$ , hence  $A|_{P\mathfrak{H}}$  and  $A|_{(1-P)\mathfrak{H}}$  are semi-bounded operators. As noticed in Remark 2.1, it is sufficient to assume for this step that  $A|_{P\mathfrak{H}}$  and  $A|_{(1-P)\mathfrak{H}}$  are the Friedrichs extensions of  $(PAP, PC)$  and  $((1-P)A(1-P), (1-P)C)$  without assuming *a priori* that they are essentially self-adjoint.

Now we argue by contradiction and assume that there exists a Weyl sequence  $\{x_n\} \in V_n^+ \oplus V_n^- \subset D(A)$  like in Lemma 1.1, for some  $\lambda \in (d_1, c_2)$ . We will write  $x_n = x_n^+ + x_n^-$  with  $x_n^+ \in V_n^+$  and  $x_n^- \in V_n^-$ . We have  $P|_{V_n^+ \oplus V_n^-} (A - \lambda)x_n \rightarrow 0$ , hence taking the scalar product with  $x_n^+$  and  $x_n^-$ , we obtain

$$\lim_{n \rightarrow \infty} \langle (A - \lambda)x_n, x_n^+ \rangle = \lim_{n \rightarrow \infty} \langle (A - \lambda)x_n, x_n^- \rangle = 0. \quad (2.9)$$

The space  $\mathcal{C}$  being a core for  $A$ , it is clear that we may assume further that  $x_n \in \mathcal{C}$  and still that (2.9) holds true. In this case we have  $x_n^+, x_n^- \in D(A)$  hence we are allowed to write

$$\langle (A - \lambda)x_n^+, x_n^+ \rangle + \langle (A - \lambda)x_n^-, x_n^+ \rangle \rightarrow 0,$$

$$\langle (A - \lambda)x_n^-, x_n^- \rangle + \overline{\langle (A - \lambda)x_n^-, x_n^+ \rangle} \rightarrow 0.$$

Taking the complex conjugate of the second line (the first term is real since  $A$  is self-adjoint) and subtracting the two quantities, we infer that

$$\langle (A - \lambda)x_n^+, x_n^+ \rangle - \langle (A - \lambda)x_n^-, x_n^- \rangle \rightarrow 0. \quad (2.10)$$

As by assumption  $\lambda \in (d_1, c_2)$ , we have as quadratic forms on  $PC$  and  $(1-P)C$ ,  $P(A - \lambda)P \geq \epsilon P - r$  and  $-(1-P)(A - \lambda)(1-P) \geq \epsilon(1-P) - r'$  for some finite-rank operators  $r$  and  $r'$  and some  $\epsilon > 0$  small enough. Hence we have

$$\langle (A - \lambda)x_n^+, x_n^+ \rangle - \langle (A - \lambda)x_n^-, x_n^- \rangle \geq \epsilon \|x_n^+\|^2 + \epsilon \|x_n^-\|^2 + o(1).$$

This shows that we must have  $x_n \rightarrow 0$  which is a contradiction.  $\square$

## 2.2. A simple criterion of no pollution

Here we give a very intuitive condition allowing to avoid pollution in a gap.

**Theorem 2.2 (Compact perturbations of spectral projector do not pollute).** *Let  $A$  be a self-adjoint operator defined on a dense domain  $D(A)$ , and let  $a < b$  be such that*

$$(a, b) \cap \sigma_{\text{ess}}(A) = \emptyset \quad \text{and} \quad \text{Tr}(\chi_{(-\infty, a]}(A)) = \text{Tr}(\chi_{[b, \infty)}(A)) = +\infty. \quad (2.11)$$

*Let  $c \in (a, b) \setminus \sigma(A)$  and denote  $\Pi := \chi_{(c, \infty)}(A)$ . Let  $P$  be an orthogonal projector satisfying the assumptions of Theorem 2.1. We furthermore assume that  $(P - \Pi)|A - c|^{1/2}$ , initially defined on  $D(|A - c|^{1/2})$ , extends to a compact operator on  $\mathfrak{H}$ . Then we have*

$$\text{Spu}(A, P) \cap (a, b) = \emptyset.$$

As we will see in Corollary 2.1, Theorem 2.2 is useful when our operator takes the form  $A + B$  where  $B$  is  $A$ -compact. Using the spectral projector  $P = \Pi$  of  $A$  will then avoid pollution for  $A + B$ , when  $A$  is bounded from below.

**Remark 2.3.** *We give an example showing that the power  $1/2$  in  $|A - c|^{1/2}$  is sharp. Consider for instance an orthonormal basis  $\{e_n^\pm\}$  of a separable Hilbert space  $\mathfrak{H}$ , and define  $A := \sum_{n \geq 1} n|e_n^+\rangle\langle e_n^+|$ . Choosing  $c = 1/2$ , we get  $\Pi = \chi_{[1/2, \infty)}(A) = \sum_{n \geq 1} |e_n^+\rangle\langle e_n^+|$ . Define now a new basis by  $f_n^+ = \cos \theta_n e_n^+ + \sin \theta_n e_n^-$ ,  $f_n^- = \sin \theta_n e_n^+ - \cos \theta_n e_n^-$ , and introduce the associated projector  $P = \sum_{n \geq 1} |f_n^+\rangle\langle f_n^+|$ . Consider then  $V_n := \text{span}\{f_1^\pm, \dots, f_{n-1}^\pm, f_n^-\}$  for which we have  $\sigma(A|_{V_n}) = \{0, 1, \dots, n-1, n \sin^2 \theta_n\}$ . On the other hand it is easily checked that  $(P - \Pi)|A - 1/2|^\alpha$  is compact if and only if  $n^\alpha \theta_n \rightarrow 0$  as  $n \rightarrow \infty$ . Hence, if  $0 \leq \alpha < 1/2$  we can take  $\theta_n = 1/\sqrt{2n}$  and we will have a polluted eigenvalue at  $1/2$  whereas  $(P - \Pi)|A - 1/2|^\alpha$  is compact.*

We now write the proof of Theorem 2.2.

**Proof.** We will prove that  $\sigma_{\text{ess}}(A|_{P\mathfrak{H}}) \subset [b, \infty)$ . This will end the proof, by Theorem 2.1 and a similar argument for  $A|_{(1-P)\mathfrak{H}}$ . Assume on the contrary that  $\lambda \in (-\infty, b) \cap \sigma_{\text{ess}}(A|_{P\mathfrak{H}})$ . Without any loss of generality, we may assume that  $c > \lambda$  (changing  $c$  if necessary). As  $PC$  is a core for  $A|_{P\mathfrak{H}}$ , there exists a sequence  $\{x_n\} \subset PC$  such that  $x_n \rightharpoonup 0$  weakly in  $\mathfrak{H}$ ,  $\|x_n\| = 1$  and  $P(A - \lambda)x_n \rightarrow 0$  strongly in  $\mathfrak{H}$ . We have

$$\begin{aligned} & \langle (P - \Pi)(A - \lambda)(P - \Pi)x_n, x_n \rangle + 2\Re \langle (P - \Pi)(A - c)\Pi x_n, x_n \rangle \\ & + \langle \Pi(A - \lambda)\Pi x_n, x_n \rangle = \langle P(A - \lambda)x_n, x_n \rangle + (\lambda - c)2\Re \langle \Pi x_n, (P - \Pi)x_n \rangle \end{aligned} \quad (2.12)$$

where we note that  $Px_n = x_n \in D(A)$  and  $\Pi x_n \in D(A)$  since  $\Pi$  stabilizes  $D(A)$ . As  $c \notin \sigma(A)$ , we have that  $|A - c|^{-1/2}$  is bounded, hence  $P - \Pi$  must be a compact operator, i.e. the last term of the right hand side of (2.12) tends to 0 as  $n \rightarrow \infty$ . By the Cauchy-Schwarz inequality we have

$$|\langle (P - \Pi)(A - c)\Pi x_n, x_n \rangle| \leq \left\| |A - c|^{1/2} \Pi x_n \right\| \left\| |A - c|^{1/2} (P - \Pi)x_n \right\|. \quad (2.13)$$

As by assumption  $(P - \Pi)|A - c|^{1/2}$  is compact, we have that  $(P - \Pi)(A - \lambda)(P - \Pi)$  and  $|A - c|^{1/2}(P - \Pi)$  are also compact operators. Hence

$$\lim_{n \rightarrow \infty} \left\| |A - c|^{1/2} (P - \Pi)x_n \right\| = \lim_{n \rightarrow \infty} \langle (P - \Pi)(A - \lambda)(P - \Pi)x_n, x_n \rangle = 0.$$

On the other hand we have  $\Pi(A - \lambda)\Pi = \Pi(A - c)\Pi + (c - \lambda)\Pi \geq \Pi|A - c|\Pi$  since we have chosen  $c$  in such a way that  $c > \lambda$ , and by the definition of  $\Pi$ . Hence by (2.12) we have an inequality of the form

$$\left\| |A - c|^{1/2}\Pi x_n \right\|^2 - 2\epsilon_n \left\| |A - c|^{1/2}\Pi x_n \right\| \leq \epsilon'_n$$

where  $\lim_{n \rightarrow \infty} \epsilon_n = \lim_{n \rightarrow \infty} \epsilon'_n = 0$ . This clearly shows that

$$\lim_{n \rightarrow \infty} \left\| |A - c|^{1/2}\Pi x_n \right\| = 0.$$

Therefore we deduce  $\Pi x_n \rightarrow 0$  strongly,  $|A - c|^{1/2}$  being invertible. Hence  $x_n = Px_n = (P - \Pi)x_n + \Pi x_n \rightarrow 0$  and we have reached a contradiction.  $\square$

We now give a simple application of the above result.

**Corollary 2.1.** *Let  $A$  be a bounded-below self-adjoint operator defined on a dense domain  $D(A)$ , and let  $a < b$  be such that*

$$(a, b) \cap \sigma_{\text{ess}}(A) = \emptyset \quad \text{and} \quad \text{Tr}(\chi_{(-\infty, a]}(A)) = \text{Tr}(\chi_{[b, \infty)}(A)) = +\infty. \quad (2.14)$$

Let  $c \in (a, b)$  be such that  $c \notin \sigma(A)$  and denote  $\Pi := \chi_{(c, \infty)}(A)$ .

Let  $B$  be a symmetric operator such that  $A + B$  is self-adjoint on  $D(A)$  and such that  $((A + B - i)^{-1} - (A - i)^{-1})|A - c|^{1/2}$ , initially defined on  $D(|A - c|^{1/2})$ , extends to a compact operator on  $\mathfrak{H}$ . Then we have

$$\text{Spu}(A + B, \Pi) \cap (a, b) = \emptyset.$$

**Proof.** Under our assumption we have that  $(A + B - i)^{-1} - (A - i)^{-1}$  is compact, hence  $\sigma_{\text{ess}}(A + B) = \sigma_{\text{ess}}(A)$  by Weyl's Theorem [30, 8] and  $A + B$  is also bounded from below. Changing  $c$  if necessary we may assume that  $c \notin \sigma(A + B) \cup \sigma(A)$ . Next we take a curve  $\mathcal{C}$  in the complex plane enclosing the whole spectrum of  $A$  and  $A + B$  below  $c$  (i.e. intersecting the real axis only at  $c$  and  $c' < \inf \sigma(A) \cup \sigma(A + B)$ ). In this case, we have by Cauchy's formula and the resolvent identity

$$\begin{aligned} \left( \Pi - \chi_{[c, \infty)}(A + B) \right) |A - c|^{1/2} &= -\frac{1}{2i\pi} \oint_{\mathcal{C}} \left( \frac{1}{A + B - z} - \frac{1}{A - z} \right) |A - c|^{1/2} dz \\ &= -\frac{1}{2i\pi} \oint_{\mathcal{C}} \frac{A + B - i}{A + B - z} \left( \frac{1}{A + B - i} - \frac{1}{A - i} \right) |A - c|^{1/2} \frac{A - i}{A - z} dz \end{aligned}$$

Since  $\mathcal{C}$  is bounded (we use here that  $A$  is bounded-below), we easily deduce that the above operator is compact, hence the result follows from Theorem 2.2.  $\square$

**Remark 2.4.** *Again the power 1/2 in  $|A - c|^{1/2}$  is optimal, as seen by taking  $B = -A + \sum_n n |f_n^+\rangle \langle f_n^+|$  where  $A, f_n^+$  and  $\theta_n$  are chosen as in Remark 2.3 and  $V_n := \{e_1^\pm, \dots, e_{n-1}^\pm, e_n^-\}$ .*

**Remark 2.5.** *Corollary 2.1 is a priori wrong when  $A$  is not semi-bounded. This is seen by taking for instance  $A = \sum_{n \geq 1} n |e_n^+\rangle \langle e_n^+| - \sum_{n \geq 1} n |e_n^-\rangle \langle e_n^-|$  and  $B = -A + \sum_{n \geq 1} n |f_n^+\rangle \langle f_n^+| - \sum_{n \geq 1} n |f_n^-\rangle \langle f_n^-|$  where  $f_n^+ = e_n^+/\sqrt{2} + e_n^-/\sqrt{2}$  and  $f_n^- = -e_n^-/\sqrt{2} + e_n^+/\sqrt{2}$ . A short calculation shows that  $((A + B)^{-1} - A^{-1})|A|^\alpha$  is compact for all  $0 \leq \alpha < 1$  whereas  $0 \in \text{Spu}(A + B, \Pi)$  which is seen by choosing again  $V_n = \{e_1^\pm, \dots, e_{n-1}^\pm, e_n^-\}$ .*

### 2.3. Applications

#### 2.3.1. Periodic Schrödinger operators in Wannier basis

In this section, we show that approximating eigenvalues in gaps of periodic Schrödinger operators using a so-called *Wannier basis* does not yield any spectral pollution. This method was already successfully applied in dimension 1 in [6] for a nonlinear model introduced in [6]. For references on pollution in this setting, we refer for example to [4].

Consider  $d$  linearly independent vectors  $a_1, \dots, a_d$  in  $\mathbb{R}^d$  and denote by

$$\mathcal{L} := a_1\mathbb{Z} \oplus \dots \oplus a_d\mathbb{Z}$$

the associated lattice. We also define the *dual lattice*

$$\mathcal{L}^* := a_1^*\mathbb{Z} \oplus \dots \oplus a_d^*\mathbb{Z} \quad \text{with} \quad \langle a_i, a_j^* \rangle = (2\pi)\delta_{ij}.$$

Finally, the Brillouin zone is defined by

$$\mathcal{B} := \left\{ x \in \mathbb{R}^d \mid \|x\| = \inf_{k \in \mathcal{L}^*} \|x - k\| \right\}.$$

Next we fix an  $\mathcal{L}$ -periodic potential  $V_{\text{per}}$ , i.e.  $V_{\text{per}}(x + a) = V_{\text{per}}(x)$  for all  $a \in \mathcal{L}$ . We will assume as usual [30] that

$$V_{\text{per}} \in L^p(\mathcal{B}) \quad \text{where} \quad \begin{cases} p = 2 & \text{if } d \leq 3, \\ p > 2 & \text{if } d = 4, \\ p = d/2 & \text{if } d \geq 5. \end{cases}$$

In this case it is known [30] that the operator

$$A_{\text{per}} = -\Delta + V_{\text{per}} \tag{2.15}$$

is self-adjoint on  $H^2(\mathbb{R}^d)$ . One has the *Bloch-Floquet decomposition*

$$A_{\text{per}} = \frac{1}{|\mathcal{B}|} \int_{\mathcal{B}}^{\oplus} A_{\text{per}}(\xi) d\xi$$

where  $A_{\text{per}}(\xi)$  is for almost all  $\xi \in \mathcal{B}$  a self-adjoint operator acting on the space

$$L_{\xi}^2 = \{u \in L_{\text{loc}}^2(\mathbb{R}^3) \mid u(x + a) = e^{-ia \cdot \xi} u(x), \forall a \in \mathcal{L}\}.$$

For any  $\xi$ , the spectrum of  $A_{\text{per}}(\xi)$  is composed of a (nondecreasing) sequence of eigenvalues of finite multiplicity  $\lambda_k(\xi) \nearrow \infty$ , hence the spectrum

$$\sigma(A_{\text{per}}) = \sigma_{\text{ess}}(A_{\text{per}}) = \bigcup_{k \geq 1} \lambda_k(\mathcal{B})$$

is composed of bands. The eigenvalues  $\lambda_k(\xi)$  are known to be real-analytic in any fixed direction when  $V_{\text{per}}$  is smooth enough [39, 30], in which case the spectrum of  $A_{\text{per}}$  is purely absolutely continuous.

The operator (2.15) may be used to describe quantum electrons in a crystal. It appears naturally for noninteracting systems in which case  $V_{\text{per}}$  is the periodic Coulomb potential induced by the nuclei of the crystal. However operators of the form (2.15) also appear in nonlinear models taking into account the interaction between the electrons. In this case, the potential  $V_{\text{per}}$  contains an additional effective (mean-field) potential induced by the electrons

themselves [7, 6]. In the presence of an impurity in the crystal, one is led to consider an operator of the form

$$A = -\Delta + V_{\text{per}} + W. \quad (2.16)$$

We will assume in the following that

$$W \in L^p(\mathbb{R}^d) + L^\infty_\epsilon(\mathbb{R}^d) \text{ for some } p > \max(d/3, 1)$$

in which case  $(A_{\text{per}} + W - i)^{-1} - (A_{\text{per}} - i)^{-1}$  is  $(1 - \Delta)^{-1/2}$ -compact as seen by the resolvent expansion [30], and one has

$$\sigma_{\text{ess}}(A) = \sigma(A_{\text{per}}).$$

However eigenvalues may appear between the bands. Intuitively, they correspond to bound states of electrons (or holes) in presence of the defect. By Theorem 1.1, their computation may lead to pollution. For a finite elements-type basis, spectral pollution was studied in [4].

Using the Bloch-Floquet decomposition, a spectral decomposition of the reference periodic operator  $A_{\text{per}}$  is easily accessible numerically. This decomposition can be used as a starting point to avoid pollution for the perturbed operator  $A$ . For simplicity we shall assume that the spectral decomposition of  $A_{\text{per}}$  is known exactly. More precisely we make the assumption that there is a gap between the  $k$ th and the  $(k + 1)$ st band:

$$a := \sup \lambda_k(\mathcal{B}) < \inf \lambda_{k+1}(\mathcal{B}) := b$$

and that the associated spectral projector

$$P_{\text{per}} := \chi_{(-\infty, c)}(A_{\text{per}}), \quad c = \frac{a + b}{2}$$

is known. The interest of this approach is the following

**Theorem 2.3 (No pollution for periodic Schrödinger operators).** *We assume  $V_{\text{per}}$  and  $W$  are as before. Then we have*

$$\text{Spu}(A, P_{\text{per}}) \cap (a, b) = \emptyset. \quad (2.17)$$

**Proof.** This is a simple application of Corollary 2.1.  $\square$

It was noticed in [6] that a very natural basis respecting the decomposition associated with  $P_{\text{per}}$  is given by a so-called *Wannier basis* [40]. Wannier functions  $\{w_k\}$  are defined in such a way that  $w_k$  belongs to the spectral subspace associated with the  $k$ th band and  $\{w_k(\cdot - a)\}_{a \in \mathcal{L}}$  forms a basis of this spectral subspace. One can take

$$w_k(x) = \frac{1}{|\mathcal{B}|} \int_{\mathcal{B}} u_k(\xi, x) d\xi \quad (2.18)$$

where  $u_k(\xi, \cdot) \in L^2_\xi$  is for any  $\xi \in \mathcal{B}$  an eigenvector of  $A_{\text{per}}(\xi)$  corresponding to the  $k$ th eigenvalue  $\lambda_k(\xi)$ . The so-defined  $\{w_k(\cdot - a)\}_{a \in \mathcal{L}}$  are mutually orthogonal. Formula (2.18) does not define  $w_k$  uniquely since the  $u_k(\xi, x)$  are in the best case only known up to a phase. Choosing the right phase, one can prove that when the  $k$ th band is isolated from other bands,  $w_k$  decays exponentially [25].

More generally, instead of using only one band (i.e. one eigenfunction  $u_k(\xi, x)$ ), one can use  $K$  different bands for which it is possible to construct  $K$  exponentially localized Wannier

functions as soon as the union of the  $K$  bands is isolated from the rest of the spectrum [26, 5]. The union of the  $K$  bands is called a *composite band*.

In our case we typically have a natural composite band corresponding to the spectrum of  $A_{\text{per}}$  which is below  $c$ , and another one corresponding to the spectrum above  $c$  (the latter is not bounded above). By Theorem 2.3, we know that using such a basis will not create any pollution in the gap of  $A$ .

We emphasize that the Wannier basis *does not* depend on the decaying potential  $W$ , and can be precalculated once and for all for a given  $\mathcal{L}$  and a given  $V_{\text{per}}$ . Another huge advantage is that since  $w_k$  decays fast, it will be localized over a certain number of unit cells of  $\mathcal{L}$ . As  $W$  represents a localized defect in the lattice, keeping only the Wannier functions  $w_k(\cdot - a)$  with  $a \in \mathcal{L} \cap B(0, R)$  for some radius  $R > 0$  should already yield a very good approximation to the spectrum in the gap (we assume that the defect is localized in a neighborhood of 0). This approximation can be improved by enlarging progressively the radius  $R$ .

Of course in practice exponentially localized Wannier functions are not simple to calculate. But some authors have defined the concept of *maximally localized Wannier functions* [24] and proposed efficient methods to find these functions numerically.

The efficiency of the computation of the eigenvalues of  $A$  in the gap using a Wannier basis (compared to that of the so-called super-cell method) were illustrated for a nonlinear model in [6].

### 2.3.2. Dirac operators in upper/lower spinor basis

The Dirac operator is a differential operator of order 1 acting on  $L^2(\mathbb{R}^3, \mathbb{C}^4)$ , defined as [38, 15]

$$D^0 = -ic \sum_{k=1}^3 \alpha_k \partial_{x_k} + mc^2 \beta := c\alpha \cdot p + mc^2 \beta. \quad (2.19)$$

Here  $\alpha_1, \alpha_2, \alpha_3$  and  $\beta$  are the so-called Pauli  $4 \times 4$  matrices [38] which are chosen to ensure that

$$(D^0)^2 = -c^2 \Delta + m^2 c^4.$$

The usual representation in  $2 \times 2$  blocks is given by

$$\beta = \begin{pmatrix} I_2 & 0 \\ 0 & -I_2 \end{pmatrix}, \quad \alpha_k = \begin{pmatrix} 0 & \sigma_k \\ \sigma_k & 0 \end{pmatrix} \quad (k = 1, 2, 3),$$

where the Pauli matrices are defined as

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \quad (2.20)$$

In the whole paper we use the common notation  $p = -i\nabla$ .

The operator  $D^0$  is self-adjoint on  $H^1(\mathbb{R}^3, \mathbb{C}^4)$  and its spectrum is symmetric with respect to zero:  $\sigma(D^0) = (-\infty, -mc^2] \cup [mc^2, \infty)$ . An important problem is to compute eigenvalues of operators of the form

$$D^V = D^0 + V$$

in the gap  $(-mc^2, mc^2)$ , where  $V$  is a multiplication operator by a real function  $x \mapsto V(x)$ . Loosely speaking, positive eigenvalues correspond to bound states of a relativistic quantum

electron in the external field  $V$ , whereas negative eigenvalues correspond to bound states of a positron, the anti-particle of the electron. In practice, spectral pollution is an important problem [13, 18, 22, 34] which is dealt with in Quantum Physics and Chemistry by means of several different methods, the most widely used being the so-called *kinetic balance* which we will study later in Section 3.2.1. We refer to [3] for a recent numerical study based on the so-called *second-order method* for the radial Dirac operator.

We now present a heuristic argument which can be made mathematically rigorous in many cases [38, 15]. First we write the equation satisfied by an eigenvector  $(\varphi, \chi)$  of  $D^0 + V$  with eigenvalue  $mc^2 + \lambda \in (-mc^2, mc^2)$  as follows:

$$\begin{cases} (mc^2 + V)\varphi + c\sigma \cdot (-i\nabla)\chi = (mc^2 + \lambda)\varphi, \\ (-mc^2 + V)\chi + c\sigma \cdot (-i\nabla)\varphi = (mc^2 + \lambda)\chi, \end{cases} \quad (2.21)$$

where we recall that  $\sigma = (\sigma_1, \sigma_2, \sigma_3)$  are the Pauli matrices defined in (2.20). Hence one deduces that (when it makes sense)

$$\chi = \frac{c}{2mc^2 + \lambda - V} \sigma \cdot (-i\nabla)\varphi. \quad (2.22)$$

If  $V$  and  $\lambda$  stay bounded, we infer that, at least formally,

$$\begin{pmatrix} \varphi \\ \chi \end{pmatrix} \sim_{c \rightarrow \infty} \begin{pmatrix} \varphi \\ \frac{1}{2mc} \sigma \cdot (-i\nabla)\varphi \end{pmatrix}. \quad (2.23)$$

Hence we see that in the nonrelativistic limit  $c \rightarrow \infty$ , the eigenvectors of  $A$  associated with a positive eigenvalue converge to a vector of the form  $\begin{pmatrix} \varphi \\ 0 \end{pmatrix}$ . Reintroducing the asymptotic formula (2.23) of  $\chi$  in the first equation of (2.21), one gets that  $\varphi$  is an eigenvector of the nonrelativistic operator  $-\Delta/(2m) + V$  in  $L^2(\mathbb{R}^3, \mathbb{C}^2)$ .

For this reason, it is very natural to consider a splitting of the Hilbert space  $L^2(\mathbb{R}^3, \mathbb{C}^4)$  into *upper* and *lower spinor* and we introduce the following orthogonal projector

$$\mathcal{P} \begin{pmatrix} \varphi \\ \chi \end{pmatrix} = \begin{pmatrix} \varphi \\ 0 \end{pmatrix}, \quad \varphi, \chi \in L^2(\mathbb{R}^3, \mathbb{C}^2). \quad (2.24)$$

This splitting is the choice of most of the methods we are aware of in Quantum Physics and Chemistry. Applying Theorem 2.1, we can characterize the spurious spectrum associated with this splitting. For simplicity we take  $m = c = 1$  in the following.

**Theorem 2.4 (Pollution in upper/lower spinor basis for Dirac operators).** *Assume that the real function  $V$  satisfies the following assumptions:*

(i) *there exist  $\{R_k\}_{k=1}^M \subset \mathbb{R}^3$  and a positive number  $r < \inf_{k \neq \ell} |R_k - R_\ell|/2$  such that*

$$\max_{k=1..K} \sup_{|x-R_k| \leq r} |x - R_k| |V(x)| < \frac{\sqrt{3}}{2}; \quad (2.25)$$

(ii) *one has<sup>c</sup>*

$$V \mathbf{1}_{\mathbb{R}^3 \setminus \cup_1^K B(R_k, r)} \in L^p(\mathbb{R}^3) + L_\epsilon^\infty(\mathbb{R}^3) \quad \text{for some } 3 < p < \infty. \quad (2.26)$$

<sup>c</sup>We use the notation of [30]:  $X + L_\epsilon^\infty = \{f \in X + L^\infty \mid \forall \epsilon > 0, \exists f_\epsilon \in X \text{ such that } \|f - f_\epsilon\|_{L^\infty} \leq \epsilon\}$ .

Let  $\mathcal{P}$  be as in (2.24). Then one has

$$\overline{\text{Spu}(D^0 + V, \mathcal{P})} = \{ \text{Conv}(\text{Ess}(1 + V)) \cup \text{Conv}(\text{Ess}(-1 + V)) \} \cap [-1, 1] \quad (2.27)$$

where  $\text{Ess}(W)$  denotes the essential range of the function  $W$ , i.e.

$$\text{Ess}(W) = \{ \lambda \in \mathbb{R} \mid |W^{-1}([\lambda - \epsilon, \lambda + \epsilon])| \neq 0 \forall \epsilon > 0 \}.$$

**Remark 2.6.** It is known that the operator  $D^0 + V$  is essentially self-adjoint on  $C_0^\infty(\mathbb{R}^3, \mathbb{C}^4)$  when (2.25) and (2.26) hold, and that its domain is simply the domain  $H^1(\mathbb{R}^3, \mathbb{C}^4)$  of the free Dirac operator. When  $\sqrt{3}/2$  is replaced by 1 in (2.25), the operator  $D^0 + V$  still has a distinguished self-adjoint extension [38] whose associated domain satisfies  $H^1(\mathbb{R}^3, \mathbb{C}^4) \subsetneq D(D^0 + V) \subset H^{1/2}(\mathbb{R}^3, \mathbb{C}^4)$ . Furthermore this domain is not stable by the projector  $\mathcal{P}$  on the upper spinor (a characterization of this domain was given in [16]). The generalization to this case is possible but it is outside the scope of this paper.

**Remark 2.7.** By Theorem 2.4, we see that  $\text{Spu}(D^0, \mathcal{P}) = \emptyset$  but  $\text{Spu}(D^0 + V, \mathcal{P}) \neq \emptyset$  for all smooth potentials  $V \neq 0$  even if  $V$  is  $D^0$ -compact. Hence spectral pollution is in general not stable under relatively compact perturbations (but it is obviously stable under compact perturbations as we have already mentioned in Remark 1.2).

Our assumptions on  $V$  cover the case of the Coulomb potential,  $V(x) = \kappa|x|^{-1}$  when  $|\kappa| < \sqrt{3}/2$ . In our units, this corresponds to nuclei which have less than 118 protons, which covers all existing atoms. On the other hand, a typical example for which  $V \in L^p(\mathbb{R}^3) \cap L^\infty(\mathbb{R}^3)$  is the case of smeared nuclei  $V = \rho * 1/|x|$  where  $\rho$  is a (sufficiently smooth) distribution of charge for the nuclei. We now give the proof of Theorem 2.4:

**Proof.** Under assumptions (i) and (ii), it is known that  $D^0 + V$  is self-adjoint on  $H^1(\mathbb{R}^3, \mathbb{C}^4)$  (which is stable under the action of  $\mathcal{P}$ ) and that  $\sigma_{\text{ess}}(D^0 + V) = (-\infty, -1] \cup [1, \infty)$ . We will simply apply Theorem 2.1 with  $\mathcal{C} = H^1(\mathbb{R}^3, \mathbb{C}^4)$ . We have in the decomposition of  $L^2(\mathbb{R}^3)$  associated with  $\mathcal{P}$ ,

$$D^0 + V = \begin{pmatrix} 1 + V & \sigma \cdot (-i\nabla) \\ \sigma \cdot (-i\nabla) & -1 + V \end{pmatrix}.$$

Hence  $\mathcal{P}(D^0 + V)\mathcal{P} = 1 + V$  and  $(1 - \mathcal{P})(D^0 + V)(1 - \mathcal{P}) = -1 + V$ , both seen as operators acting on  $L^2(\mathbb{R}^3, \mathbb{C}^2)$ . It is clear that  $D(D^0 + V) \cap \mathcal{P}L^2(\mathbb{R}^3, \mathbb{C}^4) \simeq H^1(\mathbb{R}^3, \mathbb{C}^2)$  is dense in the domain of the multiplication operator by  $V(x)$

$$D(V) = \{ f \in L^2(\mathbb{R}^3, \mathbb{C}^2) \mid Vf \in L^2(\mathbb{R}^3, \mathbb{C}^2) \},$$

for the associated norm

$$\|f\|_{G(V)}^2 = \int_{\mathbb{R}^3} (1 + |V(x)|^2) |f(x)|^2 dx.$$

Also the spectrum of  $V$  is the essential range of  $V$ . Note under our assumptions on  $V$  we have that  $0 \in \text{Ess}(V)$ . The rest follows from Theorem 2.1.  $\square$

### 2.3.3. Dirac operators in dual basis

In this section we study a generalization of the decomposition into upper and lower spinors, which was introduced by Shabaev et al [32]. For any fixed  $\epsilon$ , we consider the unitary operator

$$U_\epsilon := \frac{D^0(\epsilon p)}{|D^0(\epsilon p)|} \quad (2.28)$$

which is just a dilation of the sign of  $D^0$  (note that  $(U_\epsilon)^* = U_\epsilon$ ). Next we define the following orthogonal projector

$$\mathcal{P}_\epsilon := U_\epsilon \mathcal{P} U_\epsilon \quad (2.29)$$

where  $\mathcal{P}$  is the projector on the upper spinors as defined in (2.24). As for  $\epsilon = 0$  we have  $U_0 = 1$ , we deduce that  $\mathcal{P}_0 = \mathcal{P}$ . However, as we will see below, the limit  $\epsilon \rightarrow 0$  seems to be rather singular from the point of view of spectral pollution. We note that any vector in  $\mathcal{P}_\epsilon L^2(\mathbb{R}^3, \mathbb{C}^4)$  may be written in the following simple form

$$\begin{pmatrix} \varphi \\ \epsilon \sigma \cdot (-i\nabla)\varphi \end{pmatrix} \quad \text{with } \varphi \in H^1(\mathbb{R}^3, \mathbb{C}^2).$$

Hence for  $\epsilon \ll 1$ , the above choice just appears as a kind of correction to the simple decomposition into upper and lower spinors. Also we notice that  $\mathcal{P}_\epsilon H^1(\mathbb{R}^3, \mathbb{C}^4) \subset H^1(\mathbb{R}^3, \mathbb{C}^4)$  for every  $\epsilon$  since  $U_\epsilon$  is a multiplication operator in Fourier space and  $\mathcal{P}$  stabilizes  $H^1(\mathbb{R}^3, \mathbb{C}^4)$ .

In [32], the projector  $\mathcal{P}_\epsilon$  is considered with  $\epsilon = 1/(2mc)$  as suggested by Equation (2.23). However here we will for convenience let  $\epsilon$  free. The method was called “dual” in [32] since contrarily to the ones that we will study later on (the *kinetic* and *atomic balance* methods), the two subspaces  $\mathcal{P}_\epsilon L^2(\mathbb{R}^3, \mathbb{C}^4)$  and  $(1 - \mathcal{P}_\epsilon)L^2(\mathbb{R}^3, \mathbb{C}^4)$  play a symmetric role. For this reason, the dual method was suspected to avoid pollution in the whole gap and not only in the upper part. Our main result is the following (let us recall that  $m = c = 1$ ):

**Theorem 2.5 (Pollution in dual basis).** *Assume that the real function  $V$  satisfies the following assumptions:*

(i) *there exist  $\{R_k\}_{k=1}^M \subset \mathbb{R}^3$  and a positive number  $r < \inf_{k \neq \ell} |R_k - R_\ell|/2$  such that*

$$\max_{k=1..K} \sup_{|x-R_k| \leq r} |x - R_k| |V(x)| < \frac{\sqrt{3}}{2}; \quad (2.30)$$

(ii) *one has*

$$V \mathbb{1}_{\mathbb{R}^3 \setminus \cup_1^K B(R_k, r)} \in L^p(\mathbb{R}^3) \cap L^\infty(\mathbb{R}^3) \quad \text{for some } 3 < p < \infty. \quad (2.31)$$

Let  $0 < \epsilon \leq 1$  and  $\mathcal{P}_\epsilon$  as defined in (2.29). Then one has<sup>d</sup>

$$\overline{\text{Spu}(D^0 + V, \mathcal{P}_\epsilon)} = \left[ -1, \min \left\{ -\frac{2}{\epsilon} + 1 + \sup V, 1 \right\} \right] \cup \left[ \max \left\{ -1, \frac{2}{\epsilon} - 1 + \inf V \right\}, 1 \right]. \quad (2.32)$$

Our result shows that contrarily to the decomposition into upper and lower spinors studied in the previous section, the use of  $\mathcal{P}_\epsilon$  indeed allows to avoid spectral pollution under

<sup>d</sup>Recall that  $[a, b] = \emptyset$  if  $b < a$ .

the condition that  $V$  is a bounded potential and that  $\epsilon$  is small enough:

$$\epsilon \leq \frac{2}{2 + |V|}.$$

This mathematically justifies a claim of [32]. However we see that for Coulomb potentials, we will again get pollution in the whole gap, independently of the choice of  $\epsilon$ . Also for large but bounded potentials (like the ones approximating a Coulomb potential), one might need to take  $\epsilon$  so small that this could give rise to a numerical instability.

**Proof.** We will again apply Theorem 2.1. We choose  $\mathcal{C} = U_\epsilon C_0^\infty(\mathbb{R}^3, \mathbb{C}^4)$ . Note that  $\mathcal{C}$  is a core for  $D^0 + V$  (its domain is simply  $H^1(\mathbb{R}^3, \mathbb{C}^4)$ ) and that  $\mathcal{P}_\epsilon \mathcal{C} \subset \cap_{s \geq 0} H^s(\mathbb{R}^3, \mathbb{C}^4)$  since  $\mathcal{P}_\epsilon$  and  $U_\epsilon$  commute with the operator  $p = -i\nabla$ . An easy computation yields

$$U_\epsilon(D^0 + V)|_{\mathcal{P}_\epsilon \mathcal{C}} U_\epsilon \simeq 1 + \frac{1}{\sqrt{1 + \epsilon^2 |p|^2}} V \frac{1}{\sqrt{1 + \epsilon^2 |p|^2}} + \frac{\epsilon \sigma \cdot p}{\sqrt{1 + \epsilon^2 |p|^2}} \left( \frac{2}{\epsilon} - 2 + V \right) \frac{\epsilon \sigma \cdot p}{\sqrt{1 + \epsilon^2 |p|^2}} := A_1 \quad (2.33)$$

and

$$U_\epsilon(D^0 + V)|_{(1 - \mathcal{P}_\epsilon) \mathcal{C}} U_\epsilon \simeq -1 + \frac{1}{\sqrt{1 + \epsilon^2 |p|^2}} V \frac{1}{\sqrt{1 + \epsilon^2 |p|^2}} + \frac{\epsilon \sigma \cdot p}{\sqrt{1 + \epsilon^2 |p|^2}} \left( -\frac{2}{\epsilon} + 2 + V \right) \frac{\epsilon \sigma \cdot p}{\sqrt{1 + \epsilon^2 |p|^2}} := A_2. \quad (2.34)$$

Strictly speaking these operators should be defined on  $\mathcal{P}U_\epsilon \mathcal{C}$  and  $(1 - \mathcal{P})U_\epsilon \mathcal{C}$  but we have made the identification  $\mathcal{P}U_\epsilon \mathcal{C} \simeq (1 - \mathcal{P})U_\epsilon \mathcal{C} \simeq C_0^\infty(\mathbb{R}^3, \mathbb{C}^2)$ . Let us remark that for  $\epsilon > 0$  the term  $K := (1 + \epsilon^2 |p|^2)^{-1/2} V (1 + \epsilon^2 |p|^2)^{-1/2}$  is indeed compact under our assumptions on  $V$ , hence it does not contribute to the polluted spectrum. On the other hand for  $\epsilon = 0$  it is the only term yielding pollution as we have seen before.

Theorem 2.5 is then a consequence of Theorem 2.1 and of the following

**Lemma 2.3 (Properties of  $(D^0 + V)|_{\mathcal{P}_\epsilon \mathcal{C}}$  and  $(D^0 + V)|_{(1 - \mathcal{P}_\epsilon) \mathcal{C}}$ ).** *The operators  $A_1$  and  $A_2$  defined in (2.33) and (2.34) are self-adjoint on the domain*

$$\mathcal{D} := \left\{ \varphi \in L^2(\mathbb{R}^3, \mathbb{C}^2) \mid V(\sigma \cdot p)(1 + \epsilon^2 |p|^2)^{-1/2} \varphi \in L^2(\mathbb{R}^3, \mathbb{C}^2) \right\}.$$

*They are both essentially self-adjoint on  $C_0^\infty(\mathbb{R}^3, \mathbb{C}^2)$ . Moreover, we have*

$$\begin{aligned} \text{Conv Ess} \left( \frac{2}{\epsilon} - 1 + V \right) &\subseteq \text{Conv } \sigma_{\text{ess}}(A_1) \subseteq \\ &\subseteq \left[ \min \left\{ 1, \frac{2}{\epsilon} - 1 + \inf V \right\}, \max \left\{ 1, \frac{2}{\epsilon} - 1 + \sup V \right\} \right] \end{aligned}$$

and

$$\begin{aligned} \text{Conv Ess} \left( -\frac{2}{\epsilon} + 1 + V \right) &\subseteq \text{Conv } \sigma_{\text{ess}}(A_2) \subseteq \\ &\subseteq \left[ \min \left\{ -1, -\frac{2}{\epsilon} + 1 + \inf V \right\}, \max \left\{ -1, -\frac{2}{\epsilon} + 1 + \sup V \right\} \right]. \end{aligned}$$

**Proof.** The operator  $K = (1 + \epsilon^2|p|^2)^{-1/2}V(1 + \epsilon^2|p|^2)^{-1/2}$  being compact, it suffices to prove the statement for  $\mathcal{L}_\epsilon(-2/\epsilon + 2 + V)\mathcal{L}_\epsilon$ , where we have introduced the notation  $\mathcal{L}_\epsilon := \epsilon\sigma \cdot p(1 + \epsilon^2|p|^2)^{-1/2}$ . The argument is exactly similar for  $\mathcal{L}_\epsilon(2/\epsilon - 2 + V)\mathcal{L}_\epsilon$ . We denote  $W := -2/\epsilon + 2 + V$  and we introduce  $A = \mathcal{L}_\epsilon W \mathcal{L}_\epsilon$  which is a symmetric operator defined on  $\mathcal{D}$ . We also note that  $\mathcal{D}$  is dense in  $L^2$ .

Let  $f \in D(A^*)$ , i.e. such that  $|\langle f, \mathcal{L}_\epsilon W \mathcal{L}_\epsilon \varphi \rangle| \leq C \|\varphi\|$ ,  $\forall \varphi \in \mathcal{D}$ . We introduce  $\chi := \mathbb{1}_{\cup_{k=1}^M B(R_k, r)}$ , a localizing function around the singularities of  $V$ , and we recall that  $V$  is bounded away from the  $R_k$ 's. Hence we also have  $|\langle f, \mathcal{L}_\epsilon \chi W \mathcal{L}_\epsilon \varphi \rangle| \leq C' \|\varphi\|$  for all  $\varphi \in \mathcal{D}$ . Then we notice that under our assumptions on  $V$ , we have  $W\chi \in L^2$ , hence  $g := \chi W \mathcal{L}_\epsilon f \in L^1$  and  $\widehat{g} \in L^\infty$ . In Fourier space the property  $|\int \widehat{g} \widehat{\mathcal{L}_\epsilon \varphi}| \leq C' \|\varphi\|$  for all  $\varphi$  in a dense subspace of  $L^2$  means that  $\epsilon\sigma \cdot p(1 + \epsilon^2|p|^2)^{-1/2}\widehat{g}(p) \in L^2$ , hence  $\widehat{g} \in L^2(\mathbb{R}^3 \setminus B(0, 1))$ . As by construction  $\widehat{g} \in L^\infty$ , we finally deduce that  $\widehat{g} \in L^2$ , hence  $W\mathcal{L}_\epsilon f \in L^2$ . We have proven that  $D(A^*) \subseteq \mathcal{D}$ , hence  $A$  is self-adjoint on  $\mathcal{D}$ . The essential self-adjointness is easily verified.

The next step is to identify the essential spectrum of  $A$ . We consider a smooth normalized function  $\zeta \in C_0^\infty(\mathbb{R}^3, \mathbb{R})$  and we introduce  $\varphi_1 = (1 + \sigma \cdot p/|p|)(\zeta, 0)$ . We notice that  $\varphi_1 \in H^s(\mathbb{R}^3, \mathbb{C}^2)$  for all  $s > 0$ . Then we let  $\varphi_n(x) := n^{3/2}\varphi_1(n(x-x_0))$  and note that  $(\sigma \cdot p/|p|)\varphi_n = \varphi_n$ . We take for  $x_0 \in \mathbb{R}^3$  some fixed Lebesgue point of  $V$ , i.e. such that

$$\lim_{r \rightarrow 0} \frac{1}{|B(x_0, r)|} \int_{B(x_0, r)} |V(x) - V(x_0)| dx = 0. \quad (2.35)$$

First we notice that

$$\lim_{n \rightarrow \infty} \left\| \left( \frac{\epsilon\sigma \cdot p}{\sqrt{1 + \epsilon^2|p|^2}} - 1 \right) \varphi_n \right\|_{H^1} = \lim_{n \rightarrow \infty} \left\| \left( \frac{\epsilon|p|}{\sqrt{1 + \epsilon^2|p|^2}} - 1 \right) \varphi_n \right\|_{H^1} = 0$$

as is seen by Fourier transform and Lebesgue's dominated convergence theorem. Therefore,

$$\lim_{n \rightarrow \infty} \left\| W \left( \frac{\epsilon\sigma \cdot p}{\sqrt{1 + \epsilon^2|p|^2}} - 1 \right) \zeta_n \right\|_{L^2} = 0 \quad (2.36)$$

since we have  $W \in L^2 + L^\infty$ . On the other hand we have  $\lim_{n \rightarrow \infty} \|(W - W(x_0))\zeta_n\|_{L^2} = 0$ . Using this to estimate cross terms we obtain  $\lim_{n \rightarrow \infty} \|(A - W(x_0))\zeta_n\|_{L^2} = 0$ . This proves that  $\text{Ess}(W) \subseteq \sigma_{\text{ess}}(A)$ . Let us remark that  $0 \in \sigma_{\text{ess}}(A)$  as seen by taking  $\varphi'_n(x) = n^{-3/2}\varphi_1(x/n)$ .

The last step is to show that  $\sigma_{\text{ess}}(A) \subseteq [\min\{0, \inf(W)\}, \max\{0, \sup(W)\}]$ . When  $\sup(W) < \infty$ , we estimate  $A \leq \sup(W)\mathcal{L}_\epsilon^2$ . If  $W \leq 0$ , then we just get  $A \leq 0$ , hence  $\sigma(A) \subseteq (-\infty, 0]$ . If  $0 < \sup(W) < \infty$ , we can estimate  $\mathcal{L}_\epsilon^2 \leq 1$  and get  $\sigma(A) \subset (-\infty, \sup(W)]$ . Repeating the argument for the lower bound, this ends the proof of Lemma 2.3.  $\square$

#### 2.3.4. Dirac operators in free basis

In this section, we prove that a way to avoid pollution in the *whole gap* is to take a basis associated with the spectral decomposition of the free Dirac operator, i.e. choosing as projector  $P_+^0 := \chi_{(0, \infty)}(D^0)$ . As we will see this choice does not rely on the size of  $V$  like in the previous section. Its main disadvantage compared to the dual method making use of  $\mathcal{P}_\epsilon$ , is that constructing a basis preserving the decomposition induced by  $P_+^0$  requires a Fourier transform, which might increase the computational cost dramatically. First we treat the case of a 'smooth' enough potential.

**Theorem 2.6 (No pollution in free basis - nonsingular case).** *Assume that  $V$  is a real function such that*

$$V \in L^p(\mathbb{R}^3) + \left( L^r(\mathbb{R}^3) \cap \dot{W}^{1,q}(\mathbb{R}^3) \right) + L^\infty_\epsilon(\mathbb{R}^3)$$

for some  $6 < p < \infty$ , some  $3 < r \leq 6$  and some  $2 < q < \infty$ . Then one has

$$\text{Spu}(D^0 + V, P_+^0) = \emptyset.$$

**Remark 2.8.** *We have used the notation*

$$L^r(\mathbb{R}^3) \cap \dot{W}^{1,q}(\mathbb{R}^3) = \{V \in L^r(\mathbb{R}^3) \mid \nabla V \in L^q(\mathbb{R}^3)\}.$$

**Remark 2.9.** *A physical situation for which the potential  $V$  satisfies the assumptions of the theorem is  $V = \rho * \frac{1}{|x|}$  with  $\rho \in L^1(\mathbb{R}^3) \cap L^2(\mathbb{R}^3)$ .*

**Proof.** Under the above assumptions on the potential  $V$ , it is easily seen that the operator  $D^0 + V$  is self-adjoint with domain  $H^1(\mathbb{R}^3, \mathbb{C}^4)$ , the same as  $D^0$ , and that  $\sigma_{\text{ess}}(D^0 + V) = \sigma(D^0) = (-\infty, -1] \cup [1, \infty)$  (these claims are indeed a consequence of the calculation below). Hence  $(-1, 1)$  only contains eigenvalues of finite multiplicity of  $D^0 + V$  and we may find a  $c \in (-1, 1) \setminus \sigma(D^0 + V)$ . In the following we shall assume for simplicity that  $c = 0$ . The argument is very similar if  $0 \in \sigma(D^0 + V)$ . We will denote  $\Pi = \chi_{[0, \infty)}(D^0 + V)$  and prove that  $(P_+^0 - \Pi)|D^0 + V|^{1/2}$  is compact. This will end the proof, by Theorem 2.2.

As  $0 \notin \sigma(D^0 + V)$ , we have that  $|D^0 + V| \geq \epsilon$  for some  $\epsilon > 0$ . Also, we have

$$\epsilon|D^0|^2 + C_1 \leq (D^0 + V)^2 \leq \epsilon|D^0|^2 + C_2$$

for  $\epsilon \geq 0$  small enough. Taking the square root of the above inequality, this proves that  $|D^0|^{-1/2}|D^0 + V|^{1/2}$  and its inverse are both bounded operators.

Next we use the resolvent formula together with Cauchy's formula like in [19] to infer

$$\begin{aligned} (P_+^0 - \Pi)|D^0 + V|^{1/2} &= -\frac{1}{2\pi} \int_{-\infty}^{\infty} \left( \frac{1}{D^0 + V + i\eta} - \frac{1}{D^0 + i\eta} \right) |D^0 + V|^{1/2} d\eta \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{1}{D^0 + i\eta} V \frac{|D^0 + V|^{1/2}}{D^0 + V + i\eta} d\eta. \end{aligned}$$

Let us now write  $V = V_1^n + V_2^n + V_3^n$  with  $V_1^n \in L^p(\mathbb{R}^3)$  for  $6 < p < \infty$ ,  $V_2^n \in L^r(\mathbb{R}^3)$  and  $\nabla V_2^n \in L^q(\mathbb{R}^3)$  for  $3 < r \leq 6$ ,  $2 < q < \infty$ , and  $\|V_3^n\|_{L^\infty(\mathbb{R}^3)} \rightarrow 0$  as  $n \rightarrow \infty$ . We write

$$(P_+^0 - \Pi)|D^0 + V|^{1/2} = K(V_1^n) + K(V_2^n) + K(V_3^n)$$

with

$$K(W) := \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{1}{D^0 + i\eta} W \frac{|D^0 + V|^{1/2}}{D^0 + V + i\eta} d\eta$$

and estimate each term in an appropriate trace norm. We denote by  $\mathfrak{S}_p$  the usual Schatten class [33, 30] of operators  $A$  having a finite  $p$ -trace,  $\|A\|_{\mathfrak{S}_p} = \text{Tr}(|A|^p)^{1/p} < \infty$ . Let us recall the Kato-Seiler-Simon inequality (see [31] and Thm 4.1 in [33])

$$\forall p \geq 2, \quad \|f(-i\nabla)g(x)\|_{\mathfrak{S}_p} \leq (2\pi)^{-3/p} \|f\|_{L^p(\mathbb{R}^3)} \|g\|_{L^p(\mathbb{R}^3)}. \quad (2.37)$$

The term  $K(V_1^n)$  is treated as follows:

$$\|K(V_1^n)\|_{\mathfrak{S}_p} \leq \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{d\eta}{(\epsilon^2 + \eta^2)^{1/4}} \|(D^0 + i\eta)^{-1} V_1^n\|_{\mathfrak{S}_p}$$

where we have used that

$$\left\| \frac{|D^0 + V|^{1/2}}{D^0 + V + i\eta} \right\| \leq \frac{1}{(\epsilon^2 + \eta^2)^{1/4}}.$$

By (2.37) we have

$$\begin{aligned} \|(D^0 + i\eta)^{-1}V_1^n\|_{\mathfrak{S}_p} &\leq (2\pi)^{-3/p} \|V_1^n\|_{L^p(\mathbb{R}^3)} \left( \int_{\mathbb{R}^3} \frac{dk}{(1 + |k|^2 + \eta^2)^{p/2}} \right)^{1/p} \\ &\leq \frac{C}{1 + \eta^{1-\frac{3}{p}}} \|V_1^n\|_{L^p(\mathbb{R}^3)}. \end{aligned} \quad (2.38)$$

Since  $6 < p < \infty$ , this finally proves that  $\|K(V_1^n)\|_{\mathfrak{S}_p} \leq C \|V_1^n\|_{L^p(\mathbb{R}^3)}$ , hence this term is a compact operator for any  $n$ .

The term involving  $V_2^n$  is more complicated to handle. First we use the formula [19, 20]

$$\begin{aligned} \int_{-\infty}^{\infty} \frac{1}{D^0 + i\eta} V_2^n \frac{1}{D^0 + i\eta} d\eta &= \int_{-\infty}^{\infty} \frac{1}{D^0 + i\eta} \left[ V_2^n, \frac{1}{D^0 + i\eta} \right] d\eta \\ &= \int_{-\infty}^{\infty} \frac{1}{(D^0 + i\eta)^2} [D^0, V_2^n] \frac{1}{D^0 + i\eta} d\eta \\ &= -i \int_{-\infty}^{\infty} \frac{1}{(D^0 + i\eta)^2} (\alpha \cdot \nabla V_2^n) \frac{1}{D^0 + i\eta} d\eta. \end{aligned}$$

Iterating the resolvent formula we arrive at

$$\begin{aligned} K(V_2^n) &= -\frac{i}{2\pi} \left( \int_{-\infty}^{\infty} \frac{1}{(D^0 + i\eta)^2} (\alpha \cdot \nabla V_2^n) \frac{|D^0|^{1/2}}{D^0 + i\eta} d\eta \right) |D^0|^{-1/2} |D^0 + V|^{1/2} \\ &\quad - \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{1}{D^0 + i\eta} V_2^n \frac{1}{D^0 + i\eta} V \frac{|D^0 + V|^{1/2}}{D^0 + V + i\eta} d\eta. \end{aligned} \quad (2.39)$$

The first term can be estimated as before by (recall that  $|D^0|^{-1/2}|D^0 + V|^{1/2}$  is bounded)

$$\left\| \frac{i}{2\pi} \int_{-\infty}^{\infty} \frac{1}{(D^0 + i\eta)^2} (\alpha \cdot \nabla V_2^n) \frac{|D^0|^{1/2}}{D^0 + i\eta} d\eta \right\|_{\mathfrak{S}_q} \leq C \|\nabla V_2^n\|_{L^q(\mathbb{R}^3)} \int_{-\infty}^{\infty} \frac{d\eta}{1 + \eta^{1+3\frac{q-2}{2}}}$$

which is convergent since  $q > 2$  by assumption.

The next step is to expand the last term of (2.39) using again the resolvent expansion:

$$\begin{aligned} &\int_{-\infty}^{\infty} \frac{1}{D^0 + i\eta} V_2^n \frac{1}{D^0 + i\eta} V \frac{|D^0 + V|^{1/2}}{D^0 + V + i\eta} d\eta \\ &= \sum_{j=1}^{k-1} (-1)^{j+1} \left( \int_{-\infty}^{\infty} \frac{1}{D^0 + i\eta} V_2^n \left( \frac{1}{D^0 + i\eta} V \right)^j \frac{1}{D^0 + i\eta} |D^0|^{1/2} d\eta \right) |D^0|^{-1/2} |D^0 + V|^{1/2} \\ &\quad - (-1)^k \int_{-\infty}^{\infty} \frac{1}{D^0 + i\eta} V_2^n \left( \frac{1}{D^0 + i\eta} V \right)^k \frac{1}{D^0 + V + i\eta} |D^0 + V|^{1/2} d\eta. \end{aligned} \quad (2.40)$$

By (2.38) we see that the last term belongs to  $\mathfrak{S}_{kr}$  when  $k$  is chosen large enough such that  $k(1 - 3/r) > 1/2$  (which is possible since  $r > 3$ ).

We now have to prove that the other terms corresponding to  $j = 2 \dots k - 1$  in (2.40) are also compact. We will only consider the term  $j = 2$ , the others being handled similarly. Writing  $V = V_1^n + V_2^n + V_3^n$  the terms containing  $V_1^n$  and  $V_3^n$  are treated using previous

ideas. For the term which only contains  $V_2^n$ , the idea is, as done previously in [19], to insert  $P_+^0 + P_-^0 = 1$  as follows

$$\int_{-\infty}^{\infty} \frac{P_+^0 + P_-^0}{D^0 + i\eta} V_2^n \frac{P_+^0 + P_-^0}{D^0 + i\eta} V_2^n \frac{P_+^0 + P_-^0}{D^0 + i\eta} |D^0|^{1/2} d\eta.$$

The next step is to expand and note that, by the residuum formula, the ones which contains only  $P_+^0$  or only  $P_-^0$  vanish. Hence we only have to treat terms which contain two different  $P_{\pm}^0$ . We will consider for instance

$$\begin{aligned} & \int_{-\infty}^{\infty} \frac{P_-^0}{D^0 + i\eta} V_2^n \frac{P_+^0}{D^0 + i\eta} V_2^n \frac{P_+^0}{D^0 + i\eta} |D^0|^{1/2} d\eta \\ &= \int_{-\infty}^{\infty} \frac{P_-^0}{D^0 + i\eta} [P_-^0, V_2^n] \frac{P_+^0}{D^0 + i\eta} V_2^n \frac{P_+^0}{D^0 + i\eta} |D^0|^{1/2} d\eta. \end{aligned} \quad (2.41)$$

Now we have using again a Cauchy formula for  $P_-^0$

$$|D^0|^{-1/2} [P_-^0, V_2^n] = -\frac{i}{2\pi} \int_{-\infty}^{\infty} \frac{|D^0|^{-1/2}}{D^0 + i\eta} \sigma \cdot \nabla V_2^n \frac{1}{D^0 + i\eta} d\eta.$$

The Kato-Seiler-Simon inequality (2.37) yields as before

$$\left\| |D^0|^{-1/2} [P_-^0, V_2^n] \right\|_{\mathfrak{S}_q} \leq C \|\nabla V_2^n\| \int_{-\infty}^{\infty} \frac{d\eta}{1 + \eta^{1+3\frac{q-2}{2}}}$$

Inserting this in (2.41) and using that  $V_2^n \in L^r$ , we see that the corresponding operator is compact.

Eventually we have by a trivial estimate

$$\left\| (P_+^0 - \Pi) |D^0 + V|^{1/2} - K(V_1^n) - K(V_2^n) \right\| = \|K(V_3^n)\| \leq C \|V_3^n\|_{L^\infty(\mathbb{R}^3)} \rightarrow_{n \rightarrow \infty} 0.$$

As  $(P_+^0 - \Pi) |D^0 + V|^{1/2}$  is a limit in the operator norm of a sequence of compact operators, it must be compact.  $\square$

We treat separately the case of a Coulomb-type singularity, for which  $(P_+^0 - \Pi) |D^0 + V|^{1/2}$  is not compact, hence we cannot use Theorem 2.2 directly.

**Theorem 2.7 (No pollution in free basis - Coulomb case).** *Let  $|\kappa| < \sqrt{3}/2$ . Then*

$$\text{Spu} \left( D^0 + \frac{\kappa}{|x|}, P_+^0 \right) \cap (-1, 1) = \emptyset.$$

**Proof.** The operators  $P_+^0(D^0 + \kappa|x|^{-1})P_+^0$  and  $P_-^0(D^0 + \kappa|x|^{-1})P_-^0$  are known to have a self-adjoint Friedrichs extension as soon as  $|\kappa| < 2/(\pi/2 + 2/\pi)$ , see [17]. Furthermore one has  $\sigma_{\text{ess}}(D^0 + \kappa|x|^{-1})|_{P_+^0 L^2} = [1, \infty)$  and  $\sigma_{\text{ess}}(D^0 + \kappa|x|^{-1})|_{P_-^0 L^2} = (-\infty, -1]$ , see Theorem 2 in [17]. As  $\sqrt{3}/2 < 2/(\pi/2 + 2/\pi)$ , the result immediately follows from Theorem 2.1 and Remark 2.1.  $\square$

### 3. Balanced basis

In Section 2 we have studied and characterized spectral pollution in the case of a spitting  $\mathfrak{H} = P\mathfrak{H} \oplus (1-P)\mathfrak{H}$  of the main Hilbert space. In particular for the case of the Dirac operator  $D^0 + V$  we have seen that the simple decomposition into upper and lower spinors may yield to pollution as soon as  $V \neq 0$ . In this section we study an abstract theory (inspired of methods used in Physics and Chemistry) in which one tries to avoid pollution by *imposing a relation between the vectors of the basis in  $P\mathfrak{H}$  and in  $(1-P)\mathfrak{H}$* , modelled by one operator  $L : P\mathfrak{H} \rightarrow (1-P)\mathfrak{H}$ . We call such basis a *balanced basis*.

#### 3.1. General results

Consider an orthogonal projection  $P : \mathfrak{H} \rightarrow \mathfrak{H}$ . Let  $L : D(L) \subset P\mathfrak{H} \rightarrow (1-P)\mathfrak{H}$  be a (possibly unbounded) operator which we call *balanced operator*. We assume that

- $L$  is an injection: if  $Lx = 0$  for  $x \in D(L)$ , then  $x = 0$ ;
- $D(L) \oplus LD(L)$  is a core for  $A$ .

**Definition 3.1 (Spurious eigenvalues in balanced basis).** *We say that  $\lambda \in \mathbb{R}$  is a  $(P, L)$ -spurious eigenvalue of the operator  $A$  if there exist a sequence of finite dimensional spaces  $\{V_n^+\}_{n \geq 1} \subset D(L)$  with  $V_n^+ \subset V_{n+1}^+$  for any  $n$ , such that*

- (1)  $\overline{\cup_{n \geq 1} (V_n^+ \oplus LV_n^+)}^{D(A)} = D(A)$ ;
- (2)  $\lim_{n \rightarrow \infty} \text{dist} \left( \lambda, \sigma \left( A|_{(V_n^+ \oplus LV_n^+)} \right) \right) = 0$ ;
- (3)  $\lambda \notin \sigma(A)$ .

We denote by  $\text{Spu}(A, P, L)$  the set of  $(P, L)$ -spurious eigenvalues of the operator  $A$ .

**Remark 3.1.** *Another possible definition would be to only ask that for all  $n$ ,  $V_n^-$  contains  $LV_n^+$ . This would actually also correspond to some methods used by chemists (like the so-called unrestricted kinetic balance [14]). The study of these methods is similar but simpler than the one given by Definition 3.1.*

Contrarily to the previous section, we will not characterize completely  $(P, L)$ -spurious eigenvalues. We will only give some necessary or sufficient conditions which will be enough for the examples we are interested in and which we study in the next section. We will assume as in the previous section that  $PAP$  (resp.  $(1-P)A(1-P)$ ) is essentially self-adjoint on  $D(L)$  (resp. on  $LD(L)$ ) with closure denoted as  $A|_{P\mathfrak{H}}$  (resp.  $A|_{(1-P)\mathfrak{H}}$ ).

##### 3.1.1. Sufficient conditions

We start by exhibiting a very simple part of the polluted spectrum. For any fixed  $0 \neq x \in D(L)$ , we consider the  $2 \times 2$  matrix  $M(x)$  of  $A$  restricted to the 2-dimensional space  $x\mathbb{C} \oplus Lx\mathbb{C}$ , and we denote by  $\mu_1(x) \leq \mu_2(x)$  its eigenvalues. Note that  $\mu_i$  is homogeneous for  $i = 1, 2$ ,  $\mu_i(\lambda x) = \mu_i(x)$ .

**Theorem 3.1 (Pollution in balanced basis - sufficient condition).** *Let  $A, P, L$  as*

before and define  $m_i, M_i \in \mathbb{R} \cup \{\pm\infty\}$ ,  $i = 1, 2$ , as follows:

$$m_1 := \inf_{\substack{\{x_n^+\} \subset D(L) \setminus \{0\}, \\ x_n^+ \|x_n^+\|^{-1} \rightarrow 0, \\ Lx_n^+ \|Lx_n^+\|^{-1} \rightarrow 0}} \liminf_{n \rightarrow \infty} \mu_1(x_n^+), \quad M_1 := \sup_{\substack{\{x_n^+\} \subset D(L) \setminus \{0\}, \\ x_n^+ \|x_n^+\|^{-1} \rightarrow 0, \\ Lx_n^+ \|Lx_n^+\|^{-1} \rightarrow 0}} \limsup_{n \rightarrow \infty} \mu_1(x_n^+), \quad (3.1)$$

$$m_2 := \inf_{\substack{\{x_n^+\} \subset D(L) \setminus \{0\}, \\ x_n^+ \|x_n^+\|^{-1} \rightarrow 0, \\ Lx_n^+ \|Lx_n^+\|^{-1} \rightarrow 0}} \liminf_{n \rightarrow \infty} \mu_2(x_n^+), \quad M_2 := \sup_{\substack{\{x_n^+\} \subset D(L) \setminus \{0\}, \\ x_n^+ \|x_n^+\|^{-1} \rightarrow 0, \\ Lx_n^+ \|Lx_n^+\|^{-1} \rightarrow 0}} \limsup_{n \rightarrow \infty} \mu_2(x_n^+). \quad (3.2)$$

Then we have:

$$[m_1, M_1] \cup [m_2, M_2] \subseteq \overline{\text{Spu}(A, P, L)} \cup \hat{\sigma}_{\text{ess}}(A). \quad (3.3)$$

We supplement the above result by the following

**Remark 3.2.** *The two diagonal elements of the matrix  $A(x_n^+)$  being  $\langle Ax_n^+, x_n^+ \rangle \|x_n^+\|^{-2}$  and  $\langle ALx_n^+, Lx_n^+ \rangle \|Lx_n^+\|^{-2}$ , it is clear that we have*

$$m_2 \geq m_1 \geq \max(\inf \hat{\sigma}_{\text{ess}}(A_{|(1-P)\mathfrak{H}}), \inf \hat{\sigma}_{\text{ess}}(A_{|P\mathfrak{H}})),$$

$$M_1 \leq M_2 \leq \min(\sup \hat{\sigma}_{\text{ess}}(A_{|(1-P)\mathfrak{H}}), \sup \hat{\sigma}_{\text{ess}}(A_{|P\mathfrak{H}})),$$

which is compatible with Theorem 2.1, since we must of course have  $\text{Spu}(A, P, L) \subset \text{Spu}(A, P)$ .

**Proof.** We will use the following

**Lemma 3.1.** *Assume that  $A, P$  and  $L$  are as above. Let  $\{V_n\} \subset D(L)$  be a sequence of  $K$ -dimensional spaces with orthonormal basis  $(x_n^1, \dots, x_n^K)$ . Let  $(y_n^1, \dots, y_n^K)$  be an orthonormal basis of  $LV_n \subset (1-P)\mathfrak{H}$ . We assume that  $x_n^k \rightarrow 0$  and  $y_n^k \rightarrow 0$  weakly for every  $k = 1..K$ , as  $n \rightarrow \infty$ . If  $\lambda \in \mathbb{R}$  is such that  $\lim_{n \rightarrow \infty} \text{dist}(\lambda, \sigma(A_{|V_n \oplus LV_n})) = 0$ , then  $\lambda \in \text{Spu}(A, P, L) \cup \sigma(A)$ .*

The proof of Lemma 3.1 will be omitted, it is very similar to that of Lemma 1.2. We notice that the two sets

$$K_i := \left\{ \mu \in \mathbb{R} \cup \{\pm\infty\} : \exists \{x_n^+\} \subset D(L), x_n^+ \|x_n^+\|^{-1} \rightarrow 0, \right. \\ \left. Lx_n^+ \|Lx_n^+\|^{-1} \rightarrow 0, \mu_i(x_n^+) \rightarrow \mu \right\}. \quad (3.4)$$

are closed convex sets, for  $i = 1, 2$ . Indeed, assume for instance that  $\lambda_1, \lambda_2 \in K_1$  and let be  $\{x_n\}$  and  $\{y_n\}$  such that  $\mu_1(x_n) \rightarrow \lambda_1$  and  $\mu_1(y_n) \rightarrow \lambda_2$ . By the homogeneity of  $\mu_1$  we may assume that  $\|x_n\| = \|y_n\| = 1$  for all  $n$ . Also, extracting a subsequence from  $\{y_n\}$ , we may always assume that

$$\lim_{n \rightarrow \infty} \langle x_n, y_n \rangle = \lim_{n \rightarrow \infty} \left\langle \frac{Lx_n}{\|Lx_n\|}, \frac{Ly_n}{\|Ly_n\|} \right\rangle = 0.$$

Fix some  $\lambda \in (\lambda_1, \lambda_2)$  and consider as usual  $z_n(\theta) = \cos \theta x_n + \sin \theta y_n$ . By continuity of the first eigenvalue of the  $2 \times 2$  matrix of  $A$  in the space spanned by  $z_n(\theta)$  and  $Lz_n(\theta)$ , we know that there exists (for  $n$  large enough) a  $\theta_n \in (0, 2\pi)$  such that  $\mu_1(\theta_n) = \lambda$ . Note

that  $\|z_n(\theta_n)\| = 1 + o(1)$ . Writing  $Lz_n(\theta_n) \|Lz_n(\theta_n)\|^{-1} = \alpha_n Lx_n \|Lx_n\|^{-1} + \beta_n Ly_n \|Ly_n\|^{-1}$  we see that both  $\alpha_n$  and  $\beta_n$  are bounded and satisfy  $\alpha_n^2 + \beta_n^2 \rightarrow 1$ , hence  $\|Lz_n(\theta_n)\| \rightarrow 1$ . It is then clear that  $z_n(\theta_n) \|z_n(\theta_n)\|^{-1} \rightarrow 0$  and that  $Lz_n(\theta_n) \|Lz_n(\theta_n)\|^{-1} \rightarrow 0$ . Therefore  $\lambda = \lim_{n \rightarrow \infty} \mu_2(z_n(\theta_n)) \in K_1$ . The argument is the same for  $K_2$ . As Lemma 3.1 tells us that  $K_1 \cup K_2 \subset \text{Spu}(A, P, L) \cup \sigma(A)$ , this ends the proof of Theorem 3.1.  $\square$

### 3.1.2. Necessary conditions

Let us emphasize that, contrarily to  $P$ -spurious eigenvalues, for  $(P, L)$ -spurious eigenvalues the two spaces  $P\mathfrak{H}$  and  $(1 - P)\mathfrak{H}$  do not play anymore a symmetric role due to the introduction of the operator  $L$ . For this reason we shall concentrate on pollution occurring in the *upper part of the spectrum* and we will not give necessary conditions for the lower part<sup>e</sup>. Loosely speaking, obtaining an information on the lower part would need to study the operator  $L^{-1}$ . In the applications of the next section, we will simply compute the lower polluted spectrum explicitly using Theorem 3.1. Let us introduce

$$d := \sup \sigma(A_{(1-P)\mathfrak{H}}). \quad (3.5)$$

and assume that  $d < \infty$ . In the sequel we will only study  $(P, L)$ -spurious eigenvalues in  $(c, \infty)$ . Note that due to Theorem 2.1, it would be more natural to let instead  $d := \sup \hat{\sigma}_{\text{ess}}(A_{(1-P)\mathfrak{H}})$  but this will actually not change anything for the examples we want to treat: in the Dirac case  $D^0 + V$  and for  $P = \mathcal{P}$ , the orthogonal projector on the upper spinor defined in (2.24), the spectrum of  $(D^0 + V)|_{(1-P)L^2(\mathbb{R}^3, \mathbb{C}^4)} = -1 + V$  is only composed of essential spectrum. We do not know how to handle the case of an operator  $A|_{(1-P)\mathfrak{H}}$  which has a nonempty discrete spectrum above its essential spectrum. Our main result is the following

**Theorem 3.2 (Pollution in balanced basis - necessary conditions).** *Let  $A, P, L$  as before. We recall that the real number  $d < \infty$  was defined in (3.5).*

(i) *Let us define*

$$m_2'' = \inf_{x^+ \in D(L) \setminus \{0\}} \mu_2(x^+) \quad (3.6)$$

and assume that  $m_2'' > d$ . Then we have

$$\text{Spu}(A, P, L) \cap (d, m_2'') = \emptyset.$$

(ii) *Let us define*

$$m_2' = \inf_{\substack{\{x_n^+\} \subset D(L) \setminus \{0\}, \\ x_n^+ \rightarrow 0, \|x_n^+\| = 1}} \liminf_{n \rightarrow \infty} \mu_2(x_n^+) \quad (3.7)$$

and assume that  $m_2' > d$ . We also assume that the following additional continuity property holds for some real number  $b > d$ :

$$\left. \begin{array}{l} \{x_n^+\} \subset D(L) \\ x_n^+ \rightarrow 0 \\ \limsup_{n \rightarrow \infty} \mu_2(x_n^+) < b \end{array} \right\} \implies \langle Ax_n^+, x_n^+ \rangle \rightarrow 0. \quad (3.8)$$

<sup>e</sup>As we have mentioned before we always assume for simplicity that  $\inf \hat{\sigma}_{\text{ess}}(A|_{(1-P)\mathfrak{H}}) \leq \inf \hat{\sigma}_{\text{ess}}(A|_{P\mathfrak{H}})$ , i.e. that  $1 - P$  is responsible from the pollution occurring in the lower part of the spectrum.

Then we have

$$\text{Spu}(A, P, L) \cap (d, \min(m'_2, b)) = \emptyset.$$

**Remark 3.3.** The property (3.8) is a kind of compactness property at 0 of the set  $\{x^+ \in D(L) \mid \mu_2(x^+) < b\}$  for the quadratic-form norm of the operator  $A|_{P\mathfrak{H}}$ .

**Remark 3.4.** Note that (3.8) holds true for  $b = +\infty > d$  when  $A|_{P\mathfrak{H}}$  is a bounded operator.

Theorem 3.2 has many similarities with the characterization of eigenvalues in a gap which was proved by Dolbeault, Esteban and Séré in [12] (where our number  $d = \sup \sigma(A_{(1-P)\mathfrak{H}})$  was denoted by ‘ $a$ ’). In particular the reader should compare the assumptions  $d < m'_2$  and  $d < m''_2$  with (iii) at the bottom of p. 209 in [12]. The proof indeed uses many ideas of [12]. Note that [12] was itself inspired by an important Physics paper of Talman [37] who introduced a minimax principle for the Dirac equation in order to avoid spectral pollution.

**Proof.** Assume that  $\lambda \in \text{Spu}(A, P, L) \cap (d, \infty)$ . We consider a Weyl sequence  $\{x_n\}$  like in Lemma 1.1, i.e. such that

$$P_{V_n^+ \oplus LV_n^+}(A - \lambda_n)x_n = 0 \quad (3.9)$$

for some  $x_n \rightharpoonup 0$  with  $\|x_n\| = 1$  and some  $\lambda_n \rightarrow \lambda$ . We write  $x_n = x_n^+ + x_n^-$  where  $x_n^- = Ly_n^+$  for some  $y_n^+ \in V_n^+$ . Now, like in [12] we consider the following functional defined on  $LV_n^+$ :

$$Q(x^-) := \langle A(x_n^+ + x^-), x_n^+ + x^- \rangle - \lambda_n \|x_n^+ + x^-\|^2.$$

Using the equation  $P_{V_n^+ \oplus LV_n^+}(A - \lambda_n)x_n = 0$ , we deduce that

$$\forall x^- \in LV_n^+, \quad Q(x^-) = \langle (A - \lambda_n)(x^- - x_n^-), x^- - x_n^- \rangle.$$

By definition of  $d$  we obtain

$$\forall x^- \in LV_n^+, \quad Q(x^-) \leq (d - \lambda_n) \|x^- - x_n^-\|^2. \quad (3.10)$$

Consider the  $2 \times 2$  matrix  $M(x_n^+)$  of  $A$  restricted to  $x_n^+ \oplus Lx_n^+$  and recall that  $\mu_2(x_n^+)$  is by definition its second eigenvalue, hence

$$\mu_2(x_n^+) = \sup_{\theta \in \mathbb{R}} \frac{\langle A(x_n^+ + \theta Lx_n^+), x_n^+ + \theta Lx_n^+ \rangle}{\|x_n^+ + \theta Lx_n^+\|^2}$$

(the sup is not necessarily attained). There exists  $\theta_n \in \mathbb{R}$  such that for  $n$  large enough

$$\frac{\langle A(x_n^+ + \theta_n Lx_n^+), x_n^+ + \theta_n Lx_n^+ \rangle}{\|x_n^+ + \theta_n Lx_n^+\|^2} \geq \mu_2(x_n^+) - 1/n. \quad (3.11)$$

Inserting  $x^- = \theta_n Lx_n^+$  in (3.10) we obtain for  $n$  large enough,

$$(\mu_2(x_n^+) - \lambda_n - 1/n) \left( \|x_n^+\|^2 + \theta_n^2 \|Lx_n^+\|^2 \right) + (\lambda_n - d) \|\theta_n Lx_n^+ - x_n^-\|^2 \leq 0. \quad (3.12)$$

Let us assume we are in case (i) for which  $m''_2 > d$ . Using the obvious estimate  $\mu_2(x_n^+) \geq m''_2$  we see that if  $\lambda \in (d, m''_2)$ , then for  $n$  large enough we must have  $x_n^+ = \theta_n Lx_n^+ = \theta_n Lx_n^+ - x_n^- = 0$ , thus  $x_n = 0$  which is a contradiction with  $\|x_n\| = 1$ . Hence  $\text{Spu}(A, P, L) \cap (d, m''_2) = \emptyset$ .

Let us now treat case (ii) for which we assume  $m'_2 > d$  and that (3.8) holds for some  $b > d$ . Let  $\lambda \in \text{Spu}(A, P, L) \cap (d, \min(b, m'_2))$ . From (3.12) we see that necessarily  $\mu_2(x_n^+) \leq \lambda_n + 1/n$  (except if  $x_n = 0$  which is a contradiction). Therefore we have  $\limsup_{n \rightarrow \infty} \mu_2(x_n^+) < b$ .

Assume first that  $x_n^+ \rightarrow 0$  strongly. Using our assumption (3.8), we deduce that  $\lim_{n \rightarrow \infty} \langle Ax_n^+, x_n^+ \rangle = 0$ . Next we argue like in the 3rd step of the proof of Theorem 2.1. First, taking the scalar product of (3.9) with  $x_n^+$ , we deduce that  $\lim_{n \rightarrow \infty} \langle Ax_n^-, x_n^+ \rangle = 0$ . Taking then the scalar product with  $x_n^-$  we deduce that

$$\lim_{n \rightarrow \infty} \langle (A - \lambda_n)x_n^-, x_n^- \rangle = 0.$$

As  $\langle (A - \lambda_n)x_n^-, x_n^- \rangle \leq (d - \lambda + o(1)) \|x_n^-\|^2$  and  $d - \lambda < 0$  we deduce that  $x_n^- \rightarrow 0$  which is a contradiction with  $\|x_n\| = 1$ .

Hence we must have  $x_n^+ \not\rightarrow 0$ , which implies that  $x_n^+ \|x_n^+\|^{-1} \rightharpoonup 0$ , up to a subsequence. Therefore we have  $\liminf_{n \rightarrow \infty} \mu_2(x_n^+) = \liminf_{n \rightarrow \infty} \mu_2(x_n^+ \|x_n^+\|^{-1}) \geq m'_2$  by definition of  $m'_2$ . Inserting this information in (3.12), we again arrive at a contradiction, similarly as before. This ends the proof of Theorem 3.2.  $\square$

### 3.2. Application to Dirac operator

In this section, we consider the Dirac operator  $A = D^0 + V$  for a potential satisfying the assumptions (2.25) and (2.26) of Theorem 2.4 and

$$\sup(V) < 2 \tag{3.13}$$

We will indeed for simplicity concentrate ourselves on the case for which either  $V$  is bounded, or  $V$  is a purely attractive Coulomb potential,  $V(x) = -\kappa/|x|$ ,  $0 < \kappa < \sqrt{3}/2$ . The generalization to potentials having several singularities is rather straightforward.

Like in Section 2.3.2, we start by choosing  $P = \mathcal{P}$ , the projector on the upper spinors as defined in (2.24). As already noticed in Section 2.3.2 we then have  $\mathcal{P}A\mathcal{P} = 1 + V$  and  $(1 - \mathcal{P})A(1 - \mathcal{P}) = -1 + V$  on the appropriate domain. This shows that the number  $d$  introduced in the previous sections is  $d = -1 + \sup V < 1$  by (3.13).

We will now study different balanced operators  $L$  which we have found in the Quantum Chemistry literature. Note that we can always see  $L$  as an operator defined on 2-spinors  $D(L) \subset L^2(\mathbb{R}^3, \mathbb{C}^2)$  with values in the same Hilbert space  $L^2(\mathbb{R}^3, \mathbb{C}^2)$ , which we will do in the rest of the paper.

We will describe the polluted spectrum  $\text{Spu}(D^0 + V, \mathcal{P}, L)$  using the results presented in the previous sections. We note that the number  $\mu_2(\varphi)$  is the largest solution to the following equation [12]

$$\langle (1 + V)\varphi, \varphi \rangle + \frac{(\Re \langle L\varphi, \sigma \cdot (-i\nabla)\varphi \rangle)^2}{\langle (\mu + 1 - V)L\varphi, L\varphi \rangle} = \mu \|\varphi\|^2 \tag{3.14}$$

where the denominator of the second term does not vanish when  $\mu_2(\varphi) > d = \sup(V) - 1$ . Note the term on the left is decreasing with respect to  $\mu$ , whereas the term on the right is increasing with respect to  $\mu$ . Hence we have  $\mu_2(\varphi) \geq 1$  if and only if

$$\langle V\varphi, \varphi \rangle + \frac{(\Re \langle L\varphi, \sigma \cdot (-i\nabla)\varphi \rangle)^2}{\langle (2 - V)L\varphi, L\varphi \rangle} \geq 0 \tag{3.15}$$

where the denominator of the second term does not vanish due to (3.13). Note that (3.15) takes the form of a Hardy-type inequality similar to those which were found in [12, 11]. In the following we will have to study this kind of inequalities for sequences  $\varphi_n$  which converge weakly to 0. The Hardy inequalities of [12, 11] will indeed be an important tool as we will see below.

Concerning the choice of the operator  $L$ , several possibilities exist, although the main method is without any doubt the so-called *kinetic balance* which we will study in the next section. All the methods from Quantum Chemistry or Physics are based on the following formula for an eigenfunction  $(\varphi, \chi)$  with eigenvalue  $mc^2 + \lambda$  (we reintroduce the speed of light  $c$  and the mass  $m$  for convenience) and which we have already formally derived before in Section 2.3.2:

$$\chi = \frac{c}{2mc^2 + \lambda - V} \sigma \cdot (-i\nabla) \varphi. \quad (3.16)$$

This equation suggests that for an eigenvector to be represented correctly, the basis of the lower spinor should contain  $c(2mc^2 + \lambda - V)^{-1} \sigma \cdot (-i\nabla)$  applied to the elements of the basis for the upper spinor. However we cannot choose in principle  $L = c(2mc^2 + \lambda - V)^{-1} \sigma \cdot (-i\nabla)$  because  $\lambda$  is simply unknown. For this reason, one often takes the first order approximation in the nonrelativistic limit which is nothing but

$$L_{KB} = \frac{1}{2mc} \sigma \cdot (-i\nabla).$$

The choice of this balanced operator is (by far) the most widespread method in Quantum Physics and Chemistry. It will be studied in details in Section 3.2.1.

It seems a well-known fact in Quantum Chemistry and Physics [14, 27] that the kinetic balance method consisting in choosing  $L = L_{KB}$  is not well-behaved for pointwise nuclei. The reason is that the behaviour at zero of  $c(2mc^2 + \lambda - V)^{-1} \sigma \cdot (-i\nabla)$  is not properly captured by  $\sigma \cdot (-i\nabla)$ , if  $V(x) = -\kappa|x|^{-1}$ . Indeed we will prove that the kinetic balance method allows to avoid pollution in the upper part of the spectrum for ‘regular’ potentials, but not for Coulomb potentials, which justifies the aforementioned intuition.

To better capture the behaviour at zero, we study another method in Section 3.2.2 which we call *atomic balance*<sup>f</sup> and which consists in choosing

$$L_{AB} = \frac{c}{2mc^2 - V} \sigma \cdot (-i\nabla).$$

Although this operator does not depend on  $\lambda$ , it will be shown to completely avoid pollution in the upper part of the spectrum, even for Coulomb potentials. It is very likely that any other reasonable choice with the same behaviour at zero would have the same effect but we have not studied this question more deeply.

In the following we again work in units for which  $m = c = 1$ .

### 3.2.1. Kinetic Balance

The most common method is the so-called *kinetic balance* [13, 18, 22, 34]. It consists in choosing as balanced operator

$$\boxed{L_{KB} = -i\sigma \cdot \nabla} \quad (3.17)$$

We can for instance define  $L_{KB}$  on the domain  $D(L_{KB}) = C_0^\infty(\mathbb{R}^3, \mathbb{C}^2)$ , in which case  $L_{KB}$  satisfies all the assumptions of Section 3. Our main result is the following

**Theorem 3.3 (Kinetic Balance).** *(i) Bounded potential.* Assume that  $V \in L^p(\mathbb{R}^3)$  for some  $p > 3$ , that  $\lim_{|x| \rightarrow \infty} V(x) = 0$ , and that

$$-1 + \sup(V) < 1 + \inf(V). \quad (3.18)$$

<sup>f</sup>The relation (2.22) is usually called *exact atomic balance*.

Then we have

$$\overline{\text{Spu}(D^0 + V, \mathcal{P}, L_{KB})} = [-1, -1 + \sup V].$$

(ii) **Coulomb potential.** Assume that  $0 < \kappa < \sqrt{3}/2$ . Then we have

$$\overline{\text{Spu}\left(D^0 - \frac{\kappa}{|x|}, \mathcal{P}, L_{KB}\right)} = [-1, 1]. \quad (3.19)$$

**Remark 3.5.** The conclusion (3.19) also holds if  $V$  is such that  $V \in L^p(\mathbb{R}^3) \cap L^\infty(\mathbb{R}^3 \setminus B(x_0, r))$  for some  $p > 3$  and

$$-\frac{\kappa}{|x - x_0|} \leq V(x) \leq -\frac{\kappa'}{|x - x_0|^a} \text{ on } B(x_0, r)$$

for some  $0 < a \leq 1$  and some  $\kappa < \sqrt{3}/2$ , as is obviously seen from the proof.

We have proved that the widely used kinetic balance method allows to *avoid pollution in the upper part of the gap for smooth potentials*, hence for instance for  $V = -\rho * |x|^{-1}$  where  $\rho \geq 0$  is the distribution of charge for smeared nuclei. However, the kinetic balance method *does not avoid spectral pollution in the case of pointwise nuclei* (Coulomb potential).

**Proof.**

**Case (i).** We assume that  $V \in L^p(\mathbb{R}^3) \cap L^\infty(\mathbb{R}^3)$  satisfies (3.18). Clearly we have  $\sup_\varphi \mu_1(\varphi) \leq -1 + \sup(V) =: d$  and  $m_2'' = \inf_\varphi \mu_2(\varphi) \geq 1 + \inf(V)$ . Hence we necessarily have  $m_2' \geq m_2'' > d$  as requested by Theorem 3.2. Also since  $V$  is bounded by assumption,  $(D^0 + V)|_{\mathcal{P}L^2(\mathbb{R}^3, \mathbb{C}^4)} = 1 + V$  is bounded, hence (3.8) holds for  $b = 1$ . We deduce that

$$\text{Spu}(D^0 + V, \mathcal{P}, L_{KB}) \cap (c, 1) \subset [m_2', 1).$$

Now we claim that  $m_2' \geq 1$ . Indeed, let us argue by contradiction and assume that there exists a sequence  $\varphi_n \in C_0^\infty(\mathbb{R}^3, \mathbb{C}^2)$  such that  $\varphi_n \rightharpoonup 0$  in  $L^2$ ,  $\|\varphi_n\| = 1$  and  $\mu_2(\varphi_n) \rightarrow \lambda \in (c, 1)$ . The number  $\mu_2(\varphi_n)$  is characterized by the equality

$$\int_{\mathbb{R}^3} V|\varphi_n|^2 + \frac{\left(\int_{\mathbb{R}^3} |\sigma \cdot \nabla \varphi_n|^2\right)^2}{\int_{\mathbb{R}^3} (\mu_2(\varphi_n) + 1 - V)|\sigma \cdot \nabla \varphi_n|^2} = \mu_2(\varphi_n) - 1. \quad (3.20)$$

Since  $V$  is bounded and  $\|\varphi_n\| = 1$  we get

$$\left| \int_{\mathbb{R}^3} |\sigma \cdot \nabla \varphi_n|^2 \right|^2 \leq (1 - \lambda + o(1) + \|V\|_\infty)(1 - \lambda + o(1) + \|V\|_\infty) \int_{\mathbb{R}^3} |\sigma \cdot \nabla \varphi_n|^2$$

which proves that  $\{\varphi_n\}$  is bounded in  $H^1(\mathbb{R}^3, \mathbb{C}^2)$ . We deduce that  $\varphi_n \rightharpoonup 0$  in  $L^p(\mathbb{R}^3, \mathbb{C}^2)$  weakly for all  $2 \leq p \leq 6$  and strongly in  $L_{\text{loc}}^p(\mathbb{R}^3, \mathbb{C}^2)$  for all  $2 \leq p < 6$ . Under our assumption on  $V$ , this shows that  $\lim_{n \rightarrow \infty} \int V|\varphi_n|^2 = 0$ . For  $n$  large enough, we thus have

$$\frac{\left(\int_{\mathbb{R}^3} |\sigma \cdot \nabla \varphi_n|^2\right)^2}{\int_{\mathbb{R}^3} (\mu_2(\varphi_n) + 1 - V)|\sigma \cdot \nabla \varphi_n|^2} \leq \frac{\lambda - 1}{2} < 0 \quad (3.21)$$

which is a contradiction since by assumption  $\mu_2(\varphi_n) = \lambda + o(1) > d \geq V - 1$ . Hence we have proved that  $\text{Spu}(D^0 + V, \mathcal{P}, L_{KB}) \cap (d, 1) = \emptyset$ .

Now we assume  $\sup(V) > 0$  (otherwise there is nothing else to prove since  $d = -1$ ) and prove that  $(-1, -1 + \sup(V)] \subset \overline{\text{Spu}(D^0 + V, \mathcal{P}, L_{KB})}$ . Let  $x_0$  be a Lebesgue point of  $V$ , with  $V(x_0) > 0$  (hence  $V(x) \geq 0$  on a neighborhood of  $x_0$ ). Consider a smooth radial nonnegative function  $\zeta$  which is equal to 1 on the annulus  $\{2 \leq |x| \leq 3\}$  and 0 outside the annulus  $\{1 \leq |x| \leq 4\}$ . We define for some fixed  $\delta > 0$

$$\varphi_n(x) = \left( n^{1/2} \zeta(n(x - x_0)) + \frac{\delta^{1/2}}{(4n)^{3/2}} \zeta\left(\frac{x - x_0}{4n}\right) \right) \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

where we have chosen the scaling in such a way that the above two functions have a disjoint support. We note that

$$\int |\varphi_n|^2 = \delta N + O(n^{-2}), \quad \int |\sigma \cdot \nabla \varphi_n|^2 = D + O(\delta n^{-2})$$

where we have introduced  $N := \int \zeta^2$  and  $D = \int |\nabla \zeta|^2$ . Similarly, we have, using (2.35) and that  $V \rightarrow 0$  at infinity,

$$\langle (1 + V)\varphi_n, \varphi_n \rangle = \delta(N + o(1)) + O(n^{-2}),$$

$$\langle (-1 + V)L_{KB}\varphi_n, L_{KB}\varphi_n \rangle = (-1 + V(x_0))D + O(n^{-2}).$$

Hence the matrix of  $D^0 + V$  in the basis  $\{(\varphi_n, 0), (0, L_{KB}\varphi_n)\}$  converges as  $n \rightarrow \infty$  towards the following  $2 \times 2$  matrix:

$$\begin{pmatrix} 1 & (\frac{D}{N\delta})^{1/2} \\ (\frac{D}{N\delta})^{1/2} & -1 + V(x_0) \end{pmatrix}.$$

Eventually we note that  $\varphi_n \|\varphi_n\|^{-1} \rightarrow 0$  and  $\sigma \cdot \nabla \varphi_n \|\sigma \cdot \nabla \varphi_n\|^{-1} \rightarrow 0$ . Hence, varying  $\delta$  and  $x_0$ , we see that  $M_1 = -1 + \sup(V)$  and  $m_1 \leq -1$  where  $m_1$  and  $M_1$  were defined in (3.1). This ends the proof of (i), by Theorem 3.1.

**Case (ii).** We will use again Theorem 3.1. More precisely we will show that  $m_2 = -\infty < -1$  and  $M_2 \geq 1$ , where  $m_2$  and  $M_2$  have been defined in (3.2). This time we define

$$\varphi_n(x) = \left( n^{1/2} \zeta(nx) + (\delta n)^{1/2} \zeta(\delta nx) \right) \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad (3.22)$$

where  $\delta \geq 4$  is a fixed constant (note the above two functions then have a disjoint support). Similarly as before, we compute

$$\int |\varphi_n|^2 = \frac{1 + \delta^{-2}}{n^2} N, \quad \int |\sigma \cdot \nabla \varphi_n|^2 = 2D,$$

$$\langle (1 + V)\varphi_n, \varphi_n \rangle = \frac{1 + \delta^{-2}}{n^2} N - \kappa \frac{1 + \delta^{-1}}{n} C_1,$$

$$\langle (-1 + V)L_{KB}\varphi_n, L_{KB}\varphi_n \rangle = -2D - \kappa(1 + \delta)nC_2,$$

where  $N$  and  $D$  are defined as above and

$$C_1 = \int_{\mathbb{R}^3} \frac{|\zeta(x)|^2}{|x|} dx, \quad C_2 = \int_{\mathbb{R}^3} \frac{|\sigma \cdot \nabla \zeta(x)|^2}{|x|} dx.$$

Hence, the matrix of  $D^0 - \kappa|x|^{-1}$  in the associated basis reads

$$A_n(\delta) := \begin{pmatrix} 1 - \kappa n \frac{1+\delta^{-1}}{1+\delta^{-2}} \frac{C_1}{N} & n \left( \frac{2D}{(1+\delta^{-2})N} \right)^{1/2} \\ n \left( \frac{2D}{(1+\delta^{-2})N} \right)^{1/2} & -1 - \kappa(1+\delta)n \frac{C_2}{2D} \end{pmatrix}.$$

Let us now choose  $\delta \geq 4$  large enough such that  $\kappa^2(1+\delta^{-1})(1+\delta)C_1C_2 - 2D^2 > 0$ . Then

$$\det(A_n(\delta)) = \frac{\kappa^2(1+\delta^{-1})(1+\delta)C_1C_2 - 2D^2}{(1+\delta^{-2})ND} n^2 + O(n) \quad (3.23)$$

hence  $\det(A_n(\delta)) \rightarrow +\infty$  as  $n \rightarrow \infty$ . Note that the first eigenvalue  $\mu_1(\varphi_n)$  of  $A_n(\delta)$  satisfies

$$\mu_1(\varphi_n) \leq -1 - \kappa(1+\delta)n \frac{C_2}{2D}$$

hence  $\mu_1(\varphi_n) \rightarrow -\infty$  as  $n \rightarrow \infty$ . Therefore we must have  $\mu_2(\varphi_n) < 0$  for  $n$  large enough. More precisely

$$\mu_1(\varphi_n) \geq -1 - \kappa(1+\delta)n \frac{C_2}{2D} - n \left( \frac{2D}{(1+\delta^{-2})N} \right)^{1/2}$$

therefore, multiplying by  $\mu_2(\varphi_n)$  and using (3.23) we deduce that

$$\mu_2(\varphi_n) \leq -\frac{\kappa^2(1+\delta^{-1})(1+\delta)C_1C_2 - 2D^2}{\kappa(1+\delta)(1+\delta^{-2})C_2N/2 + D(2(1+\delta^{-2})N)^{1/2}} n + O(1),$$

which eventually proves that  $\mu_2(\varphi_n) \rightarrow -\infty$ . As it is clear that  $\varphi_n \|\varphi_n\|^{-1} \rightarrow 0$  and  $\sigma \cdot \nabla \varphi_n \|\sigma \cdot \nabla \varphi_n\|^{-1} \rightarrow 0$ , we have shown that  $m_2 = -\infty$ .

The proof that  $M_2 \geq 1$  is simpler, it suffices to use

$$\varphi_n(x) = n^{-3/2} \zeta \left( \frac{x}{n} \right)$$

whose associated matrix of  $A$  reads

$$A_n := \begin{pmatrix} 1 - \kappa \frac{C_1}{Nn} & \sqrt{\frac{D}{N}} \frac{1}{n} \\ \sqrt{\frac{D}{N}} \frac{1}{n} & -1 - \kappa \frac{C_2}{Dn} \end{pmatrix}.$$

Therefore the result follows from Theorem 3.1.  $\square$

### 3.2.2. Atomic Balance

We have proved in the previous section that the kinetic balance method allows to avoid spectral pollution in the case of a smooth potential, but that it does not solve the pollution issue for a Coulomb potential. In this section we consider another method called *atomic balance*. It consists in taking

$$\boxed{L_{AB} = \frac{1}{2-V} \sigma \cdot (-i\nabla)} \quad (3.24)$$

where we recall that we have assumed  $2 > \sup(V)$ . Provided that  $V$  is smooth enough, we can define  $L_{AB}$  on the domain  $D(L_{AB}) = C_0^\infty(\mathbb{R}^3 \setminus \{0\}, \mathbb{C}^2)$ , in which case  $L_{AB}$  satisfies all the assumptions of Section 3. Our main result is the following

**Theorem 3.4 (Atomic Balance).** *Let  $V$  be such that  $\sup(V) < 2$ ,  $(2-V)^{-2}\nabla V \in L^\infty(\mathbb{R}^3)$  and*

$$-\frac{\kappa}{|x|} \leq V(x)$$

for some  $0 \leq \kappa < \sqrt{3}/2$ . We also assume that the positive part  $\max(V, 0)$  is in  $L^p(\mathbb{R}^3)$  for some  $p > 3$  and that  $\lim_{|x| \rightarrow \infty} V(x) = 0$ . Then we have

$$\overline{\text{Spu}(D^0 + V, \mathcal{P}, L_{AB})} = [-1, -1 + \sup V].$$

**Remark 3.6.** We define the operator  $L_{AB}$  on  $D(L_{AB}) = C_0^\infty(\mathbb{R}^3 \setminus \{0\}, \mathbb{C}^2)$ . Note that under our assumptions on  $V$  we have that  $L_{AB}D(L_{AB})$  is dense in  $H^1(\mathbb{R}^3, \mathbb{C}^2)$  for the associated Sobolev norm, hence  $L_{AB}$  satisfies the properties required in Section 3.1.2. The above conditions on  $V$  are probably far from being optimal.

**Remark 3.7.** The choice of ‘2’ in the definition of  $L_{AB}$  is somewhat arbitrary. It can be seen that our result still holds true if  $\sup(V) < 1$  and  $L_{AB}$  is replaced by  $(\theta - V)^{-1}\sigma \cdot p$  for some fixed  $\theta \geq 1$ . The proof is the same when  $\theta \geq 2$  but it is slightly more technical when  $1 \leq \theta < 2$ .

As we will explain in the proof, a very important tool is the Hardy-type inequality:

$$\int_{\mathbb{R}^3} \frac{c^2 |\sigma \cdot \nabla \varphi(x)|^2}{c^2 + \frac{\nu}{|x|} + \sqrt{c^4 - \nu^2 c^2}} dx + (c^2 - \sqrt{c^4 - \nu^2 c^2}) \int_{\mathbb{R}^3} |\varphi(x)|^2 dx \geq \nu \int_{\mathbb{R}^3} \frac{|\varphi(x)|^2}{|x|} dx. \quad (3.25)$$

This inequality was obtained in [12] by using a min-max characterization of the first eigenvalue of  $-ic\alpha \cdot \nabla + c^2\beta - \nu/|x|$ . Indeed (3.25) is an equality when  $\varphi$  is equal to the upper spinor of the eigenfunction corresponding to the first eigenvalue in  $(-1, 1)$  of  $-ic\alpha \cdot \nabla + c^2\beta - \nu/|x|$ . The inequality (3.25) was then proved by a direct analytical method in [11]. Introducing  $m = c(1 + \sqrt{1 - (\nu/c)^2})$  and  $\kappa = \nu/c$  we can rewrite (3.25) in the following form

$$\int_{\mathbb{R}^3} \frac{|\sigma \cdot \nabla \varphi(x)|^2}{m + \frac{\kappa}{|x|}} dx + m \frac{1 - \sqrt{1 - \kappa^2}}{1 + \sqrt{1 - \kappa^2}} \int_{\mathbb{R}^3} |\varphi(x)|^2 dx \geq \kappa \int_{\mathbb{R}^3} \frac{|\varphi(x)|^2}{|x|} dx. \quad (3.26)$$

We now provide the proof of Theorem 3.4.

**Proof.** Let us first prove that when  $\sup(V) > 0$ , then we have  $(-1, -1 + \sup V] \subset \overline{\text{Spu}(D^0 + V, \mathcal{P}, L_{AB})}$ . The proof is indeed the same as that of Theorem 3.3: we define for some fixed  $\delta > 0$

$$\varphi_n(x) = \left( n^{1/2} \zeta(n(x - x_0)) + \frac{\delta^{1/2}}{(4n)^{3/2}} \zeta\left(\frac{x - x_0}{4n}\right) \right) \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

where  $x_0$  is a Lebesgue point of  $V$  such that  $0 < V(x_0) < 2$ . One can prove that the matrix of  $D^0 + V$  in  $\{(\varphi_n, 0), (0, L_{AB}\varphi_n)\}$  converges as  $n \rightarrow \infty$  towards the following  $2 \times 2$  matrix:

$$\begin{pmatrix} 1 & (\frac{D}{N\delta})^{1/2} \\ (\frac{D}{N\delta})^{1/2} & -1 + V(x_0) \end{pmatrix}.$$

Hence we have again, by Theorem 3.1,  $(-1, -1 + \sup V] \subset \overline{\text{Spu}(D^0 + V, \mathcal{P}, L_{AB})}$ .

The second part consists in proving that there is no spectral pollution above  $-1 + \sup(V)$ . As a first illustration of the usefulness of the Hardy-type inequality (3.26), we start by proving the following

**Lemma 3.2.** *We have*

$$m_2'' = \inf_{\varphi \in D(L_{AB})} \mu_2(\varphi) \geq 1 - 2 \frac{1 - \sqrt{1 - \kappa^2}}{1 + \sqrt{1 - \kappa^2}}. \quad (3.27)$$

**Remark 3.8.** We note that the right hand side of (3.27) is always  $\geq 1/3$  when  $0 \leq \kappa < \sqrt{3}/2$ , and it converges to 1 as  $\kappa \rightarrow 0$ , as it should be.

**Proof.** The number  $\mu_2(\varphi)$  is the largest solution of the equation

$$\int_{\mathbb{R}^3} (1 + V(x)) |\varphi(x)|^2 + \frac{\left( \int_{\mathbb{R}^3} \frac{|\sigma \cdot \nabla \varphi(x)|^2}{2 - V(x)} dx \right)^2}{\int_{\mathbb{R}^3} \frac{(1 + \mu - V(x)) |\sigma \cdot \nabla \varphi(x)|^2}{(2 - V(x))^2} dx} = \mu \int_{\mathbb{R}^3} |\varphi(x)|^2 dx. \quad (3.28)$$

Clearly we must always have

$$\mu_2(\varphi) > \mu_c(\varphi) := -1 + \frac{\int_{\mathbb{R}^3} \frac{V(x)}{(2 - V(x))^2} |\sigma \cdot \nabla \varphi(x)|^2 dx}{\int_{\mathbb{R}^3} \frac{|\sigma \cdot \nabla \varphi(x)|^2}{(2 - V(x))^2} dx}.$$

Let be  $\mu_c(\varphi) < \mu < 1$ . We estimate:

$$\begin{aligned} \int_{\mathbb{R}^3} (1 + V(x) - \mu) |\varphi(x)|^2 dx + \frac{\left( \int_{\mathbb{R}^3} \frac{|\sigma \cdot \nabla \varphi(x)|^2}{2 - V(x)} dx \right)^2}{\int_{\mathbb{R}^3} \frac{(1 + \mu - V(x)) |\sigma \cdot \nabla \varphi(x)|^2}{(2 - V(x))^2} dx} \\ \geq \int_{\mathbb{R}^3} (1 + V(x) - \mu) |\varphi(x)|^2 dx + \int_{\mathbb{R}^3} \frac{|\sigma \cdot \nabla \varphi(x)|^2}{2 + \frac{\kappa}{|x|}} dx \\ \geq \left( 1 - 2 \frac{1 - \sqrt{1 - \kappa^2}}{1 + \sqrt{1 - \kappa^2}} - \mu \right) \int_{\mathbb{R}^3} |\varphi(x)|^2 dx \end{aligned} \quad (3.29)$$

where in the last line we have used (3.26) and the fact that  $\kappa|x|^{-1} + V(x) \geq 0$ . From this we deduce that

$$\mu_2(\varphi) \geq \max \left( 1 - 2 \frac{1 - \sqrt{1 - \kappa^2}}{1 + \sqrt{1 - \kappa^2}}, \mu_c(\varphi) \right).$$

This ends the proof of Lemma 3.2 □

The next step is to prove that property (3.8) is satisfied.

**Lemma 3.3.** Property (3.8) holds true for  $b = \infty$ : if  $\{\varphi_n\} \subset C_0^\infty(\mathbb{R}^3, \mathbb{C}^2)$  is such that  $\varphi_n \rightarrow 0$  in  $L^2$  and  $\mu_2(\varphi_n) \rightarrow \ell < \infty$ , then  $\int_{\mathbb{R}^3} V |\varphi_n|^2 \rightarrow 0$ .

**Proof.** Note that necessarily  $\ell \geq 1/3$  by Lemma 3.2, hence  $\ell$  must be finite. We use the estimate (3.29), with  $\mu = \mu_2(\varphi_n)$  to get

$$0 \geq \int_{\mathbb{R}^3} (1 + V(x) - \mu_2(\varphi_n)) |\varphi_n(x)|^2 dx + \int_{\mathbb{R}^3} \frac{|\sigma \cdot \nabla \varphi_n(x)|^2}{2 + \frac{\kappa}{|x|}} dx. \quad (3.30)$$

Now we write

$$\begin{aligned} \int_{\mathbb{R}^3} \frac{|\sigma \cdot \nabla \varphi(x)|^2}{2 + \frac{\kappa}{|x|}} dx &= (1 - \kappa^2) \int_{\mathbb{R}^3} \frac{|\sigma \cdot \nabla \varphi(x)|^2}{2 + \frac{\kappa}{|x|}} dx + \kappa \int_{\mathbb{R}^3} \frac{|\sigma \cdot \nabla \varphi(x)|^2}{\frac{2}{\kappa} + \frac{1}{|x|}} dx \\ &\geq (1 - \kappa^2) \int_{\mathbb{R}^3} \frac{|\sigma \cdot \nabla \varphi(x)|^2}{2 + \frac{\kappa}{|x|}} dx + \kappa \int_{\mathbb{R}^3} \frac{|\varphi(x)|^2}{|x|} dx - \int_{\mathbb{R}^3} |\varphi(x)|^2 dx \end{aligned} \quad (3.31)$$

where we have used (3.26) with  $m \leftrightarrow 2/\kappa$  and  $\kappa \leftrightarrow 1$ . We deduce that

$$\int_{\mathbb{R}^3} \left( \frac{\kappa}{|x|} + V(x) \right) |\varphi_n(x)|^2 dx + (1 - \kappa^2) \int_{\mathbb{R}^3} \frac{|\sigma \cdot \nabla \varphi_n(x)|^2}{2 + \frac{\kappa}{|x|}} dx \leq \mu_2(\varphi_n) \int_{\mathbb{R}^3} |\varphi_n|^2. \quad (3.32)$$

Using that  $\mu_2(\varphi_n) \rightarrow \ell$ , that  $V \geq -\kappa|x|^{-1}$  and  $\varphi_n \rightarrow 0$  we deduce that

$$\lim_{n \rightarrow \infty} \int_{\mathbb{R}^3} \frac{|\sigma \cdot \nabla \varphi_n(x)|^2}{2 + \frac{\kappa}{|x|}} dx = 0.$$

Using again (3.31) with  $\varphi = \varphi_n$  we finally get the result.  $\square$

We will now prove the following

**Lemma 3.4.** *We have  $m'_2 \geq 1$  where  $m'_2$  was defined in (3.7).*

**Proof.** Consider a sequence  $\{\varphi_n\} \subset C_0^\infty(\mathbb{R}^3, \mathbb{C}^2)$  such that  $\|\varphi_n\| = 1$  and  $\varphi_n \rightarrow 0$ . We will argue by contradiction and suppose that, up to a subsequence,  $\mu_2(\varphi_n) \rightarrow \ell \in [1/3, 1)$ . Similarly as in the proof of Lemma 3.3,  $\{\varphi_n\}$  must satisfy (3.32), from which we infer that

$$\int_{\mathbb{R}^3} \frac{|\sigma \cdot \nabla \varphi_n(x)|^2}{2 + \frac{\kappa}{|x|}} dx \leq C,$$

hence  $\{\varphi_n\}$  is bounded in  $H^1$ . Therefore, up to a subsequence we may assume that  $\varphi_n \rightarrow 0$  strongly in  $L_{loc}^p(\mathbb{R}^3)$  for  $2 \leq p < 6$ .

Let us now fix a smooth partition of unity  $\xi_0^2 + \xi_1^2 + \xi_2^2 = 1$  where each  $\xi_i$  is  $\geq 0$ ,  $\xi_0 \equiv 1$  on the ball  $B(0, r)$  and  $\xi_0 \equiv 0$  outside the ball  $B(0, 2r)$ ,  $\xi_2 \equiv 1$  outside the ball  $B(0, 2R)$  and  $\xi_2 \equiv 0$  in the ball  $B(0, R)$ . We fix  $R$  large enough such that

$$\forall |x| \geq R, \quad |V(x)| \leq \frac{1 - \ell}{3}$$

and  $r$  small enough such that

$$m - \epsilon \leq \frac{\epsilon}{2r}$$

where  $\epsilon$  is a fixed constant chosen such that  $1 - \ell - \epsilon/3 > (1 - \ell)/3$  and  $\kappa + \epsilon < \sqrt{3}/2$ .

Next we use the (pointwise) IMS formula

$$|\nabla \varphi(x)|^2 = \sum_{i=0}^2 |\nabla(\xi_i \varphi)(x)|^2 - |\varphi(x)|^2 \sum_{i=0}^2 |\nabla \xi_i(x)|^2$$

and (3.30) to infer, denoting  $\varphi_n^i := \varphi_n \xi_i$  and  $\eta = \sum_{i=0}^2 |\nabla \xi_i(x)|^2$ ,

$$\begin{aligned} \sum_{i=0}^2 \left( \int_{\mathbb{R}^3} (1 + V(x) - \mu_2(\varphi_n)) |\varphi_n^i(x)|^2 dx + \int_{\mathbb{R}^3} \frac{|\sigma \cdot \nabla \varphi_n^i(x)|^2}{2 + \frac{\kappa}{|x|}} dx \right) \\ \leq \int_{\mathbb{R}^3} \frac{\eta(x) |\varphi_n(x)|^2}{2 + \frac{\kappa}{|x|}} dx. \end{aligned} \quad (3.33)$$

Next we note that for  $n$  large enough, by our definition of  $R$ ,

$$\int_{\mathbb{R}^3} (1 + V(x) - \mu_2(\varphi_n)) |\varphi_n^2(x)|^2 dx \geq \frac{1 - \ell}{3} \|\varphi_n^2\|^2. \quad (3.34)$$

Similarly we have by definition of  $r$  and  $\epsilon$  (using that  $\varphi_n^0$  has its support in the ball  $B(0, 2r)$ )

$$\int_{\mathbb{R}^3} \frac{|\sigma \cdot \nabla \varphi_n^0(x)|^2}{2 + \frac{\kappa}{|x|}} dx \geq \int_{\mathbb{R}^3} \frac{|\sigma \cdot \nabla \varphi_n^0(x)|^2}{\epsilon + \frac{\kappa + \epsilon}{|x|}} dx \geq \kappa \int_{\mathbb{R}^3} \frac{|\varphi_n^0(x)|^2}{|x|} - \frac{\epsilon}{3} \int_{\mathbb{R}^3} |\varphi_n^0(x)|^2 dx$$

where for the last inequality we have used (3.26) and  $\kappa + \epsilon < \sqrt{3}/2$ . Using again that  $V \geq -\kappa|x|^{-1}$ , we infer the lower bound, for  $n$  large enough,

$$\int_{\mathbb{R}^3} (1 + V(x) - \mu_2(\varphi_n)) |\varphi_n^0(x)|^2 dx + \int_{\mathbb{R}^3} \frac{|\sigma \cdot \nabla \varphi_n^0(x)|^2}{2 + \frac{\kappa}{|x|}} dx \geq \frac{1 - \ell}{3} \|\varphi_n^0\|^2. \quad (3.35)$$

Inserting (3.34) and (3.35) in (3.33), we obtain

$$\frac{1 - \ell}{3} \left( \|\varphi_n^2\|^2 + \|\varphi_n^0\|^2 \right) \leq \int_{\mathbb{R}^3} \frac{\eta(x) |\varphi_n(x)|^2}{2 + \frac{\kappa}{|x|}} dx + \|V \mathbb{1}_{r \leq |x| \leq 2R}\|_{L^\infty} \|\varphi_n \mathbb{1}_{r \leq |x| \leq 2R}\|_{L^2}^2.$$

Using the strong local convergence of  $\varphi_n$ , we finally deduce that  $\lim_{n \rightarrow \infty} \|\varphi_n^2\| = \lim_{n \rightarrow \infty} \|\varphi_n^0\| = 0$  which is a contradiction with  $\|\varphi_n\| = 1$ .  $\square$

The conclusion follows from Theorem 3.2 (ii). This ends the proof of Theorem 3.4.  $\square$

### 3.2.3. Dual Kinetic Balance

Let us now study the method which was introduced in [32], based this time on the splitting of the Hilbert space induced by the projector  $\mathcal{P}_\epsilon$  defined in (2.29). We have seen in Theorem 2.5 that pollution might occur when  $\epsilon$  is not small enough. We prove below that introducing a balance as proposed in [32] does not in general decrease the polluted spectrum.

Let us introduce the following operator

$$J \begin{pmatrix} \varphi \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ \varphi \end{pmatrix}$$

defined on  $\mathcal{P}L^2(\mathbb{R}^3, \mathbb{C}^4)$  with values in  $(1 - \mathcal{P})L^2(\mathbb{R}^3, \mathbb{C}^4)$ . Next we introduce the following balance operator [32]

$$\boxed{L_{DKB} = U_\epsilon J U_\epsilon} \quad (3.36)$$

which is an isometry defined on  $\mathcal{P}_\epsilon L^2(\mathbb{R}^3, \mathbb{C}^4)$  with values in  $(1 - \mathcal{P}_\epsilon)L^2(\mathbb{R}^3, \mathbb{C}^4)$ . A calculation shows that, like in [32], formulas (24) and (25),

$$L_{DKB} \begin{pmatrix} \varphi \\ \epsilon \sigma(-i\nabla)\varphi \end{pmatrix} = \begin{pmatrix} \epsilon \sigma(-i\nabla)\varphi \\ -\varphi \end{pmatrix}.$$

As before we may define  $L_{DKB}$  on  $\mathcal{C} = U_\epsilon C_0^\infty(\mathbb{R}^3, \mathbb{C}^4)$ .

**Theorem 3.5 (Dual Kinetic Balance).** *Assume that the real function  $V$  satisfies the same assumptions as in Theorem 2.4. Assume also that  $\mathcal{P}_\epsilon$  and  $L_{DKB}$  are defined as in (2.29) and (3.36) for some  $0 < \epsilon \leq 1$ . Then one has*

$$\begin{aligned} \overline{\text{Spu}(D^0 + V, \mathcal{P}_\epsilon, L_{DKB})} &= \overline{\text{Spu}(D^0 + V, \mathcal{P}_\epsilon)} \\ &= \left[ -1, \min \left\{ -\frac{2}{\epsilon} + 1 + \sup V, 1 \right\} \right] \cup \left[ \max \left\{ -1, \frac{2}{\epsilon} - 1 + \inf V \right\}, 1 \right]. \end{aligned}$$

**Proof.** We will use Theorem 3.1. Consider a radial function  $\zeta \in C_0^\infty(\mathbb{R}^3, \mathbb{R})$  and introduce the following functions:  $\varphi_1 := (\zeta, 0)$  and  $\varphi'_1 := (\sigma \cdot p)/|p|\varphi_1 \in \cap_{s>0} H^s(\mathbb{R}^3, \mathbb{C}^2)$ . We define similarly as in the proof of Theorem 2.5,  $\varphi_n(x) = n^{3/2}\varphi_1(n(x - x_0))$  and  $\varphi'_n(x) = n^{3/2}\varphi'_1(n(x - x_0))$ , where  $x_0$  is a fixed Lebesgue point of  $V$ . We note that  $\varphi'_n := (\sigma \cdot p)/|p|\varphi_n$ . Also, using that  $\widehat{\zeta}$  is radial, we get for any real function  $f$ :

$$\langle f(|p|)\varphi_n, \varphi'_n \rangle = \langle f(n|p|)\varphi_1, \varphi'_1 \rangle = \int_{S^2} \omega_1 d\omega \int_0^\infty |\widehat{\zeta}(|p|)|^2 f(n|p|)|p|^2 d|p| = 0. \quad (3.37)$$

A simple calculation shows that the  $2 \times 2$  matrix of  $D^0 + V$  in the basis  $(U_\epsilon(\varphi_n, 0), L_{DKB}U_\epsilon(\varphi_n, 0))$  reads

$$M_n = \begin{pmatrix} \langle A_{11}\varphi_n, \varphi_n \rangle & \langle A_{12}\varphi_n, \varphi_n \rangle \\ \langle A_{21}\varphi_n, \varphi_n \rangle & \langle A_{22}\varphi_n, \varphi_n \rangle \end{pmatrix}$$

where

$$\begin{aligned} A_{11} &= 1 + \frac{1}{\sqrt{1 + \epsilon^2|p|^2}} V \frac{1}{\sqrt{1 + \epsilon^2|p|^2}} + \frac{\epsilon \sigma \cdot p}{\sqrt{1 + \epsilon^2|p|^2}} \left( \frac{2}{\epsilon} - 2 + V \right) \frac{\epsilon \sigma \cdot p}{\sqrt{1 + \epsilon^2|p|^2}}, \\ A_{22} &= -1 + \frac{1}{\sqrt{1 + \epsilon^2|p|^2}} V \frac{1}{\sqrt{1 + \epsilon^2|p|^2}} + \frac{\epsilon \sigma \cdot p}{\sqrt{1 + \epsilon^2|p|^2}} \left( -\frac{2}{\epsilon} + 2 + V \right) \frac{\epsilon \sigma \cdot p}{\sqrt{1 + \epsilon^2|p|^2}}, \\ A_{12} &= (A_{21})^* = \frac{2\epsilon - 1 + \epsilon^2|p|^2}{1 + \epsilon^2|p|^2} (\sigma \cdot p) + \epsilon \frac{1}{\sqrt{1 + \epsilon^2|p|^2}} [V, \sigma \cdot p] \frac{1}{\sqrt{1 + \epsilon^2|p|^2}}. \end{aligned}$$

We infer from (3.37) that

$$\left\langle \frac{2\epsilon - 1 + \epsilon^2|p|^2}{1 + \epsilon^2|p|^2} (\sigma \cdot p) \varphi_n, \varphi_n \right\rangle = 0$$

for every  $n$ . Also we have

$$\lim_{n \rightarrow \infty} \left\| \frac{\epsilon \sigma \cdot p}{\sqrt{1 + \epsilon^2|p|^2}} \varphi_n - \varphi'_n \right\|_{H^1} = 0.$$

It is then easy to see that

$$\lim_{n \rightarrow \infty} M_n = \begin{pmatrix} \frac{2}{\epsilon} - 1 + V(x_0) & 0 \\ 0 & -\frac{2}{\epsilon} + 1 + V(x_0) \end{pmatrix}.$$

Note that  $L_{DKB}(\varphi_n, 0) \rightarrow 0$  since  $L_{DKB}$  is an isometry. The result follows from Theorem 3.1, by varying  $x_0$ .  $\square$

## References

1. L. ACETO, P. GHELARDONI, AND M. MARLETTA, *Numerical computation of eigenvalues in spectral gaps of Sturm-Liouville operators*, J. Comput. Appl. Math., 189 (2006), pp. 453–470.
2. L. BOULTON, *Non-variational approximation of discrete eigenvalues of self-adjoint operators*, IMA J. Numer. Anal., 27 (2007), pp. 102–121.
3. L. BOULTON AND N. BOUSSAID, *Non-variational computation of the eigenstates of Dirac operators with radially symmetric potentials*. arXiv:0808.0228, 2008.
4. L. BOULTON AND M. LEVITIN, *On approximation of the eigenvalues of perturbed periodic Schrödinger operators*, J. Phys. A, 40 (2007), pp. 9319–9329.
5. C. BROUDER, G. PANATI, M. CALANDRA, C. MOURougane, AND N. MARZARI, *Exponential localization of wannier functions in insulators*, Physical Review Letters, 98 (2007), p. 046402.

6. É. CANCÈS, A. DELEURENCE, AND M. LEWIN, *Non-perturbative embedding of local defects in crystalline materials*, J. Phys.: Condens. Matter, 20 (2008), p. 294213.
7. I. CATTO, C. LE BRIS, AND P.-L. LIONS, *On the thermodynamic limit for Hartree-Fock type models*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 18 (2001), pp. 687–760.
8. E. DAVIES, *Spectral theory and differential operators*, vol. 42 of Cambridge Studies in Advanced Mathematics, Cambridge University Press, Cambridge, 1995.
9. E. DAVIES AND M. PLUM, *Spectral pollution*, IMA J. Numer. Anal., 24 (2004), pp. 417–438.
10. J. DESCLOUX, *Essential numerical range of an operator with respect to a coercive form and the approximation of its spectrum by the galerkin method*, SIAM J. Numer. Anal., 18 (1981), pp. 1128–1133.
11. J. DOLBEAULT, M. J. ESTEBAN, M. LOSS, AND L. VEGA, *An analytical proof of Hardy-like inequalities related to the Dirac operator*, J. Funct. Anal., 216 (2004), pp. 1–21.
12. J. DOLBEAULT, M. J. ESTEBAN, AND É. SÉRÉ, *On the eigenvalues of operators with gaps. Application to Dirac operators*, J. Funct. Anal., 174 (2000), pp. 208–226.
13. G. W. F. DRAKE AND S. P. GOLDMAN, *Application of discrete-basis-set methods to the dirac equation*, Phys. Rev. A, 23 (1981), pp. 2093–2098.
14. K. G. DYALL AND K. FÆGRI JR, *Kinetic balance and variational bounds failure in the solution of the dirac equation in a finite gaussian basis set*, Chem. Phys. Letters, 174 (1990), pp. 25–32.
15. M. J. ESTEBAN, M. LEWIN, AND É. SÉRÉ, *Variational methods in relativistic quantum mechanics*, Bull. Amer. Math. Soc. (N.S.), 45 (2008), pp. 535–593.
16. M. J. ESTEBAN AND M. LOSS, *Self-adjointness for Dirac operators via Hardy-Dirac inequalities*, J. Math. Phys., 48 (2007), pp. 112107, 8.
17. W. D. EVANS, P. PERRY, AND H. SIEDENTOP, *The spectrum of relativistic one-electron atoms according to Bethe and Salpeter*, Commun. Math. Phys., 178 (1996), pp. 733–746.
18. I. P. GRANT, *Conditions for convergence of variational solutions of Dirac’s equation in a finite basis*, Phys. Rev. A, 25 (1982), pp. 1230–1232.
19. C. HAINZL, M. LEWIN, AND É. SÉRÉ, *Existence of a stable polarized vacuum in the Bogoliubov-Dirac-Fock approximation*, Commun. Math. Phys., 257 (2005), pp. 515–562.
20. ———, *Existence of atoms and molecules in the mean-field approximation of no-photon quantum electrodynamics*, Arch. Rational Mech. Anal., (in press).
21. A. C. HANSEN, *On the approximation of spectra of linear operators on Hilbert spaces*, J. Funct. Anal., 254 (2008), pp. 2092–2126.
22. W. KUTZELNIGG, *Basis set expansion of the dirac operator without variational collapse*, Int. J. Quantum Chemistry, 25 (1984), pp. 107–129.
23. M. LEVITIN AND E. SHARGORODSKY, *Spectral pollution and second-order relative spectra for self-adjoint operators*, IMA J. Numer. Anal., 24 (2004), pp. 393–416.
24. N. MARZARI AND D. VANDERBILT, *Maximally localized generalized wannier functions for composite energy bands*, Phys. Rev. B, 56 (1997), pp. 12847–12865.
25. G. NENCIU, *Existence of the exponentially localised Wannier functions*, Commun. Math. Phys., 91 (1983), pp. 81–85.
26. G. PANATI, *Triviality of bloch and blochdirac bundles*, Ann. Henri Poincaré, 8 (2007), pp. 995–1011.
27. G. PESTKA, *Spurious roots in the algebraic dirac equation*, Phys. Scr., 68 (2003), pp. 254–258.
28. A. POKRZYWA, *Method of orthogonal projections and approximation of the spectrum of a bounded operator*, Studia Math., 65 (1979), pp. 21–29.
29. ———, *Method of orthogonal projections and approximation of the spectrum of a bounded operator. II*, Studia Math., 70 (1981), pp. 1–9.
30. M. REED AND B. SIMON, *Methods of modern mathematical physics. IV. Analysis of operators*, Academic Press, New York, 1978.
31. E. SEILER AND B. SIMON, *Bounds in the Yukawa2 quantum field theory: upper bound on the pressure, Hamiltonian bound and linear lower bound*, Commun. Math. Phys., 45 (1975), pp. 99–114.
32. V. SHABAEV, I. I. TUPITSYN, V. A. YEROKHIN, G. PLUNIEN, AND G. SOFF, *Dual kinetic balance approach to basis-set expansions for the Dirac equation*, Phys. Rev. Lett., 93 (2004),

- p. 130405.
33. B. SIMON, *Trace ideals and their applications*, vol. 35 of London Mathematical Society Lecture Note Series, Cambridge University Press, Cambridge, 1979.
  34. R. E. STANTON AND S. HAVRILIAK, *Kinetic balance: A partial solution to the problem of variational safety in Dirac calculations*, J. Chem. Phys., 81 (1984), pp. 1910–1918.
  35. G. STOLZ AND J. WEIDMANN, *Approximation of isolated eigenvalues of ordinary differential operators*, J. Reine Angew. Math., 445 (1993), pp. 31–44.
  36. ———, *Approximation of isolated eigenvalues of general singular ordinary differential operators*, Results Math., 28 (1995), pp. 345–358.
  37. J. D. TALMAN, *Minimax principle for the Dirac equation*, Phys. Rev. Lett., 57 (1986), pp. 1091–1094.
  38. B. THALLER, *The Dirac equation*, Texts and Monographs in Physics, Springer-Verlag, Berlin, 1992.
  39. L. E. THOMAS, *Time dependent approach to scattering from impurities in a crystal*, Commun. Math. Phys., 33 (1973), pp. 335–343.
  40. G. H. WANNIER, *The structure of electronic excitation levels in insulating crystals*, Phys. Rev., 52 (1937), pp. 191–197.