

Chapter 3

Theoretical Framework

We attempt to give the flavor of the rich phenomenology of inclusive $b \rightarrow u \ell \nu$ decays. The focus is on the few remaining hurdles on the theoretical side that continue to stall efforts to deliver a well-defined, rigorous, and experimentally viable recipe for a precision extraction of the CKM element $|V_{ub}|$. The story begins with a description of heavy quark effective theory and the analysis of inclusive decays, and then moves deeper into the territory of $b \rightarrow u \ell \nu$ phenomenology to review recent proposals for measuring $|V_{ub}|$, concluding with a survey of lingering difficulties.

3.1 Motivation

Current best-knowledge [2] for $|V_{ub}|$ from inclusive measurements is

$$|V_{ub}| = (4.63 \pm 0.72) \times 10^{-3}. \quad (3.1)$$

Despite almost two decades of experimental and theoretical work on inclusive $b \rightarrow u \ell \nu$, the total precision is still at the 18% level, and remains dominated by theory errors. This simple observation prompts several immediate questions:

- What's limiting the progress? Why can't we do better at $|V_{ub}|$?
- What's been happening for the past 15 years?
- How does the current study of weak annihilation fit into the larger picture of measuring $|V_{ub}|$ to higher accuracy?
- Will the situation ever improve?

To explore these issues and the story that threads them together, we need to first review some of the theoretical machinery commonly used in the study of $b \rightarrow u \ell \nu$ decays. The operator production expansion (OPE) organizes the separation of short- and long-distance contributions to weak amplitudes, and heavy quark effective theory (HQET) provides a rigorous formalism for handling the non-perturbative physics that continually threatens to engulf any calculation. When applied to the computation of inclusive quantities such as (differential) decay rates, these tools offer a systematic framework for calculation and evaluation of correction terms and theoretical errors, especially when working in restricted regions of the full $b \rightarrow u \ell \nu$ phase space.

Once this footing is secure, we will advance to some of the proposals for extracting $|V_{ub}|$ from experimental measurements of semileptonic B decay. The endpoint of the lepton energy spectrum seems simple and very enticing: Historically, it's where the first observation of charmless b decay was made, and in this region, $b \rightarrow u \ell \nu$ has a very simple experimental signature. But as we'll see, unusual efforts are required to make any theoretical sense of decays in this part of the spectrum, and ultimately other experimental measurements (such as the photon energy spectrum in $b \rightarrow s \gamma$) are needed to extract $|V_{ub}|$ with any hope of well-understood errors. This example will expose more fully the tension between theoretical and experimental efforts to pursue $|V_{ub}|$. Two other possible approaches to $|V_{ub}|$ will further illustrate the depth and complexity of this convoluted relationship: a cut on the mass of the hadronic system (X_u), or a cut on the mass of the dilepton ($\ell \nu$) pair. These examples will serve as a springboard for discussion of the remaining challenges that still make the path to a successful, unqualified measurement of $|V_{ub}|$ a rocky, uphill climb in wind-searing cold.

3.2 Infrastructure

Even in the weak decay $b \rightarrow uW$, the effects of the strong interaction cannot be ignored. Due to a property of QCD called confinement, the b quark is always embedded in a strongly-interacting environment of other quarks and gluons, distracting from the essential weak physics of the decay. The consequences for calculation are dire: no meaningful prediction about hadrons in the real world can be made without accounting for the confinement of the quarks within them, but the physics of this binding is fundamentally non-perturbative, arising from a theory that is strongly-coupled in this regime. Using the tools of the OPE, renormalization, and HQET, however, the problem can be broken into well-constrained, manageable pieces. The key insight is the identification of several separations of scale in the weak physics describing B meson decay [23–31, 33].

3.2.1 Operator Product Expansion

There is a natural separation in scales between the typical hadronic energy scale $\mathcal{O}(1 \text{ GeV})$ that characterizes the binding of quarks into hadrons and the scale $\mathcal{O}(M_{W,Z})$ of weak interactions. The operator product expansion (OPE) is a computational and conceptual tool for taking advantage of this separation in a rigorous way, particularly when calculating the amplitude for weak quark decay. Very generally, the OPE isolates the short-distance physics that can be readily calculated with the tools of perturbation theory and the renormalization group, and replaces the long-distance physics with a set of hadronic matrix elements that wrap up the non-perturbative subtleties.

As a rather trivial but instructive example of this technique, consider the four-quark interaction $c \rightarrow su\bar{d}$,¹ shown at tree-level in Fig 3.1. The amplitude A for the process, constructed according to the usual Feynman rules, is

$$A = -\frac{G_F}{\sqrt{2}} V_{cs}^* V_{ud} \frac{M_W^2}{k^2 - M_W^2} (\bar{s}\gamma^\mu P_L c) (\bar{u}\gamma_\mu P_L d), \quad (3.2)$$

where P_L is the projection operator for the left-handed component of weak isospin, and k is the momentum transfer through the W propagator. Noting that k^2 is small compared to M_W^2 , we expand the propagator and find

$$A = -\frac{G_F}{\sqrt{2}} V_{cs}^* V_{ud} (\bar{s}\gamma^\mu P_L c) (\bar{u}\gamma_\mu P_L d) + \mathcal{O}\left(\frac{k^2}{M_W^2}\right). \quad (3.3)$$

In the limit of small k^2 , we discard the second term and approximate the full amplitude A by the first, creating a (very) simple effective theory for this charged-current weak interaction. Pictured on the right of Fig 3.1 is the new vertex that results, describing an effective four-quark interaction with a new coupling. By including the next few higher-order terms from the expansion of the propagator, we can introduce additional new local interactions and thus expand the range of the effective theory.

The discarded terms in Eqn 3.3 can alternatively be considered as higher-dimension local operators inherited from a larger theory that have been lost (or have become insignificant) in the low-energy limit. They represent *corrections* to the low-energy effective four-quark interaction we are left with at scales $k \ll M_W$, which we view as a legitimate, comprehensive theory in its own right, in this regime.

The notion of an operator product expansion fully generalizes this kind of identification in a completely rigorous and systematic way.

Note the OPE series is equivalent to the original theory when considered to all orders. The truncation of the series yields a systematic approximation scheme for describing low-energy processes, and only neglects contributions suppressed by higher powers of $1/M_W$.

More generally, for any weak process, the OPE prescription tells us how to write an *effective* weak Hamiltonian as a collection of local operators with (running) coefficients:

$$\mathcal{H}_{\text{eff}} = \frac{G_F}{\sqrt{2}} \sum_i V_{\text{CKM}}^i C_i(\mu) Q_i \quad (3.4)$$

The operators Q_i are treated as new interaction terms, and the (Wilson) coefficients $C_i(\mu)$ are viewed as scale-dependent couplings resulting from the short-distance, calculable physics, evaluated with the tools of perturbation theory and

¹This example is borrowed from a nice presentation by Buras [15].

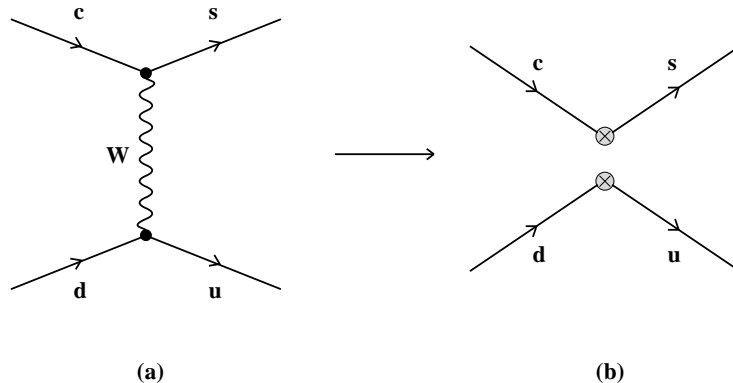


Figure 3.1: Tree-level diagram for the quark-level transition $c \rightarrow s u \bar{d}$. The diagram in the full theory is shown on the left, and after “integrating out” the W propagator, we are left with the effective interaction shown on the right.

renormalization group methods. Roughly speaking, the scale μ marks the boundary between the high-energy contributions captured in the C_i , and the low-energy contributions pieces left in the Q_i .

Clearly, in order to compute the amplitude for some weak process $i \rightarrow f$, the hadronic matrix elements $\langle f | Q_j | i \rangle$ must be evaluated. However, these quantities contain the non-perturbative physics of the strong interaction and a different approach is required. To meet these needs, tools such as lattice calculations, the $1/N$ expansion (where N is the number of colors), QCD sum rules, hadronic sum rules, and chiral perturbation theory have been developed. For the B , heavy quark effective theory (HQET) (next up; see Ch 3.2.2) and the heavy quark expansion (HQE) address these issues. Despite the myriad techniques available, however, none is without limitation: In the end, the dominant theoretical uncertainties on these calculations still come from the evaluation of the matrix elements $\langle Q_j \rangle$.²

It is unfortunate that in many cases, the hadronic matrix elements do not yield to our current theoretical tools and cannot be readily evaluated. In a few instances, it is possible to fall back on experimental measurements of several key matrix elements, or to employ suitable ratios and combinations of amplitudes in ways such that the unknown terms cancel. Finally, it is sometimes possible to derive constraints or other bounds on the matrix elements using flavor or other approximate symmetries, provided the symmetry-breaking effects can be reliably estimated.

We conclude with a short remark about renormalization schemes (RS) and scale dependence. It is a fundamental theorem of renormalization that physical observables such as cross-sections and decay rates do not depend on the choice of

²Here we include the additional uncertainty arising from the non-perturbative parameters introduced in the HQE.

RS, when computed to all orders in perturbation theory [3]. However, a truncated perturbation series *will* exhibit RS dependence, and quantities calculated to finite order will thus display some residual sensitivity to the choice of RS. At leading and next-to-leading order (NLO) in perturbation theory, the RS dependence is fully specified by the renormalization scale μ , but at next-to-next-to-leading order (NNLO),³ this is no longer sufficient and the specification of the scale μ must be augmented by the specification of the scheme as well. It is generally agreed that appropriate choices for μ can lead to reduced RS- and scale-dependence, but there are no clear criteria for determining the “best” choice, even within a given scheme.

3.2.2 Heavy Quark Effective Theory⁴

Since the development of the OPE and renormalization methods for evolving physics between scales, particle physicists have learned to view the world as a series of effective theories, to which perturbative corrections are made in a well-defined and consistent way. Indeed, the success of the Standard Model is attributed to the fact that, in some sense, it is the unique low-energy limit of whatever physics really prevails at high energies.⁵ The program set forth by Wilson—the operator product expansion just discussed—provides a rigorous framework for making calculations in a theory that seems initially plagued with infinities and unresolvable ambiguities. The trouble is that the coupling constant of QCD does not lend itself to a perturbative expansion at large distance scales: *confinement* leads to a coupling that increases in the low-energy limit, preventing bare quarks from existing individually and in isolation. Hence, even in the user-friendly Standard Model, there are calculations that are rendered impossible from first principles, since they involve fundamentally non-perturbative effects that cannot be handled with the usual assortment of theoretical tools.

A Physical Picture

Enter the heavy quark limit, an insight into QCD that arises from another separation of scales, this time springing from the (happen-chance) separation between the masses of the so-called heavy quarks (b , c) and the scale Λ_{QCD} where the strong

³Leading-order calculations typically do not offer enough precision to be useful for careful tests of the Standard Model, so the more tedious and difficult NLO and NNLO calculations are almost always required for useful theoretical prediction. Yup, life as a theorist is hard.

⁴Some of this material was prepared previously [33], but why skip the chance to have it preserved forever on microfilm by including it here?

⁵Such physics, of course, may not be described by something as straightforward as a quantum field theory. But it will give rise to a field theory in the limit of the low energies within the range of current experiment.

coupling α_s becomes of order unity.⁶ First articulated in the early 1990's, this idea essentially exploits the observation that in a hadron H_Q containing a heavy quark Q , the light degrees of freedom (ldf) are insensitive to the actual values of the mass m_Q and spin orientation s_Q of the heavy quark. A simple physical argument suffices to motivate this conclusion. Picture a heavy hadron as composed of a single heavy quark surrounded by a complicated and strongly-interacting cloud of light quarks, antiquarks, and gluons. The Compton wavelength of the heavy quark scales as the inverse of its mass, $\lambda_Q \sim 1/m_Q$. The ldf, in contrast, have momenta characterized by the scale Λ_{QCD} , corresponding to wavelengths of size $\lambda_{\text{ldf}} \sim 1/\Lambda_{\text{QCD}}$. Since $\Lambda_{\text{QCD}} \ll m_Q$, implying $\lambda_{\text{ldf}} \gg \lambda_Q$, the ldf cannot resolve any detailed features of the heavy quark beyond its conserved quantum numbers.

A similar conclusion follows from considering the situation in momentum space. Clearly, the structure of the hadron H_Q is determined by non-perturbative strong interactions. Recall that the asymptotic freedom of QCD implies that when quarks and gluons exchange momenta p much larger than the scale Λ_{QCD} , the process can be treated perturbatively in the strong coupling constant $\alpha_s(p)$. The typical momenta exchanged by the ldf with each other and with the heavy quark are of order Λ_{QCD} , for which a perturbative expansion is of no use. But in an exchange of momentum $p \ll m_Q$ between the ldf and the heavy quark, the heavy quark will essentially not recoil: it remains at rest in the frame of the hadron. In this limit, Q acts as a *static* source of electric and chromoelectric field. This latter field is what binds the constituents of H_Q together, but the field is actually independent of m_Q in this limit. Again, though the field is responsible for the clearly non-perturbative coupling that glues Q and the light degrees of freedom together, it doesn't depend in detail on the *mass* that produces it. The generalized result is that the properties of the light degrees of freedom depend only on the presence of the static gauge field, independent of the flavor and mass of the heavy quark carrying the gauge charge.

Further consideration along these lines illustrates that the light degrees of freedom are also blind to the spin orientation of the heavy quark. (Essentially, only the color electric field is relevant, since the color magnetic interaction vanishes in the $m_Q \rightarrow \infty$ limit.) These observations immediately lead to relations between the masses of various of the heavy mesons B , D , B^* , and D^* . These relations can be systematically improved by including corrections to the heavy-quark limit, as we discuss below. Excellent agreement with the data results, suggesting that this “heavy quark symmetry” is, in fact, a reasonably “good” symmetry, broken only to a modest extent.

This separation in scales is of incredible utility when making theoretical calculations. And, of course, it is satisfying to be able to make (some) predictions based

⁶For completeness, we note that current estimates place $m_b \sim 4.7$ GeV and $\Lambda_{\text{QCD}} \sim 200$ -500 MeV, although both of these parameters are subject to some ambiguity determined by the scheme within which they are applied.

on properties of QCD itself (in a certain limit). Much of the relevant physics—the short-distance effects—can be computed perturbatively using renormalization group methods; the remaining long-distance effects may simplify due to the realization of approximate symmetries which connect a large number of long-range properties to only a small number of hadronic matrix elements. In this manner, a calculation can actually make statements beyond the range of applicability of ordinary perturbation theory.

Heavy Quark Expansion

The heavy quark expansion provides a systematic treatment of this limit, organizing it into an expansion in powers of Λ/m_Q ,⁷ the ratio of characteristic scales across which the physics of the problem separates. The intuitive appeal of the picture painted above is translated into a powerful formalism for handling non-perturbative QCD in an organized fashion. In particular, this allows symmetry-breaking corrections to be computed in a controlled and model-independent fashion, and provides for the estimation of theoretical uncertainties. This latter feature of HQET represents a substantial improvement over the previous generation of quark and potential models that have been used to calculate the effects of the strong interaction. Despite the predictive power of such models, they are crippled by an inability to assess the validity of the assumptions used in the model-building phase, and it is typically not even possible to confidently assign errors within the context of a particular model. At the sacrifice of some predictive power, heavy quark symmetry grounds a computation firmly in QCD and yields useful phenomenological results.

Our presentation here is essentially a reproduction of treatments given by Neubert ([23, 26, 28]) and Falk [25]. The idea is that by explicitly separating the scales in the heavy quark field, one can derive the Lagrangian for HQET directly from QCD in a model-insensitive way. Further, one can include correction terms to the $m_Q \rightarrow \infty$ limit simply by keeping terms of order Λ/m_Q and higher during the initial calculation.

The typical starting point for this argument is to note that the heavy quark essentially carries all of the momentum of the hadron, and so it is useful to decompose the heavy quark momentum as follows

$$p_Q^\mu = m_Q v^\mu + k^\mu, \quad (3.5)$$

where v^μ is the four-velocity of the hadron, and $k^\mu \sim \Lambda_{\text{QCD}}$ is the small, fluctuating “residual momentum” contributed by the light degrees of freedom. The

⁷Here and elsewhere, it is to be assumed that, if otherwise unadorned or unlabeled, the symbol Λ refers to the scale Λ_{QCD} .

heavy quark limit suggests Q is never far from its mass shell, $p_Q^2 = M_Q^2$; explicitly requiring this leads to the constraint

$$m_Q^2 = (m_q v_\mu + k_\mu)^2 = m_Q^2 + 2m_q v \cdot k + k^2, \quad (3.6)$$

which forces us to conclude that

$$v \cdot k = 0, \quad (3.7)$$

since the term k^2 can be neglected in the heavy quark limit. Eqn 3.7 is sufficient to identify the operators

$$P_\pm = \frac{1 \pm \not{v}}{2} \quad (3.8)$$

as projection operators: They project onto the positive (P_+) and negative (P_-) frequency parts of the Dirac field of the heavy quark, $Q(x)$. From the Dirac representation in the rest frame of the heavy quark, it can be seen that P_+ and P_- project, respectively, onto the upper two and lower two components of the heavy quark spinor. In the limit $m_Q \rightarrow \infty$, in which Q is essentially on-shell, only the “large” upper components of the field Q propagate, and the “small” lower components become negligible; mixing of the upper components with the lower is suppressed by a factor of $1/2m_Q$. We can see this explicitly in the action of the projection operators on the field:

$$P_+ Q(x) = Q(x) + \mathcal{O}(1/m_Q), \quad P_- Q(x) = 0 + \mathcal{O}(1/m_Q). \quad (3.9)$$

Thus one is led naturally to introduce “large” and “small” component fields h_v and H_v given by

$$h_v(x) = e^{im_Q v \cdot x} P_+ Q(x), \quad H_v(x) = e^{im_Q v \cdot x} P_- Q(x) \quad (3.10)$$

so that

$$Q(x) = e^{-im_Q v \cdot x} [h_v(x) + H_v(x)]. \quad (3.11)$$

Note that the effective field $h_v(x)$ is independent of the heavy quark mass m_Q ; it carries only a velocity label v and is now a two-component object. The modifications to the full field $Q(x)$ basically project out the positive frequency part and ensure that the states annihilated by $h_v(x)$ have no dependence on m_Q . Similarly, the field $H_v(x)$ carries the effects of order $1/m_Q$ and vanishes in the heavy quark limit. In this same limit, we can write

$$Q(x) \sim e^{-im_Q v \cdot x} h_v(x) + \dots \quad (3.12)$$

where we mean that terms of $\mathcal{O}(1/m_Q)$ are to be dropped.

Recall now that the QCD Lagrangian for a heavy quark of mass m_Q is

$$\mathcal{L}_{QCD} = \bar{Q}(x)(i \not{D} - m_Q)Q(x), \quad (3.13)$$

where $D_\mu = \partial_\mu - igA_\mu^a T^a$ is the usual gauge-covariant derivative for the strong interaction. To explore the form of QCD in the heavy quark limit, we substitute the expression (3.11) for $Q(x)$ into the Lagrangian (3.13):

$$\mathcal{L} = \bar{h}_v iv \cdot D h_v - \bar{H}_v (iv \cdot D + 2m_Q) H_v + \bar{h}_v i \not{D}_\perp H_v + \bar{H}_v i \not{D}_\perp h_v. \quad (3.14)$$

Here we've used the popular notation $D_\perp^\mu = D^\mu - v^\mu v \cdot D$, clearly orthogonal to the heavy quark velocity. It is clear now that h_v describes massless degrees of freedom, while H_v corresponds to fluctuations with twice the heavy quark mass m_Q . By expanding according to the leading-order prescription for $Q(x)$ in Eqn 3.12, we finally obtain the Lagrangian for HQET:

$$\mathcal{L}_{HQET}^0 = \bar{h}_v iv \cdot D h \quad (3.15)$$

$$= \bar{h}_v (iv^\mu \partial_\mu + gT^a v^\mu A_\mu^a) h_v. \quad (3.16)$$

The Feynman rules for this theory are simple. The propagator for the heavy quark is simply

$$\frac{i}{v \cdot k + i\epsilon}, \quad (3.17)$$

and there is a simple quark-gluon vertex with factor

$$igT^a v^\mu A_\mu^a. \quad (3.18)$$

Of course, both propagator and vertex factor are independent of the value of the heavy quark mass, reflecting the original symmetry that we sought to expose in this limit. They also have no Dirac structure, reflecting the additional spin symmetry that arises in this limit.

It is straightforward to include power corrections to the Lagrangian \mathcal{L}_{HQET} . Starting again from the exact QCD Lagrangian in Eqn 3.14, we apply the classical equation of motion

$$(iv \cdot D + 2m_Q) H_v = i \not{D}_\perp h_v \quad (3.19)$$

(which can be obtained by taking the variation of the Lagrangian with respect to the field \bar{H}_v), and find

$$\mathcal{L} = \bar{h}_v iv \cdot D h_v + \bar{h}_v i \not{D}_\perp \frac{1}{iv \cdot D + 2m_Q} i \not{D}_\perp h_v. \quad (3.20)$$

Finally, expanding the non-local operator on the right in powers of $1/m_Q$, we arrive at the final form

$$\mathcal{L}_{HQET} = \bar{h}_v iv \cdot D h_v + \frac{1}{2m_Q} \left[\bar{h}_v (iD_\perp)^2 h_v + \frac{g}{2} \bar{h}_v \sigma^{\alpha\beta} G_{\alpha\beta} h_v \right] + \dots \quad (3.21)$$

where the first-order $1/m_Q$ corrections to the heavy quark limit are now apparent. The infinite series of local operators that has been neglected is formally $\mathcal{O}(1/m_Q^2)$; the combination as a whole forms the basis for the OPE of HQET.

These leading corrections have a simple interpretation, which becomes clearer in the rest frame where $v^\mu = (1, 0, 0, 0)$. The first term, spin-independent, becomes

$$\frac{1}{2m_Q} \mathcal{O}_{kin} \equiv \frac{1}{2m_Q} \bar{h}_v (iD_\perp)^2 h_v \rightarrow -\frac{1}{2m_Q} \bar{h}_v (i\vec{D})^2 h_v \quad (3.22)$$

which is readily identified as the gauge-covariant extension of the kinetic energy arising from the off-shell residual motion of the heavy quark. (In the rest frame, $(iD_\perp)^2$ is the operator for $-\vec{k}^2$.) The second, spin-dependent part becomes

$$\frac{1}{2m_Q} \mathcal{O}_{mag} \equiv \frac{1}{2m_Q} \frac{g}{2} \bar{h}_v \sigma^{\alpha\beta} G_{\alpha\beta} h_v \rightarrow \frac{1}{4m_Q} \bar{h}_v \sigma^{ij} T^a h_v \times g G_{ij}^a \equiv g \vec{\mu}_Q^a \cdot \vec{B}^a \quad (3.23)$$

Here we've explicitly identified the chromomagnetic moment of the heavy quark, and $B^i = -\frac{1}{2} \epsilon^{ijk} G^{jk}$ are the components of the chromomagnetic gluon field. This term represents the coupling of the spin of the heavy quark to the chromomagnetic field, akin to the familiar hyperfine interaction in atomic physics. The coupling breaks both the heavy quark flavor (mass) and spin symmetry, and is responsible, for instance, for the mass-splittings between the D - D^* and B - B^* states. In heavy quark effective theory, both of these correction terms are treated as part of an *interaction* Lagrangian, even though \mathcal{O}_{kin} has a piece which is a pure bilinear in the quark field.

Our discussion so far has dealt with the construction of the HQET Lagrangian, but it is not yet complete. In order to use the theory, short-distance physics needs also to be properly included. The typical program for doing so is to match the full Standard Model onto the four-fermion theory at the scale M_W , and then run it down to the scale relevant for the physics of interest, *e.g.* m_b . It is then that rates, differential distributions, etc. can be calculated in HQET in an expansion in powers of $1/m_Q$. The result of this “matching” is a renormalization of the coefficients of the operators in the heavy quark expansion. We jump directly to the end result: In the context of the operator product expansion, the Lagrangian for HQET is often recast as [28]

$$\mathcal{L}_{HQET} = \bar{h}_v i v \cdot D h_v + \frac{C_{kin}}{2m_Q} \bar{h}_v (iD_\perp)^2 h_v + \frac{g C_{mag}}{4m_Q} \bar{h}_v \sigma^{\alpha\beta} G_{\alpha\beta} h_v + \mathcal{O}(1/m_Q^2), \quad (3.24)$$

where the Wilson coefficients C_{kin} and C_{mag} result from short-distance effects and generally depend on the scale at which the operators are renormalized. The remaining HQET operators capture the non-perturbative physics, and present the chief computational challenge, if one can guarantee that the perturbative effects are under control.

The matrix elements of the kinetic energy and hyperfine interaction terms appear in many applications of HQET, and it is conventional to define for the B system:

$$-\lambda_1 = \frac{1}{2M_B} \langle B | \mathcal{O}_{kin} | B \rangle \quad (3.25)$$

$$3\lambda_2 = \frac{1}{2M_B} \langle B | \mathcal{O}_{mag} | B \rangle \quad (3.26)$$

These parameters characterize the order $1/m_b^2$ corrections in HQET; at $\mathcal{O}(1/m_b^3)$ seven additional operators arise in the OPE, but they are not as well constrained theoretically. Some workers in the field use the parameters μ_π^2 and μ_G^2 , which are related to the parameters above in a simple way, but receive short-distance corrections in a different fashion [31]. (Note that for the case where the meson H_Q is actually a vector particle, the definition for λ_1 is unchanged, but that for λ_2 acquires a factor of -3 on the RHS [39].)

As mentioned above, the parameter λ_2 is related to the mass-splitting between vector and pseudoscalar meson states through

$$M_{B^*}^2 - M_B^2 = 4\lambda_2 + \mathcal{O}(1/m_b) \quad (3.27)$$

which, by using the experimentally determined values of the meson masses, evaluates numerically to

$$\lambda_2 = 0.12 \text{ GeV}^2. \quad (3.28)$$

On the other hand, the parameter λ_1 cannot be determined from hadron spectroscopy, although it has been shown to be related to the difference of the pole masses for charm and bottom quarks through [28]

$$m_b - m_c = (\overline{M}_B - \overline{M}_D) + \lambda_1 \left(\frac{1}{2\overline{M}_B} - \frac{1}{2\overline{M}_D} \right) + \dots \quad (3.29)$$

Here $\overline{M}_B = \frac{1}{4}(M_B + 3M_{B^*})$ (similarly for \overline{M}_D) denotes the spin-averaged meson mass, defined so that there is no contribution from the chromomagnetic interaction. Theoretical estimates [26] for λ_1 are in the range $-(0.3 \pm 0.2) \text{ GeV}^2$. The current opinion is that this parameter is not physically meaningful and suffers from an intrinsic ambiguity of order Λ_{QCD}^2 . For further discussion of the status of these parameters, see Ref [28] and the references therein.

A third HQET parameter arises in the analysis of the heavy quark systems at this order,

$$\overline{\Lambda} = \lim_{m_b \rightarrow \infty} (M_B - m_b), \quad (3.30)$$

which measures the “mass” carried by the light degrees of freedom. It is connected to a residual mass term that can appear in the HQET Lagrangian; this presence leads to an ambiguity of order Λ_{QCD} in the heavy quark mass. With a particular choice of the expansion parameter m_Q , this residual mass term can be made to vanish, and then m_Q coincides with the pole mass used in perturbation theory. Other choices for the expansion parameter will of course adjust the value of $\overline{\Lambda}$, but this arbitrariness disappears in the calculation of actual physical quantities. This parameter is often used to re-express predictions in terms of meson masses through relations of the form [24]

$$M_Q = m_Q + \overline{\Lambda} - \frac{\lambda_1 + 3\lambda_2}{2m_Q} + \dots \quad (3.31)$$

This introduction to the formalism of heavy-quark symmetry is readily extended to make model-independent calculations on many phenomenological fronts. Correctly handling the perturbative effects and understanding the dependence on the HQET parameters are generally difficult issues, but the essential non-perturbative physics has been isolated and, to some extent, rendered tractable. We will not go into the details of these issues here; for more information, see the references mentioned above.

3.2.3 Theory of Inclusive Decays

The calculation of inclusive quantities such as (differential) decay rates is one area that has benefited from the equipment of HQET. Recall that in these calculations, we sum over all possible hadronic final states allowed by the kinematics or some set of global quantum numbers. Although the key theme is again the separation of long- and short-distance physics, there is a new ingredient: the notion of so-called “quark-hadron” duality which also relies on the heavy quark limit. From the theoretical point of view, such decays have two advantages: First, bound-state effects related to the initial state—such as the “Fermi motion” of the heavy quark within the confines of the hadron—are accounted for in a systematic way with the heavy quark expansion. Secondly, the fact that the final state is a sum over many hadronic channels eliminates bound-state effects in the final state hadrons. Essentially, the final state quarks are guaranteed with unit probability to hadronize, and so the particular details of how this comes about are irrelevant. That this process does not “interfere” with the heavy quark decay can be seen by noting that the timescale Δt for the b quark decay is of size $1/m_b$, while the timescale associated with hadronization is again set by QCD: $\Delta t \sim 1/\Lambda_{\text{QCD}}$, *i.e.* the hadronization occurs at a much later time than the initial b decay, so the physics decouples. It is typical of the success of HQET that—at leading order—its predictions match the more naïve parton-model calculations carried out in the early days, and further indicative of its power that it also includes a prescription for computing sub-leading corrections to these rates—which subsequently compare well with experiment.

It is instructive to examine the outline for an inclusive calculation to see just how the machinery of HQET is applied, and to gain some understanding of how the assumption of duality arises as a necessary part of the computation. To keep the analysis general, we consider the rate $\Gamma(B \rightarrow X)$, where X represents the set of final states consistent with whatever inclusive quantities have been specified. The discussion here once again closely follows the one presented by Falk [25] and Neubert [28]. The inclusive decay is a sum over all possible final states, which is really a sum over exclusive modes, followed by an integral over the phase space for each mode. Hence we begin by writing

$$\Gamma(B \rightarrow X) = \sum_X \int d[\text{P.S.}] |\langle X | \mathcal{O}_{bX} | B \rangle|^2 \quad (3.32)$$

where the operator \mathcal{O}_{bX} is a convenient shorthand for whatever effective operators are relevant for the process under consideration. An analog of the optical theorem for QCD allows us to express the transition rate instead as the imaginary part of a forward scattering amplitude:

$$\Gamma = -2 \operatorname{Im} i \int dx e^{-ik \cdot x} \langle B | T \{ \mathcal{O}_{bX}^\dagger(x), \mathcal{O}_{bX}(0) \} | B \rangle \equiv 2 \operatorname{Im} \mathbf{T} \quad (3.33)$$

The next step is to expand the time-ordered product $T \{ \mathcal{O}_{bX}^\dagger(x), \mathcal{O}_{bX}(0) \}$ in an operator product expansion, in which the transition operator \mathbf{T} is represented as a series of local operators containing the heavy-quark fields. The applicability of this expansion, and its computation in perturbation theory, rests on the heavy quark limit $m_b \gg \Lambda_{\text{QCD}}$.

For concreteness, we now restrict the discussion to the case where the final state X is semileptonic. In this case, the transition operator \mathcal{O}_{bX} factors into separate leptonic and hadronic currents. The core of the calculation is then the evaluation of a hadronic tensor of the form

$$T^{\mu\nu} = -i \int dx e^{-iq \cdot x} \langle B | T \{ J_{bX}^{\mu\dagger}(x), J_{bX}^\nu(0) \} | B \rangle \quad (3.34)$$

where we've introduced the total momentum q^μ transferred to the lepton system. It is standard to express this general tensor as a sum of five individual terms, each with definite Lorentz structure. When considering semileptonic B decay to massless u -quarks, only three of these are non-vanishing. However, such details take us beyond the scope of our current discussion.

A standard property of quantum field theory is that the imaginary part of a hadronic matrix element such as $T^{\mu\nu}$ is non-vanishing whenever there is a real intermediate state, that is, when the intermediate particles are all on mass-shell. Careful consideration of the possible avenues for the production of such physical intermediate states leads to the identification of poles and cuts in the complex $v \cdot q$ plane which neighbor the usual path of integration. Much can be said about the analyticity of the transition matrix, but the summary is that the location of the singularities along the real axis depends intimately on the details of QCD at long distances, since these physical states represent non-trivial bound states.⁸ Hence the evaluation of the integral depends in an essential way on physics that is incalculable. The perturbative treatment we have been pursuing suddenly seems to have broken down.

The solution is to deform the contour of integration away from the difficulties on the real axis, into complex $v \cdot q$ space. As Falk [25] argues, the momentum scale in the problem is set by m_b , so that the new contour is now predominantly a distance $m_b \gg \Lambda_{\text{QCD}}$ away from the difficult resonances, and a perturbative treatment

⁸Remember, perturbative QCD doesn't even predict these!

is again applicable, since on the new contour, $|q^2|$ is once again large. And in the small region where the contour is still close to the cut, we know the allowed phase space vanishes, so the calculation is once again on familiar perturbative ground. This claim, that the average value of a hadronic quantity can be calculated perturbatively when at each point the quantity depends on the details of non-perturbative physics, is the notion of (global) *parton-hadron* duality. This duality ranks as more than just a convenient theoretical assumption, since it is known to hold in the $m_b \rightarrow \infty$ limit, but it does not have the status of a standard approximation because there is no framework for calculating the leading corrections to this limit. For instance, it is known to hold only qualitatively [29] in $b \rightarrow c\bar{c}s$ processes. In any case, its use is key to the completion of the inclusive calculation. We will have more to say on this topic later (see Sec 3.4.1).

What remains is the construction of an operator product expansion for the transition operator to handle the short-distance (and therefore perturbative) physics properly. One can list the operators that can appear in this expansion quite generally. The lowest dimension operator that can contribute is $\bar{b}b$, with dimension $d = 3$. Likewise, there is only one possible gauge-invariant dimension 4 operator, $\bar{b}i\not{D}b$, but this one can be replaced by $m_b\bar{b}b$ with the application of the equations of motion. The first operator that is different from the simple $\bar{b}b$ has dimension 5 and includes the gluon field; it is the familiar $\bar{b}g\sigma_{\mu\nu}G^{\mu\nu}b$. (From dimensional considerations, one can see that the matrix elements of a dimension d operator are suppressed by inverse powers of the heavy quark mass.) Once these operators have been derived, the tools of HQET are used to replace the operators b and \bar{b} with the HQET fields h_v and \bar{h}_v according to the usual prescription. The Wilson coefficients associated with each operator are then computed to some order in perturbation theory. The result is that any inclusive decay rate can be written in the form [28]:

$$\Gamma(B \rightarrow X_f) = \frac{G_f^2 m_b^5}{192\pi^3} \left\{ c_3^f \left(1 + \frac{\lambda_1 + 3\lambda_2}{2m_b^2} \right) + c_5^f \frac{\lambda_2}{m_b^2} + \mathcal{O}(1/m_b^3) \right\}. \quad (3.35)$$

The prefactor is the usual result of the loop and phase space integrations, and the c_n^f are coefficient functions depending on the quantum numbers of the final state X_f ; naturally, they also contain the relevant CKM elements for the process. The leading contribution is just the result obtained from a naïve parton-level calculation. This is an instance of the KLN theorem [29] which asserts that the cancelation of the infrared singularities present in the exclusive rates gives a total inclusive width that is insensitive to the details of hadronization. It is worth noting that the kinetic energy contribution $\lambda_1/2m_b^2$ is nothing but the field-theory generalization of the expected Lorentz factor $(1 - \vec{v}_b^2)^{1/2} \simeq 1 - \vec{k}^2/2m_b^2$; this is the usual time-dilation factor that increases the lifetime $\tau = 1/\Gamma$ of a moving particle. The absence of first-order corrections of order $1/m_b$ in the above expression is also noteworthy, and is a natural consequence of the equations of motion in HQET. In this case, the terms disappear due to the choice of the residual mass term mentioned earlier [37]. In conclusion, we emphasize that the utility of HQET lies in the model-independent

prediction of quantitative *corrections* to these basic results.

We have only grazed the surface of the issues (both technical and physical) associated with the calculation of inclusive quantities. Observe that within the setting of HQET, inclusive rates depend only (up to order $1/m_b^2$) on the three HQET parameters λ_1 , λ_2 , and $\bar{\Lambda}$, and calculable coefficient functions. It should be clear that the procedure is well-established, grounded in a clear physical picture, and depends indispensably on the heavy quark limit of QCD.

3.3 Phenomenological Road to $|V_{ub}|$ ⁹

Our short review of some of the theoretical tools is now complete, and we turn to the advertised program: extraction of $|V_{ub}|$ from semileptonic B decays.

Falk [25] presents the results of applying this technology to the calculation of the total charmless semileptonic rate. Taking $m_u = 0$, he reports

$$\Gamma(B \rightarrow X_u \ell \nu) = \frac{G_F^2 |V_{ub}|^2}{192\pi^3} m_b^5 \left[1 + \left(\frac{25}{6} - \frac{2\pi^2}{2} \right) \frac{\alpha_s(m_b)}{\pi} \right. \quad (3.36) \\ \left. - (2.98\beta_0 + C_u) \left(\frac{\alpha_s(m_b)}{\pi} \right)^2 + \dots \right. \\ \left. + \frac{\lambda_1 - 9\lambda_2}{2m_b^2} + \dots \right].$$

The occurrence of the heavy quark mass m_b in the prefactor is troublesome since, as an unphysical parameter, it is subject to considerable theoretical ambiguity; in fact, this uncertainty dominates subsequent attempts to extract $|V_{ub}|$. The expression for the decay rate can be recast in several ways to reduce this sensitivity by exchanging some portion of it for additional input from experiment. For instance, one can substitute reference to the charm mass, employ an alternative kinematic mass for the b quark [14], or use a so-called ‘‘Upsilon expansion’’ to eliminate the quark mass in favor of the mass $m_{\Upsilon(1S)}$ of the $\Upsilon(1S)$ resonance [32].

The formula for the total decay rate can be inverted to provide for the extraction of $|V_{ub}|$ from a direct experimental measurement of the charmless semileptonic branching fraction. A recent assessment by the LEP VUB working group gave [12]:

$$|V_{ub}| = 0.00445 \left(\frac{\mathcal{B}(B \rightarrow X_u \ell \nu)}{0.002} \right)^{\frac{1}{2}} \left(\frac{1.55 \text{ps}}{\tau_B} \right)^{\frac{1}{2}} (1 \pm 0.020 \pm 0.052). \quad (3.37)$$

⁹We have been less than meticulous about notation and ν versus $\bar{\nu}$ conventions in preparing the following material. For a consistent and more accurate discussion of any of these topics, please consult any of the references cited at the start of this chapter.

The quoted errors arise from the OPE (first) and dependence on the mass m_b (second). This approach is theoretically quite straightforward, and results in a determination that has total theory error at about the 5% percent level. (Similar but not identical formulae are quoted in papers by Bigi [31], Ligeti [32], and Uraltsev [63].) As we shall see later, however, there are additional theoretical uncertainties not treated in this evaluation. More importantly, also overlooked is the significant impact that experimental practice can have on this rather naïve agenda.

Experimentally, we cannot neglect the need to separate the charmed semileptonic decays of the B from the charm-less ones. An assortment of kinematic variables have been considered to provide such discrimination, all of which rely on the basic fact that $m_u \ll m_c$. It is at kinematic endpoints that this difference can become significant and useful. The highest lepton energy E_ℓ , the smallest possible hadronic mass M_X , or the largest possible momentum transfer q^2 to the lepton system—for each of these, there is still some allowed phase space for $b \rightarrow u$ decays beyond the endpoint for charm. In short, $b \rightarrow u$ has more kinematic reach. The choice of discriminator and the location of the experimental cut together determine the fraction of $b \rightarrow u \ell \nu$ decays that are accessible with each method.

Working near the boundaries of phase space has its disadvantages, both experimental and theoretical. Briefly, the imposition of a cut changes the theoretical expansion parameters from α_s for perturbative QCD and Λ/m_b for the non-perturbative physics (*i.e.* HQET) to ones of the form $\alpha_s \log \rho$ and $\Lambda/(m_b \rho)$, where ρ is some measure of the area of phase space to which the analysis has been restricted. It may be the case that one or both of these new effective parameters will be of order unity, necessitating a reorganization of the entire calculation. Essentially, the heavy quark expansion must be generalized to account for the “Fermi motion” of the b quark within the B -meson; this is critical in the endpoint region because it is this residual momentum that accounts for the shift from the free-quark endpoint $m_b/2$ up to the physical endpoint $M_B/2$.

Due to the unfortunate reality of the ever-present $b \rightarrow c \ell \nu$ decays, theorists are forced to produce a calculation appropriate for the hard experimental cuts that must be made to isolate the signal events in a background-free region. Once again, various models have been employed to extrapolate from the rate measured in some kinematic window to the full charmless semileptonic rate, but these are hampered by questionable (or at least unverifiable) assumptions. In recent years, new techniques have been developed to deal with these problems; they rely heavily on the formalism of the heavy quark expansion but also owe their success to equally clever ideas for tackling both the experimental and theoretical analysis.

We will briefly review three such methods here, in rather broad strokes. The intent is not to capture the state-of-the-art, but to illustrate the common and disparate problems and triumphs that complicate the inclusive measurement of $|V_{ub}|$. For more detailed (and current and accurate) discussions of these and other techniques, see the references, *e.g.* Ref [2, 14, 72, 73]. In outline, the three methods

are:

- $E_\ell > (M_B^2 - M_D^2)/2M_B$
An analysis of the lepton energy spectrum in the endpoint region above the kinematic reach of $b \rightarrow c \ell \nu$, with non-perturbative effects described by a shape function that can be constrained from measurements of the photon spectrum in inclusive $b \rightarrow s \gamma$ decays [35].
- $M_X < M_D$
An analysis of the spectrum of the invariant mass M_X^2 of the recoiling hadronic system [44–46]. This offers the experimental advantage that charm background can be excluded while retaining a much larger fraction of $b \rightarrow u$ decays. On the other hand, the calculation is extremely sensitive to the precise value of the cut, and the shape function mentioned above is still relevant. Experimentally, this approach demands excellent resolution on the hadronic mass.
- $q^2 > (M_B - M_D)^2$
An analysis of the spectrum of the lepton invariant mass, where again charm contamination can be excluded with a wise cut on q^2 , leaving behind a reasonable fraction of the $b \rightarrow u$ signal [51]. In the region of phase space selected with this method, the theoretical calculation seems to be on excellent footing, but the scale of the intermediate physics is m_c instead of m_b , so $1/m_b$ correction terms blow up to $1/m_c$. Success on the experimental side depends on the resolution with which the lepton and neutrino kinematics can be reconstructed.

The issue in each case is addressing the perturbative and non-perturbative contributions to the decay rate in the selected region of phase space, and then estimating consistent errors for various correction terms. The general consensus [1] is that, due to the complex—and in some cases, correlated—issues that trouble these approaches, the best course is to pursue them in parallel, applying improvements as our knowledge of $b \rightarrow u \ell \nu$ grows. A well-constrained value for $|V_{ub}|$ will only be possible if our picture of $b \rightarrow u \ell \nu$ has been firmly established and tested in several independent ways.

3.3.1 Lepton Energy E_ℓ

With a precise prediction of the lepton energy spectrum for charmless semileptonic B decays, experiments could readily extrapolate from the partial rate measured in the endpoint region¹⁰ to the total inclusive rate and then extract $|V_{ub}|$

¹⁰The “endpoint region” as it is used here and in what follows refers to the region of phase space near the upper kinematic endpoint for the lepton energy E_ℓ in inclusive

according to Eqn 3.37 or similar. Typical values of such a cut accept only about 10% of the total rate. A clean theoretical analysis, however, is limited by the breakdown of the usual heavy quark expansion in the endpoint region, where an additional expansion parameter $\Lambda/(m_b - 2E_\ell)$ becomes important, making it hard to evaluate what fraction of the total rate has been sampled by a given cut. In what follows, we first briefly describe the nature of the problem, and then outline the proposal of Neubert [35], followed by others, to collapse the non-perturbative physics into a shape function that can be used to smear the quark-level decay into the real hadronic process. This shape function turns out to be universal in B -decays (up to higher-order corrections), and arises naturally in the discussion of the radiative decays $b \rightarrow s\gamma$, where it can be experimentally determined.

Difficulties in the Endpoint Region

An understanding of the lepton energy spectrum in the endpoint region begins with a calculation of the differential rate $d\Gamma/dE_\ell$. To leading order in the non-perturbative corrections, the standard theoretical result, ignoring all QCD perturbative corrections, is ([35, 38])

$$\frac{1}{2\Gamma_0} \frac{d\Gamma}{dy} = yF(y)\Theta(1-y) - \frac{\lambda_1 + 33\lambda_2}{6m_b^2} \delta(1-y) - \frac{\lambda_1}{6m_b^2} \delta'(1-y), \quad (3.38)$$

where the new variable

$$y = \frac{2E_\ell}{m_b} \quad (3.39)$$

parameterizes the differential rate in terms of the excursion from the free-quark kinematic limit of $y = 1$. The factor Γ_0 is similarly the free-quark decay width

$$\Gamma_0 = \frac{G_F^2 |V_{ub}|^2}{192\pi^3} m_b^5. \quad (3.40)$$

The symbol $\delta'(\cdot)$ denotes the derivative of the usual Dirac Delta function $\delta(\cdot)$.¹¹ The function $F(y)$ is a slowly-varying function of y given by

$$F(y) = (3 - 2y)y + \frac{5y^2}{3} \frac{\lambda_1}{m_b^2} + (6 + 5y)y \frac{\lambda_2}{m_b^2}. \quad (3.41)$$

$b \rightarrow u\ell\nu$ decays; this is the region free of charm and has size Λ/m_b . Here Λ is again taken as the typical low-energy scale of QCD.

¹¹Recall that the δ -function is rigorously defined in the theory of generalized functions (or distributions) as a linear functional L that acts on the space of more ordinary functions f (usually square-integrable), with an appropriate integral definition of the action $L(f)$. The derivative δ' of the δ -function can be defined in an analogous way, where a formal integration by parts makes it clear that while $\delta(f) = f(0)$, $\delta'(f) = -f'(0)$.

Once again, we note that there are no corrections of order $1/m_b$ due to the natural requirement that there be no residual mass term for the heavy quark field in the effective theory.

The step function is present to enforce the turn-off of the decay distribution above the tree level endpoint in the parton model. The more singular terms (the δ -function and its derivatives) arise from higher-order terms in the $1/m_b$ expansion that have y -derivatives of lower-order terms. All together, the appearance of this infinite series of singularities signals the non-perturbative effects that shift the endpoint from the free-quark endpoint to the the physical one $E_{\ell,\max}^{\text{phys}} = M_B/2$, and illustrates the difficulty in analyzing the shape of the theoretical distribution in this region; see Fig 3.2. The $1/m_b$ expansion breaks down in this region where $1 - y \sim \Lambda/m_b$, or equivalently, where $(m_b - 2E_\ell) \sim \Lambda$. The program for determining $|V_{ub}|$ is thus derailed by the fact that the experimental cuts made to suppress charm background force us directly into the regime where the terms in the expansion (3.38) are formally $\mathcal{O}(1)$ and must be resummed to all orders.

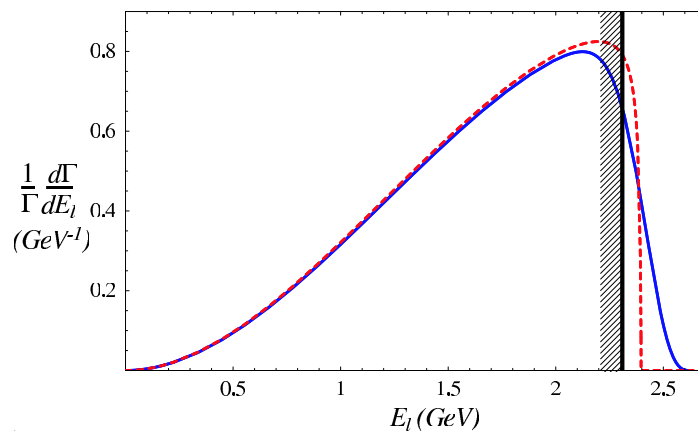


Figure 3.2: An illustration of the importance of the non-perturbative physics governing the binding of the b quark in the B meson. The red curve (dashed) shows the parton-level result for the lepton energy spectrum in $b \rightarrow u \ell \nu$, with endpoint at $m_b/2$. The blue curve (solid) shows the physics spectrum obtained by implementing the non-perturbative physics according to a model for the Fermi motion of the b quark, with endpoint higher at $M_B/2$. As is clear, the details of the hadronization significantly affect the decay rate in the region above the endpoint for semileptonic decays to charm at 2.3 GeV, indicated by the black vertical line.

To further elucidate this problem and understand the formal solution (rather, the rigorous parameterization of our ignorance), we separate the expression for the differential rate into two pieces according to the following prescription (first

suggested by Neubert) [35]:

$$\frac{1}{2\Gamma_0} \frac{d\Gamma}{dy} = F(y)\Theta(1-y) + S(y) \quad (3.42)$$

where now the function $S(y)$ captures all of the non-perturbative physics. When constructing the operator product expansion for $S(y)$, an additional expansion parameter $\Lambda/[m_b(1-y)]$ arises in addition to the usual one, Λ/m_b . Over most of the kinematically allowed region, this new parameter is small, until one approaches the endpoint region, where $(1-y)$ is of order Λ/m_b and the expansion becomes ill-behaved. From our discussion above, it should be clear that the support for the function $S(y)$ is restricted to a small interval of size 2Δ around $y = 1$, where Δ is of size Λ/m_b .

Introduction of a Shape Function

In the original experimental analyses of the lepton energy spectrum [124, 128], various models were employed to estimate the fraction of the total charmless semileptonic rate in the endpoint region, which could then be used to translate a partial rate measured in the endpoint region into an absolute total rate. Then a formula like Eqn 3.37 could be applied to determine $|V_{ub}|$. However, the value for $|V_{ub}|$ extracted in this way is heavily dependent on the model that is used, and there is also no clear prescription for evaluating the effective systematic error inherent in the use of one model over another.¹² Although these initial observations indicated quite clearly that $|V_{ub}| \neq 0$, there was little that could be done to either properly assess or improve the associated theoretical uncertainties.

The first theoretical inroads on this problem were made by Neubert [35], who was able to resum the *leading* singular contributions to the function $S(y)$ to all orders in $\Lambda/[m_b(1-y)]$.¹³ The result is a smooth—but rapidly varying—shape function or form factor that describes the fall-off of the lepton spectrum in the endpoint region in a genuinely non-perturbative way.

A formal definition of the shape function $S(y)$ can be obtained from an operator product expansion of the differential decay rate, in conjunction with an expansion of the hadronic matrix elements in powers of $1/m_b$, as provided by HQET. The result yields a formal expression for $S(y)$ as a sum of forward matrix elements of

¹²The original CLEO paper [128] considered a range of models that gave a spread of more than 50% in $|V_{ub}|$, but could find no defensible recipe for translating this spread into some sort of confidence interval around the central value.

¹³Mannel and Neubert extend this in Ref [37] to include the case of final-state quarks with finite mass $m_q > 0$, which introduces a parameter $\rho = m_q^2/m_b^2$ in the new shape function.

local operators:

$$S(y) = \sum_{n=1}^{\infty} \frac{1}{n!} \frac{A_n}{m_b^n} \delta^{(n-1)}(1-y) + \text{less singular terms} \quad (3.43)$$

where the coefficients A_n are as defined in Ref [35] and contain the aforementioned hadronic matrix elements. The conclusion of this analysis is that, by comparing to the initial expression for the rate in Eqn 3.38, the first few moments of the shape function $S(y)$ can be deduced, leading to the (non-trivial) identification

$$A_0 = 1, \quad A_1 = 0, \quad \text{and} \quad A_2 = -\lambda_1/3. \quad (3.44)$$

These first few moments already significantly constrain the broad features of the shape function; while higher-order moments have a small impact on integrated quantities such as the total decay rate, they can still substantially alter the resulting spectral shape in the endpoint region. It is important to note that in this region, the theoretical spectrum can be drastically unfaithful to the real one, and yet—frustratingly so—this area is the same place where experimentalists must focus to fight contamination from leptons produced in $b \rightarrow c$ decays.

Later treatments of this same approach develop a function $f(k_+)$ that determines (to leading order in $1/m_b$) the probability of finding a b quark with residual light-cone momentum k_+ inside the B -meson, *i.e.*

$$b_B(x)dx = \{f(k_+) + \mathcal{O}(1/m_b)\}dk_+ \quad (3.45)$$

where k_+ and x are related through $k_+ = M_B x - m_b$. In terms of the formalism just developed, this light-cone distribution function is defined by

$$f(k_+) = \langle B(v) | \bar{h}_v \delta(k_+ - iD_+) h_v | B(v) \rangle \quad (3.46)$$

and is naturally related to the shape function discussed previously:

$$\begin{aligned} \Theta(1-y) + S(y) &= \left\langle \Theta \left(1 - y + \frac{in \cdot D}{m_b} \right) \right\rangle + \text{less singular terms} \quad (3.47) \\ &= \int dk_+ \Theta \left(1 - y + \frac{k_+}{m_b} \right) \{f(k_+) + \mathcal{O}(1/m_b)\} \end{aligned}$$

Once again, it is the moments of this new shape function that are related directly to matrix elements in QCD; one can write [36]

$$f(k_+) = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} A_n \delta^{(n)}(k_+) = \delta(k_+) - \frac{\lambda_1}{6} \delta''(k_+) - \frac{A_3}{6} \delta'''(k_+) + \dots \quad (3.48)$$

where

$$A_n = \int dk_+ k_+^n f(k_+) = i^n \tilde{f}^{(n)}(0) = \langle B(v) | \bar{h}_v (iD_+)^n h_v | B(v) \rangle, \quad (3.49)$$

and $\tilde{f}^{(n)}(0)$ stands for the evaluation of the n -th derivative of the Fourier transform $\tilde{f}(t)$ at the point $t = 0$. Ref [37] presents an analysis of the first few of these moments, and concludes that the light-cone structure function is centered around $k_+ = 0$ with a width of order 200-300 MeV determined by the average kinetic energy of the b quark within the B -meson. It also follows that in the heavy quark limit, the range of k_+ is given by $-\infty < k_+ \leq \bar{\Lambda}$, and that $f(k_+)$ should be exponentially small for $k_+ \ll \bar{\Lambda}$.

The decay rate can now be expressed more clearly as a convolution of the simpler parton-level result with this function $f(k_+)$, with the requirement that in the partonic result, the b quark mass m_b is replaced by the mass $m_b^* = m_b + k_+$ [37].

$$\frac{d\Gamma}{dE_\ell} = \int dk_+ f(k_+) \frac{d\Gamma_{\text{parton}}}{dE_\ell} \Big|_{m_b \rightarrow m_b^*} \quad (3.50)$$

We thus discover that the leading bound-state corrections amount to an averaging of the parton-model rate for the decay of a quark with mass $m_b^*(k_+)$ over the distribution $f(k_+)$. The free-quark result is recovered in the obvious limit $f(k_+) \rightarrow \delta(k_+)$.

The analysis of the endpoint region is thus reduced to the determination of a smearing function that captures the non-perturbative physics of the b quark's Fermi motion within the B meson. Further, it is faithful to the underlying theory of QCD in a clear, definite way. However, the theoretical considerations do not deliver the smearing function in explicit form; rather, it is loosely determined by the values of its first few moments, which are related to quantities like m_b (or $\bar{\Lambda}$) and λ_1 .

Further work with the shape function resorts to the building of toy models that mock up the effects of the Fermi smearing. Functional forms are chosen to conform to the general limiting behavior described above, and free parameters are tuned to comply with the few constraints on the moments (*e.g.* as listed in Eqn 3.44). Three models as found in the recent literature are compared in Fig 3.3 [99, 100].

It is worth emphasizing the physical picture that has resulted from this analysis. The shape function $S(y)$, or equivalently the light-cone distribution function $f(k_+)$, was introduced to capture the non-perturbative subtleties in the lepton-energy spectrum near its endpoint; effectively, the infinite number of (leading) singularities in the perturbative expansion are summed into a single, unknown function. A surprisingly intuitive interpretation results: the the parton-level result is averaged or smeared out by a weighting function that encodes the non-perturbative bound-state effects present within the environment of the B -meson. Further, as we might expect, this smearing function contributes only over a range $-\bar{\Lambda} < k_+ < \bar{\Lambda}$, and pushes the parton-level endpoint (set by m_b) up to the physical endpoint (set by M_B). As Falk [38] notes, it is easy to see how this change comes about. If the energy of the b quark is allowed to fluctuate from its on-shell value, then it will occasionally have an energy larger than its free value m_b . This fluctuation can

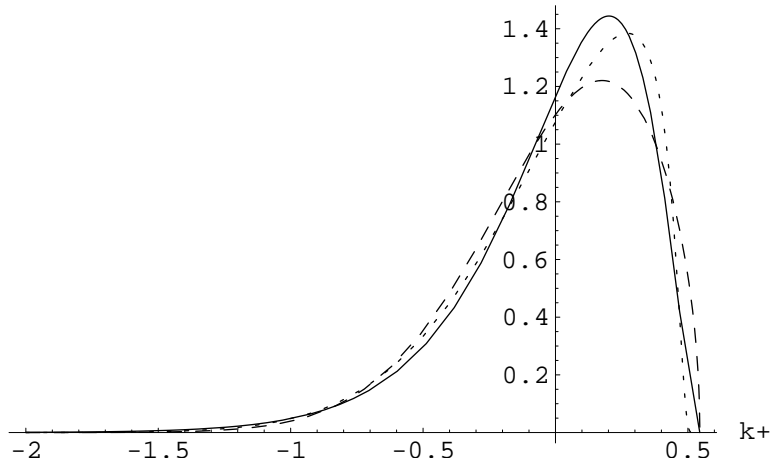


Figure 3.3: Comparison of three models for the non-perturbative shape function $f(k_+)$. Parameters are chosen to meet the requirements $\bar{\Lambda} = 0.545$ GeV and $\lambda_1 = -0.33$ GeV². The solid line is the exponential form; dashed, Gaussian; dotted, Roman [100].

be considered a result of a situation in which the quark has temporarily absorbed some energy (of order Λ) from the light degrees of freedom in the B -meson. If the quark decays at this moment, the resulting lepton is capable of having an energy greater than the limit $E_\ell = m_b/2$ set in the parton model.

The shape function describes the Fermi motion inside the B meson, and as such, is a universal property of B decays, independent of the particular decay channel. Consequently, it can in principle be extracted from a different heavy \rightarrow light process and then employed in inclusive $B \rightarrow X_u \ell \nu$. Such additional input from experiment further reduces the dependence of this method on shape-function modeling, as we shall see next.

Connection to $b \rightarrow s\gamma$

Much of the work outlined above is not specific to the case of inclusive $b \rightarrow u \ell \nu$ decays, suggesting that there may be some universal features which will help resolve the unknown details of the non-perturbative shape function $f(k_+)$. An analysis addressing the possibility of using the photon spectrum from radiative $b \rightarrow s\gamma$ decays was first presented by Neubert [36] and subsequently further explored and expanded by Leibovich *et al.* and others [40–43]. (These radiative decays are interesting in their own right because they are induced by penguin diagrams with virtual top or charm quark exchange. Interestingly, the experimental analysis also requires a hard experimental cut, this time on the photon energy E_γ to limit the systematic error from the $B\bar{B}$ background subtraction.)

In a manner similar to the one applied in the semileptonic case, the leading non-perturbative corrections to the photon spectrum can be summed to all orders in $1/m_b$. The resulting spectrum is determined by an analogous fundamental structure function describing the distribution of the light-cone momentum of the b quark within the hadron.¹⁴ A connection between the radiative and semileptonic decays is readily made at this point.

We omit most of the steps taken in arriving at the $b \rightarrow s\gamma$ spectrum; they are essentially an application of the inclusive formalism described in Section 3.2.3. A spectral function $s(y, \rho)$ is introduced to describe the inclusive spectrum according to

$$\frac{d\Gamma}{dy} = \frac{G_F^2 \alpha m_b^5}{32\pi^4} |V_{tb}V_{ts}^*|^2 |c_\gamma(m_b)|^2 (1+\rho)(1-\rho)^3 \eta_b s(y, \rho) \quad (3.51)$$

where here $y = 2E_\gamma/m_b$, $\rho = m_s^2/m_b^2$, and the parameter η_b captures the leading-order non-perturbative effects. Once again the leading singular terms in $s(y, \rho)$ are summed to all orders to arrive at [36]

$$\begin{aligned} s(y) &= \left\langle \delta \left(1 - y + \frac{iD_+}{m_b} \right) \right\rangle + \text{less singular terms} \\ &= \int dk_+ \delta(1 - y + \frac{k_+}{m_b}) \{f(k_+) + \mathcal{O}(1/m_b)\}, \end{aligned} \quad (3.52)$$

which is reminiscent of the corresponding expression arrived at in the analysis of $b \rightarrow u\ell\nu$, Eqn 3.47. This formal resemblance is a demonstration of the common role the shape function plays (to leading order) in these two different decay processes.

The experimental approach to date has pursued the use of information from the observed $b \rightarrow s\gamma$ photon spectrum to further constrain the unknown attributes of the shape function [128]: a candidate form for $f(k_+)$ is chosen, and its parameters are determined by a fit to the experimental data.

Alternatively, the parallels between the two analyses can be exploited directly at the phenomenological level, leading to the remarkable relation [36]

$$\left| \frac{V_{ub}}{V_{cb}} \right|^2 \simeq \left| \frac{V_{ub}}{V_{tb}V_{ts}^*} \right|^2 = \frac{3\alpha}{\pi} |c_\gamma(m_b)|^2 \eta_{QCD} \frac{\hat{\Gamma}_u(E_0)}{\hat{\Gamma}_s(E_0)} + \mathcal{O}\left(\frac{\Lambda_{QCD}}{m_b}\right). \quad (3.53)$$

Here the η_{QCD} factor contains radiative corrections, and the two integrals over the endpoint region are defined via

$$\hat{\Gamma}_u(E_0) \equiv \int_{E_0}^{\infty} dE_\ell \frac{d\Gamma(B \rightarrow X_u \ell \nu)}{dE_\ell} \quad (3.54)$$

$$\hat{\Gamma}_s(E_0) \equiv \frac{2}{M_B} \int_{E_0}^{\infty} dE_\gamma (E_\gamma - E_0) \frac{d\Gamma(B \rightarrow X_s \gamma)}{dE_\gamma}, \quad (3.55)$$

¹⁴The appearance of this function is the way that QCD naturally accounts for the ‘‘Fermi motion’’ of the initial state quark within the meson; in the past, various models have introduced this artifact by hand in a rather *ad hoc* fashion.

where E_0 is the experimental cut applied in the lepton endpoint region. The conclusion is that a measurement of the integrated quantities $\hat{\Gamma}_u(E_0)$ and $\hat{\Gamma}_s(E_0)$ yields a direct determination of the ratio $|V_{ub}/V_{cb}|$. The value of this technique is that intermediate extraction of the non-perturbative structure function is side-stepped completely. In addition, the use of the ratio helps reduce various hadronic uncertainties to the level of smaller $1/m_b$ power corrections.

Prospects

The basic strategy for determining $|V_{ub}|$ by looking at high momentum leptons dates back to the original observation of charmless semileptonic B decays, and has greatly evolved in experimental and theoretical sophistication and precision since then. However, even with the parameterization of our non-perturbative ignorance with a QCD-based shape function, whose details can be extracted from experimental features of $b \rightarrow s\gamma$, the approach still suffers from three as-yet unavoidable weaknesses [14]:

1. The identification between the shape function in $B \rightarrow X_u \ell \nu$ and $B \rightarrow X_s \gamma$ decay only holds to leading order in the $1/m_b$ and α_s expansions [14]. Perturbative QCD and “higher-twist” corrections inevitably confound the “universal” nature of the shape function. These complications can be ameliorated slightly by the use of relations like that in Eqn 3.53 [42], which avoids the extraction of the shape function as an intermediate step, but its application is limited on the experimental side by the fact that current lepton spectrum measurements are in the $\Upsilon(4S)$ rest frame rather than the B frame.¹⁵
2. There are contributions specific to $b \rightarrow u \ell \nu$ that are still not calculated or not calculable in principle, any of which could be quite sizeable. These include contributions from dimension-6 operators such as the one describing weak annihilation and other sub-leading corrections to the shape function. (See later Sec 3.4.2 for more discussion.)
3. The endpoint region contains such a small part of the full $b \rightarrow u \ell \nu$ phase space that it may be vulnerable to violations of (local) quark-hadron duality. (See later Sec 3.4.1 for more discussion.)

Experimental interest in this approach has not wavered, and the analysis techniques have seen considerable sophistication as well. Recently, CLEO and the B factories [128–130] have released a new round of results that combine the latest information from inclusive $b \rightarrow s\gamma$ with measurements of the semileptonic rate above

¹⁵With the advent of B -tagging and other- B reconstruction methods at the B -factories, the application of these relationships or improved ones will only become more practical.

~ 2 GeV to extract $|V_{ub}|$ with much better-controlled uncertainties than ever before. They deliver values for $|V_{ub}|$ with a properly-estimated precision at better than the 20% level.

3.3.2 Hadronic Mass M_X^2

One of the disadvantages of working in the endpoint region of the lepton energy spectrum is that only about 10% of $b \rightarrow u \ell \nu$ decays lie above the kinematic limit for charm. A different and kinematically more direct approach was first suggested by Barger *et al.* in 1990 [44]. Recall that the essential idea behind the lepton energy cut $E_\ell > (M_B^2 - M_D^2)/2M_B$ is to exclude semileptonic decays where a heavy, charm-containing meson is produced. But since there is no unique, single-valued mapping between M_X and E_ℓ (due to the presence of the neutrino), the cut on E_ℓ is not particularly efficient at implementing the implicit bound on M_X . Indeed, a more efficient method is to reconstruct the hadronic system and apply the intended cut directly, requiring $M_X < M_D$. The gain in efficiency is significant: a simple cut $M_X < M_D$ suffices to eliminate the lightest possible charm-containing hadronic system, yet retains as much as 90% of the $b \rightarrow u$ events. On the experimental side, one needs to somehow infer the invariant mass of the recoiling hadronic system, a more demanding task than identifying high-momentum leptons. Resolution issues will force the actual cut on the reconstructed mass to be somewhat lower than M_D to reduce leakage from charm down into the signal region for $b \rightarrow u \ell \nu$ thus defined.

The extraction of $|V_{ub}|$ proceeds directly from a calculation of the differential rate $d\Gamma/dM_X$ (or equivalently, $d\Gamma/dM_X^2$). An experiment essentially counts the number of $b \rightarrow u$ decays with $M_X < M_X^{\max}$, which translates to a measurement of the partial rate for that mass window. If the fraction of the total rate that falls into the given window is well-predicted theoretically, an extrapolation to the total rate can be done readily, and then Eqn 3.37 once again provides access to $|V_{ub}|$. The clear obligation on the theoretical side is thus to provide this fraction

$$\Phi(M_X^{\max}) = \frac{1}{\Gamma(B \rightarrow X_u \ell \nu)} \int_0^{M_X^{\max}} dM_X \frac{d\Gamma}{dM_X}, \quad (3.56)$$

with reasonable uncertainties.

The advantage in the theoretical evaluation of the spectrum $d\Gamma/dM_X^2$ compared to $d\Gamma/dE_\ell$ is that a cut on M_X^2 does not preferentially weight the light hadronic states: all states satisfying $M_X < M_D$ are accepted on equal footing. This selection contrasts with the case of a cut on the lepton energy, where the endpoint is dominated by contributions from the lightest states, such as π and ρ . Consequently, it is much more likely that the first few terms of the OPE will be an adequate description of the decay rate in the region $M_X^2 < M_D^2$ than in the region

$E_\ell > (M_B^2 - M_D^2)/2M_B$. This observation also leads to the expectation that the (restricted) mass spectrum will be less sensitive to duality violations.

Standard application of the OPE results in a differential rate $d^2\Gamma/ds_h dE_h$ calculated in perturbation theory to some order, in terms of the *partonic* variables $s_h \equiv m_h^2$ and E_h , the invariant mass and energy of the partons arising from the b quark decay. By integrating over the allowed range for the partonic energy E_h , one then naturally obtains the spectrum in the partonic invariant mass variable s_h . However, in order to apply a cut on the physical hadronic mass, one needs to change to the hadronic variables

$$\frac{s_H}{m_b^2} = s_h + \epsilon E_h + \epsilon^2 \quad (3.57)$$

$$\frac{E_H}{m_b} = E_h + \epsilon, \quad (3.58)$$

where $\epsilon = \bar{\Lambda}/m_b$ captures the difference between the partonic and hadronic variables.

Unfortunately, after this transformation, a double logarithmic singularity appears at $s_H = \bar{\Lambda}M_B$, and at higher orders, more logarithmic singularities appear [45–50]. These problems make the differential rate $d\Gamma/ds_H$ hard to predict reliably even in perturbation theory. The behavior of the spectrum near this singularity is less important for observables that average over larger regions of the spectrum significantly beyond $\bar{\Lambda}M_B$. Thus keeping the experimentally-imposed cut as high as possible is crucial to control of the perturbative expansion. The problem is that for reasons of finite experimental resolution, the ideal cut at $M_D^2 \approx 3.5 \text{ GeV}^2$ will need to be lowered, bringing it very close to $\bar{\Lambda}M_B$, which is numerically already at the same scale.

Non-perturbative effects—which are significant across the entire low mass region $s_H \leq \bar{\Lambda}M_B$ —are applied using the usual smearing procedures developed in the analysis of the lepton endpoint region. Since the shape function is essentially unknown, one can only study these effects only in the context of toy models. As in the lepton endpoint region, the smearing has a significant effect on the spectrum, re-distributing the rate across any cut at the scale of M_D^2 . Fig 3.4 illustrates the effects of this non-perturbative physics with a model implementation of the shape function.

Alternatively, as suggested by Leibovich *et al* [48], one can again consider the use of radiative $b \rightarrow s\gamma$ decays to determine the non-perturbative structure function and instead extract the ratio $|V_{ub}|/|V_{ts}|$. Much like the suggestion in the endpoint analysis, the idea is to eliminate ignorance of the Fermi motion of the heavy quark by incorporating measurements from another B decay that involves the same fundamental distribution function. But this procedure only mitigates the crippling ignorance of the non-perturbative effects, and cannot resolve the fundamental problems in the perturbative expansion.

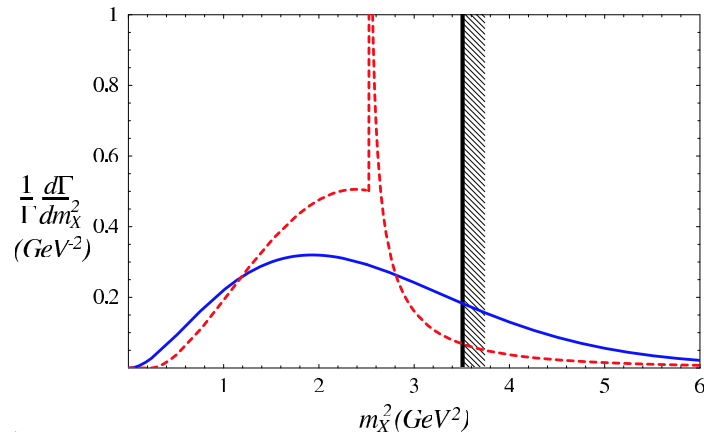


Figure 3.4: Smoothing effect of non-perturbative shape function on the hadronic mass spectrum. The red (dashed) curve shows the first-order perturbative result, after translation to hadronic variables, giving rise to a singularity at $s_H \sim \bar{\Lambda}M_B$. The blue (solid) curves shows the physical spectrum after convolution with the shape function that effectively implements the Fermi motion of the b quark in the initial-state B meson.

In summary, a straightforward cut $M_X < M_D$ is quite efficient at singling out $b \rightarrow u$ decays from $b \rightarrow c$ and accepts a large patch of phase space. This last fact leads one to hope that the analysis will be less sensitive to local duality violations. However, as we have seen, the ability to even perform the theoretical calculations is extremely sensitive to the placement of the experimental cut. The numerical coincidence that $M_D^2 \simeq \bar{\Lambda}M_B$ gives rise to significant non-perturbative corrections in the region retained by the cut, and perturbative effects lead to an unphysical singularity at the same point. The result is a direct conflict between the experimental need to apply a more restrictive cut $M_X^{\text{cut}} < M_D$ and the theoretical desire to save the calculation by raising the cut as high as possible.

3.3.3 Dilepton Mass q^2

Bauer, Ligeti, and Luke [51, 52] have proposed an inclusive determination of $|V_{ub}|$ from the dilepton invariant mass spectrum. With a cut $q^2 > (M_B - M_D)^2$, manifestly above the charm endpoint, $b \rightarrow c$ decays can be entirely eliminated while retaining close to 20% of the $b \rightarrow u$ rate. As just discussed, the cuts on lepton energy or hadronic mass required to isolate $b \rightarrow u \ell \nu$ from $b \rightarrow c \ell \nu$ spoil the convergence of the OPE, and hence require an infinite set of non-perturbative contributions to be resummed into an unknown structure function. In contrast, this proposal uses an observable for which the OPE does not break down in the region that is free from the charm background.

The key benefit is that in the restricted region of phase space singled out by a cut on q^2 , the influence of the non-perturbative shape function is reduced and the OPE expansion is much better-behaved. The non-perturbative smearing of the q^2 -spectrum is demoted to a sub-leading effect; indeed, the nature of the cut on q^2 is to forbid the hadronic final state from moving fast in the B rest frame, so the light-cone expansion which gives rise to the shape function is rendered irrelevant. Non-perturbative effects are only significant at large values of q^2 , where the difference between the hadronic and partonic endpoints is roughly $M_B^2 - m_b^2 \simeq 2\bar{\Lambda}M_B$, and the contribution of these additional terms will only be important if the spectrum is integrated over a region of similarly small width. Fig 3.5 illustrates these claims with a comparison of the parton-level q^2 spectrum before and after the application of non-perturbative smearing via a model shape function.

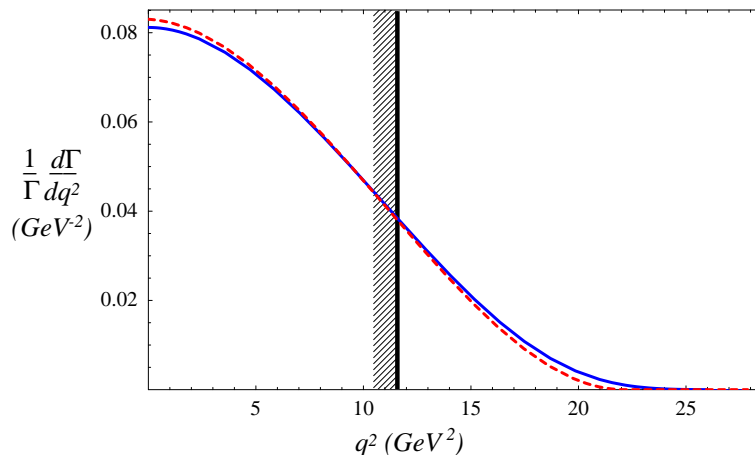


Figure 3.5: Impact of non-perturbative “smearing” on the parton-level q^2 spectrum. The interpretation of colors and line styles is the same as in the previous two figures. The message here is that the non-perturbative corrections are parametrically suppressed with this approach, so that shape function effects are far less important.

In the initial proposal, Bauer *et al.* calculate the fraction $F(q_0^2)$ of all $b \rightarrow u \ell \nu$ events that lie above a cut $q^2 > q_0^2$ by integrating the differential rate $d\Gamma/dq^2$, including perturbative and non-perturbative corrections to $\mathcal{O}(1/m_b^2)$ and $\mathcal{O}(\alpha_s^2\beta_0)$.¹⁶ Normalizing to the well-known full (charmless) semileptonic width, they use (a variation of) the result in Eqn 3.37 to extract $|V_{ub}|$. Re-expressing the dependence on the b quark mass with the use of the epsilon expansion, they find

$$|V_{ub}| = (3.04 \pm 0.06 \pm 0.08) \times 10^{-3} \left[\frac{\mathcal{B}(B \rightarrow X\ell\nu)|_{q^2 > q_0^2} 1.6 \text{ ps}}{10^{-3} F(q_0^2) \tau_B} \right]^{1/2}, \quad (3.59)$$

¹⁶The function $F(q_0^2)$ is the analog for the dilepton mass analysis of $\Phi(M_x^{\max})$ for the hadronic mass analysis.

where the errors are estimates of the perturbative and non-perturbative uncertainties, respectively, in the epsilon expansion. The dominant uncertainty remains the error on the short-distance b quark mass, whichever way it is defined.

For the nominal cut at $q_0^2 = (M_B - M_D)^2 \approx 11.6 \text{ GeV}^2$, they find 18.6% of the full rate is accepted (using $\bar{\Lambda} = 0.4 \text{ GeV}$, $\lambda_1 = -0.2 \text{ GeV}^2$, and $\alpha_s(m_b) = 0.22$). As the cut is raised, the fraction of the rate that is sampled falls rather steeply; for instance, with a cut $q_0^2 = 15 \text{ GeV}^2$, the fraction is reduced to 6.7%. Thus, the advantage of not having to employ information from $b \rightarrow s\gamma$ and the absence of $1/m_b$ power corrections is offset by the fact that the cut eliminates a large fraction of the rate. Worries about duality violations begin to lurk once more. In addition, issues of finite experimental resolution may require the cut to be significantly higher than the charm endpoint value; since the theoretical error on $|V_{ub}|$ in this scheme grows rapidly as the cut is increased, the analysis may be pushed into a regime of low acceptance and large errors.

In a subsequent analysis, Neubert [53, 54] has investigated in more detail the structure of the heavy quark expansion as it applies to the computation of $F(q_0^2)$. He finds that the relevant mass scale μ_c controlling the size of the corrections is really of order the charm quark mass m_c rather than the heavier b mass. This conclusion follows from the observation that the largest values of the hadronic mass and energy accessible in the presence of this cut are only of order m_c .¹⁷ The corrections must then be analyzed using a two-step, hybrid expansion in terms of the parameters μ_c/m_b and Λ/μ_c . The scale μ_c is the typical hadronic momentum scale, of order m_c or less, and depends on the value q_0^2 of the cut. A modified version of HQET can then be used to disentangle the physics associated with the scales m_b and μ_c , while the OPE is sufficient for separating the physics at the scale μ_c from the truly long-distance physics.

Using a renormalization group-improved (RG) result and a potential-subtracted mass m_b^{PS} that improves aspects of the perturbative calculation, Neubert revises the estimates of the various uncertainties considered by Bauer *et al.* in their analysis. In addition to a heightened sensitivity to m_b , he finds that the higher-order power corrections now at the lower scale of $(\Lambda/\mu_c)^3$ are of increased significance. As the cut on q^2 is raised from the efficient value of $(M_B - M_D)^2$, the size of these latter terms changes rapidly from sub-leading to dominant, worsening from an 8% projected error on $|V_{ub}|^2$ to 25% for $q_0^2 = 15 \text{ GeV}^2$; the total estimated theoretical uncertainty on $|V_{ub}|$ grows from 10% to almost 20%. He concludes that the q^2 proposal is sound, but the initial promise of 5–10% resolution on $|V_{ub}|$ is too optimistic. With a reasonable cut in the vicinity of $q_0^2 = 12.5 \text{ GeV}^2$, a precision

¹⁷Starting from Eqn 2.5 defining q^2 in terms of the hadronic variables, we can apply the inequality $q^2 > (M_B - M_D)^2$ and immediately arrive at the bound $E_X \leq M_D + (M_X^2 - M_D^2)/2M_B$, showing that indeed the hadronic energy is of the order of the charm mass. Since the invariant mass of the hadronic system cannot exceed its energy, the same bound applies to the mass M_X .

better than 15% is still theoretically possible.

Recently, a strategy relying on combined cuts on the dilepton and hadronic masses has been proposed [55]. The basis for the suggestion is the observation that a twist expansion is required if the phase space surviving a cut is dominated by hadronic states with energy much larger than their invariant mass. This problem arises for the cases of a cut on E_ℓ or a cut on M_X^2 . However, a cut on the dilepton mass rejects the low q^2 region, and restricts the hadronic energy to $E_X < M_B - \sqrt{q_0^2} < M_D$, which eliminates the “dangerous” region of phase space. In essence, the strategy employs a hadronic mass cut to reduce the charm background while keeping a large fraction of the $b \rightarrow u$ rate, and simultaneously applies a lower cut on q^2 to avoid the shape function region where non-perturbative effects can erode the theoretical calculation. Fig 3.6 illustrates these considerations. Compared to a pure q^2 cut, the uncertainty from unknown $1/m_b^3$ terms in the OPE is significantly reduced and the fraction of $b \rightarrow u$ events retained is roughly doubled; compared to a pure M_X^2 cut, the uncertainty attributed to ignorance of the shape function is significantly reduced as well. The originators of this new mixed strategy estimate that determining $|V_{ub}|$ with a theoretical error of 5–10% should be possible.

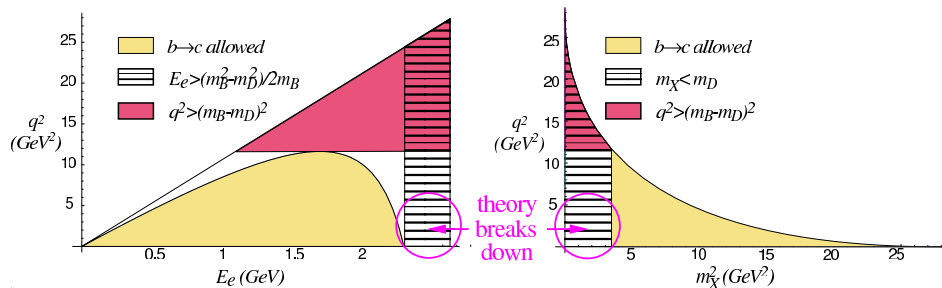


Figure 3.6: A Dalitz plot for $b \rightarrow u \ell \nu$ in the q^2 vs. M_X^2 , E_ℓ planes. The regions free from charm are highlighted, as well as the areas passing various experimental cuts. The circled region in the corner of each phase space plot indicates where the local OPE breaks down and a twist expansion is required. From Ref [55].

3.3.4 Summary

The theoretical challenge in extracting $|V_{ub}|$ is kindled by the experimental requirement that the measurement be made in some restricted region of phase space, where background $b \rightarrow c$ decays are kinematically forbidden. This restriction increases the relevance of non-perturbative corrections, which can be large enough to cripple the usual OPE in these extreme regions. Table 3.1 summarizes this tension by listing the size of the region over which non-perturbative effects are important

along with the size of the region remaining after the phase space cut to capture the $b \rightarrow u$ decays.

Table 3.1: A comparison between the three methods for extracting $|V_{ub}|$ discussed in this chapter. Each strategy relies on a cut on a different kinematic distribution to select a region free from contamination from $b \rightarrow c$ decays; the size of this region is appears in the second column. But for each spectrum, there is also a natural scale at which the contribution of non-perturbative (or hadronic) effects becomes important, indicated in the third column. Since $M_D^2 \sim \Lambda M_B$, the width of this region for the first two strategies is in fact comparable to the charm-free region selected by the cut, causing calculations in this experimentally-accessible region to break down. The final column lists the estimated efficiency for each strategy. Experimental issues are not addressed in this table. (Modeled after a similar table presented in Ref. [51].)

Decay distribution	Region free of charm	Region where had. effects signif.	$b \rightarrow u$ Eff.
$d\Gamma/dE_\ell$	$\Delta E_\ell = M_D^2/2M_B$	$\Delta E_\ell \sim \Lambda$	$\sim 10\%$
$d\Gamma/dM_X^2$	$\Delta M_X^2 = M_D^2$	$\Delta M_X^2 \sim \Lambda m_b$	$\sim 80\%$
$d\Gamma/dq^2$	$\Delta q^2 = 2M_B M_D - M_D^2$	$\Delta q^2 \sim \Lambda m_b$	$\sim 20\%$

The general feature that, numerically, M_D^2 is the same size as ΛM_B is an unfortunate coincidence,¹⁸ since it leads to (large) non-perturbative effects in both the endpoint of lepton energy spectrum, where $E_\ell \sim m_b/2 + \Lambda$, and in the region $M_X \sim M_D$, which suffers equally from the same non-perturbative physics. (Neglecting the neutrino, it is easy to see that $M_X \sim \Lambda M_B \Rightarrow E_\ell \sim M_B/2 \pm \Lambda$.) As we have seen, the non-perturbative effects can be packaged into a universal shape function that describes the Fermi motion of the b quark within the initial-state meson, but this function is essentially unknown beyond its first few moments. The size of the region over which this shape function is smeared controls the size of the contributions from the higher-dimension operators it conceals. The lepton endpoint and the region $M_X < M_D$ in hadronic mass are both inadequate for diluting these contributions, and so the shape of these spectra depend heavily on the details of this function. Both of these methods must rely on input data from *other* inclusive B decays to proceed. The q^2 approach improves on these earlier proposals slightly, but the expansion parameter in the OPE for a q^2 analysis turns out to be of size Λ/m_c , which tends to inflate higher-order contributions, and a cut on q^2 much beyond the charm endpoint is rather inefficient. A combined q^2 - M_X^2 analysis may prove more effective than any of these single-variable analyses.

¹⁸I place the blame for this unlucky state of affairs on the Higgs boson. It's responsible for so much of the trouble in the (particle physics) world today.

The question of which measurement technique to use to determine $|V_{ub}|$ has no single answer. While none of the preceding proposals for extracting $|V_{ub}|$ is markedly superior to the others, the combined q^2 - M_X^2 approach has been well-received as the first step to constructing a strategy that is effective on multiple theoretical fronts while also respecting experimental limitations [2]. The approach has reduced dependence on the shape function yet still includes a sufficient fraction of the full rate to soften concerns about weak annihilation contributions and violations of duality. On the experimental side, the large acceptance and excellent discrimination offered by the hadronic mass cut is valuable, as is the utility of the q^2 cut in rejecting non- $B\bar{B}$ backgrounds.

It is likely that the optimal measurement strategy may only be realized after the fact, when a final analysis of $b \rightarrow u \ell \nu$ is complete [56]. Meeting the challenges posed by $b \rightarrow u \ell \nu$ requires an investigation of theoretical errors in conjunction with experimental uncertainties, for they are deeply entwined in a fundamental and unavoidable way, a feature made startlingly clear in the study of the endpoint of the lepton energy spectrum. Depending on the details of experimental efficiency and resolution, some combination of cuts on all three variables could turn out to be the most effective way to minimize the overall uncertainty on $|V_{ub}|$.¹⁹

The first step toward identifying an optimal strategy is a dedicated and careful analysis of *all* sources of theoretical uncertainty,²⁰ which includes a realistic structure function well-constrained by data, and exhaustive, conservative errors that assess contributions from all expected—and neglected—correction terms.²¹

It would be remiss not to include mention of recent work by Bosch, Lange, Neubert, and Paz that uses the tools of effective field theory to compute, for the first time, a complete next-to-leading order prediction for decay rates and spectra in the shape function region [57]. They employ a systematic treatment that first matches QCD onto soft-collinear effective theory (SCET), integrating out hard quantum fluctuations; then they match SCET onto HQET, integrating out

¹⁹Any experimental analysis will, practically speaking, be forced to cut on all three variables, if only indirectly due to acceptance or other detector effects.

²⁰One could argue that a large part of the progress on $b \rightarrow u \ell \nu$ in the past 20 years has been in recognizing new sources of theoretical uncertainty. It is time to put this program into practice and start delivering.

²¹Such an effort is likely to be tough work, and it's not entirely clear where to start. An old Sufi mondo comes to mind:

Once some pilgrims came upon an old Sufi digging through the dirt on the side of a road. When asked what he was looking for, he replied: "A treasure I have lost." So all the pilgrims joined in and searched until they were hungry and tired. Finally, one of them asked the Sufi if he was sure this was where he lost his treasure. "Oh my, no! I lost it over there, on the other side of those mountains." To which they replied: "Then why in the name of Allah are you looking for it here?" He answered, "Because here there is more light."

the hard-collinear fluctuations. The final theory provides new model-independent insights into the nature of the shape function, significantly reduces the model dependence usually incurred in relating $b \rightarrow s\gamma$ to inclusive $b \rightarrow u\ell\nu$ via the shape function, and suggests a new method for a high-precision determination of $|V_{ub}|$ using the hadronic light-cone momentum variable $P_+ = E_X - |\vec{p}_X|$. Further discussion of this important progress²² is beyond the scope of this work.

3.4 Untamed Uncertainties

The phenomenological understanding of $b \rightarrow u\ell\nu$ has evolved considerably since the first observation in 1990 of high-momentum leptons from B decay beyond the endpoint for $b \rightarrow c\ell\nu$ [123, 124]. Somewhat naïve, model-based extractions of $|V_{ub}|$ have been replaced by a rigorous treatment using the tools of HQET and the OPE, and this framework has allowed for a significant increase in confidence and precision in efforts to extract $|V_{ub}|$ from semileptonic B decay. The breadth and depth of this program has hopefully been amply demonstrated in the preceding discussion of approaches for extraction of $|V_{ub}|$. Nonetheless, there remain some theoretical blindspots that only recently have been clearly identified. Unfortunately, these uncertainties are not yet fully quantified, partly because they are non-parametric by nature, and constraints of such a sort are difficult to come by. The growing consensus in the theoretical community [14], however, is that these lingering issues need to be resolved to meet the hopes for a precision value of $|V_{ub}|$ with well-constrained errors.

The concerns are tied to basic assumptions that are implicit in almost any systematic treatment of inclusive $b \rightarrow u\ell\nu$ decay, and thus are pervasive, ingrained, and correspondingly difficult to isolate and quantify. Although the categorization below is not unique, and could be re-grouped in multiple ways under different headings, we follow [2] and list the most important outstanding theoretical issues as follows:

- **Quark-hadron duality.** Within the framework described in Sec 3.2.3, the assumption of quark-hadron duality underpins the calculation of any inclusive quantity, and in particular, the notion of *local* duality is vital to calculation of differential spectra or partial rates in reduced regions of phase space. Yet it is clear that this duality is by no means guaranteed or exact.²³

²²Arguably, these calculations represent a “quantum leap” forward in theoretical understanding of these decays in the shape function regime. . . . !

²³For instance, high-momentum leptons are clearly seen above the parton-level limit of $m_b/2$ for $b \rightarrow u\ell\nu$ decays, a region where the parton-level rate vanishes. That is, with too tight a lepton energy cut, local duality simply cannot hold—the existence of such leptons would be ruled out!

- **Neglect of weak annihilation.** One contribution to the $B \rightarrow X_u \ell \nu$ decay rate is a series of annihilation-like diagrams similar in part to valence quark annihilation in the purely leptonic decay of a charged B . The magnitude of this process is expected to be small compared to the usual $b \rightarrow u \ell \nu$ rate computed at lower order, but it will be concentrated in a small region of phase space, where its neglect could have significant impact on the extraction of $|V_{ub}|$. *Constraining this uncertainty is the subject of this thesis.*
- **Unknown corrections to the shape function.** Sub-leading corrections to the non-perturbative shape function are still poorly estimated and remain somewhat controversial, but formalisms for estimating these corrections are now emerging.

We elaborate on these issues briefly, trying to place them appropriately in the context already developed.

3.4.1 Quark-Hadron Duality

The notion of quark-hadron duality (or simply “duality”) is rooted in the idea that a quark-level calculation should, in some real sense, approximate physical hadronic observables. It is implicitly invoked repeatedly to connect quantities evaluated at the quark-gluon level to the observable world of hadrons, but, particularly as it applies to the study of $b \rightarrow u \ell \nu$, it stands really only as an approximation, for which the attendant uncertainty still needs to be properly assessed.²⁴

In order to understand the limitations of duality, it must be clear to what exactly the concept refers. Several versions have been described in the literature, and there is now a categorical distinction²⁵ between at least two types: *global* duality and *local* duality. (The discussion followed here is based on that in Ref [14].)

The classic example of the explicit application of duality arises in the discussion of the cross section for the scattering process $e^+e^- \rightarrow$ hadrons. Naively, for large center-of-mass energies $\sqrt{s} \gg \Lambda_{\text{QCD}}$ it would seem that the calculation could be handled perturbatively in terms of quarks and gluons, due to the asymptotic freedom of QCD. However, the presence of production thresholds, such as those for charm, corrupts the straightforward computation of $\sigma(e^+e^- \rightarrow$ hadrons), by introducing relevant but intrinsically non-perturbative physics. The resolution is

²⁴Historically, concerns about the adequacy of duality assumptions were raised essentially at the same time that it was first used in making inclusive calculations at the dawn of the OPE and the recognition of the heavy quark symmetry, specifically in the context of deep inelastic scattering. Its role in inclusive B decays and the potential impact of its violations for values of $|V_{ub}|$ extracted from $b \rightarrow u \ell \nu$, however, remains still incompletely understood.

²⁵The distinction is not without its dissenters; see *e.g.* Shifman’s critique in Ref [59].

to posit that equality between the quark-level result and the physical world holds after an averaging or “smearing” over an appropriate energy interval is applied. In general, for an observable T , the claim is that

$$\langle T^{\text{hadronic}} \rangle_w \simeq \langle T^{\text{partonic}} \rangle_w \quad (3.60)$$

where $\langle \dots \rangle_w$ denotes the smearing using a smooth weight function $w(s)$:

$$\langle \dots \rangle_w = \int ds \dots w(s). \quad (3.61)$$

The extent to which we can trust that $\langle T^{\text{partonic}} \rangle_w$ actually represents the physical quantity $\langle T^{\text{hadronic}} \rangle_w$ depends on the details of the weight function—namely, its width. It might be broad compared to the structures that appear in the hadronic spectrum, or it could be quite narrow; in the extreme, we could have $w(s) \sim \delta(s - s_0)$. The convention is to refer to the first case (averaging over a region large compared to hadronic structure) as relying on *global parton-hadron duality*, and to the second case (point-by-point equality of the hadronic and partonic quantities) as an instance for *local parton-hadron duality*. The nomenclature “parton-hadron” emphasizes that the duality is between the physical, hadronic quantity and the partonic result which is computed from quark *and* gluon contributions. (However, we will still adopt the shorter term “quark-hadron.”) Clearly quantities like total widths will rely on global duality, since they involve integration over all available phase space, while a differential rate that integrates out only some degrees of freedom will rely on the validity of local duality.

The features of duality violation have been clarified by theoretical consideration over the past several years [58–61]. A precise definition of duality requires a more technical treatment than we can afford space and patience for here, but the discussion above in Sec 3.2.3 introduced some of the essential ingredients. The essence of the OPE is an expansion of a combination of Green’s functions as a series of local operators with short-distance coefficients. The entire analysis is conducted in the Euclidean domain, far from any singularities induced by hadronic thresholds, and has to be continued analytically into Minkowski space to connect to observable hadronic quantities. The assumption of duality is essentially the statement that the truncated series in Euclidean space corresponds in a natural way to—or is dual to—the truncated physical calculation in Minkowski space. The extent to which this assumption is flawed measures the extent to which duality is violated. Typically, neglected terms in the Euclidean domain are exponentially small, but under analytic continuation, they are mapped to oscillating functions of momentum transfer which are only power-suppressed to an extent that depends on the details of the process under consideration. These neglected terms embody the local violations of duality.

Duality then is not an additional assumption applied in the construction of the OPE, but a natural result of translating the OPE back to physical space-time

coordinates. Violations of duality hinge on our ignorance of the exact analytic solution to QCD in the Euclidean regime, and the subsequent uncertainty that arises upon the analytic continuation of a *finite* number of terms from the OPE result back into physical Minkowski space.

Bigi and Uraltsev [58] also emphasize that the effects of local duality violation must oscillate as a function of energy scale and have vanishing averages or be exponentially small. So it cannot be blamed for the systematic excess or deficit of decay rates.²⁶

Current progress on understanding duality violations is along two fronts. Theorists are exploring toy models (far simpler than full QCD) for which exact solutions are known and the extent of duality violations can be explicitly determined. Mapping these conclusions back onto QCD is still a challenging enterprise. The second approach is the oft-quoted experimental dictum: determine the same quantity in several independent ways and compare the results. Heavy quark physics lends itself well to such tests, since it can predict numerous decay rates and other spectral properties from only a small handful of basic input parameters (specifically: quark masses and hadronic expectation values). One test of duality in this regard is a comparison of the inclusive and exclusive determinations of the CKM elements $|V_{cb}|$ and $|V_{ub}|$. For both, the inclusive and exclusive values are consistent [2]. Further support for the operator product expansion used in these analyses comes from the general agreement seen for HQET parameters extracted from numerous moments of various semileptonic spectra. However, other uncertainties are also present in these measurements, making it hard to accept this comparison as a yardstick for measuring the size of duality violations, rather than simply as a (limited) test of general theoretical control over the program as a whole.

Violations of duality will have different impact depending on the experimental route taken for the identification of $b \rightarrow u \ell \nu$ and the subsequent extraction of $|V_{ub}|$. It is an adequate but not necessary rule that as a measurement becomes more inclusive, covering more of phase space (uniformly), the impact of duality violations will diminish. In particular, the endpoint analysis may be subject to violations of local duality since it considers such a narrow slice of phase space (only $\sim 10\%$ of the total $b \rightarrow u \ell \nu$ rate), but the hadronic mass analysis should be much less sensitive, since it accepts 80% of the available rate.

In general, recent attempts to quantify this particular uncertainty suggest that it will be sub-dominant compared to other theoretical uncertainties that currently limit our knowledge of $|V_{ub}|$. It is likely that future understanding of duality

²⁶Before theoretical scrutiny was brought to bear on the question of duality, it was proposed that the discrepancy between the measured and expected B semileptonic widths could perhaps be attributed to duality violation. The current understanding rules out this possible escape hatch, since the extent attributable to duality violations is constrained to be less than a fraction of a percent [14].

violations will keep pace with attempts to reduce these other theoretical errors, suggesting that uncertainty about quark-hadron duality is unlikely to ever be the limiting factor in efforts to measure $|V_{ub}|$.

3.4.2 Weak Annihilation

When the OPE for the charmless semileptonic width is carried to higher order in Λ/m_b , new terms naturally emerge. A few of these can be broadly characterized as “spectator-dependent” corrections to the lower-order terms since, for the first time, the light spectator quark in the B meson becomes involved. Some of these corrections are simply the $1/m_b$ -suppressed pieces of the kinetic and chromomagnetic operators, and are expected to be rather small compared to their leading-order contributions [63]. But two new operators arise as well: One is the so-called Darwin term,

$$\rho_D^3 = -\frac{1}{2M_B} \langle B | \frac{g_s^2}{2} \bar{b} \gamma_\alpha t^a b \sum_q \bar{q} \gamma_\alpha t^a q | B \rangle, \quad (3.62)$$

and is estimated to be a small (downward) correction to the total width of about $-(1-2)\%$ [63]. The second operator is also of a four-fermion form, but includes only the u quark and has a different color and Lorentz structure:

$$\frac{1}{2M_B} \langle B | \bar{b}_L \gamma_\alpha u_L \bar{u}_L \gamma_\beta b_L | B \rangle (\delta_{\alpha\beta} - v_\alpha v_\beta), \quad (3.63)$$

where the new quantity $v_\mu = p_\mu^B/M_B$ is the four-velocity of the B . This term describes the process of *weak annihilation* in the B meson, similar to the leptonic decay of the B^\pm where the constituent b and u quarks annihilate. In general, it affects charged and neutral B mesons differently, and so can lead to a decay rate difference between B^\pm and B^0 mesons. The magnitude of this contribution cannot be formally calculated, but some initial qualitative observations can be made. Because there are only two particles in the final state, the term is enhanced by a factor of $16\pi^2$ relative to the usual terms in the OPE for $b \rightarrow u \ell \nu$. Further, the weak annihilation channel contributes formally at the endpoint of the q^2 spectrum,

$$\frac{d\Gamma_{\text{WA}}}{dq^2} \sim \delta(q^2 - m_b^2). \quad (3.64)$$

The relative contribution of this effect is thus maximal for a partial rate measurement that focuses only on the q^2 endpoint, and is only diluted by the inclusion of additional phase space. In other words, the corresponding uncertainty from our ignorance of weak annihilation actually *grows* as a q^2 -like or E_ℓ -like cut is tightened (so long as the high q^2 region is included). The localization of this process to a tiny region of phase space makes it effectively independent of cuts on M_X^2 or

q^2 . In particular, there is potential for a large impact on an endpoint measurement, which is typically confined to the same small region where weak annihilation contributes in its entirety.

The contribution from weak annihilation vanishes in the QCD factorization approximation, so estimates of its size can be made by exploring violations of this hypothesis [62–65]. Specifically, the contribution to the total charmless semileptonic width from WA reads as [65]

$$\delta\Gamma(B \rightarrow X_u \ell \nu) = \frac{G_F^2 |V_{ub}|^2 f_B^2 m_b^2 M_B}{12\pi} (B_2 - B_1) \quad (3.65)$$

where f_B is the B meson annihilation constant, and the phenomenological “bag constants” $B_{1,2}$ parameterize the matrix elements of two four-quark operators as shown below,

$$\begin{aligned} \langle B | \bar{b}_L \gamma_\alpha u_L \bar{u}_L \gamma_\alpha b_L | B \rangle &= \frac{f_B^2 M_B}{8} B_1 \\ \langle B | \bar{b}_L \gamma_\alpha u_L \bar{u}_L \gamma_\alpha b_R | B \rangle &= \frac{f_B^2 M_B}{8} B_2 \end{aligned} \quad (3.66)$$

In the limit of naïve factorization, the product of the bilinear operators is saturated by the simple vacuum insertion ($|0\rangle\langle 0|$), and the two parameters are equal, depending only on the flavor of the light quark in the B meson. For B_u , the result is $B_1 = B_2 = 1$, while for $B_{d,s}$, $B_1 = B_2 = 0$. Clearly, in this limit, the effect on the rate expressed in Eqn 3.65 vanishes.

However, the conventional assumption in the literature is that the factorization approximation holds to no better than about 10% accuracy; that is, correctly handling non-factorizable terms²⁷ will amount to corrections of about this magnitude. The standard arena for estimating the validity of factorization is charmed D mesons, where violations can potentially explain differences between the expected and observed values for the D semileptonic width, and perhaps also lifetime differences between the D_s and the D .²⁸ Using ingredients from several other theoretical investigations, Uraltsev estimates [63] that

$$\frac{\delta(\Gamma_{\text{SL}}(b \rightarrow u))_{\text{WA}}}{\Gamma_{\text{SL}}(b \rightarrow u)} \approx 1\%, \quad (3.67)$$

²⁷Perhaps rather obviously, “non-factorizable” labels a term that is non-vanishing only in the case that factorization is violated.

²⁸Some recent analyses attribute all of the “missing” 50% of the D semileptonic width to contributions from non-factorizable terms [65], but it is by no means conclusively decided that violations of factorization are the only explanation for either the problem of the D semileptonic width or the lifetime difference $\tau_{D_s} - \tau_D$. Other possibilities include flavor SU(3) symmetry-breaking, the failure or inaccuracy of the OPE at scales as low as the charm mass, or some version of duality violations.

which only sets the scale for the significance of the potential contributions; he allows for a factor of two increase in the effects. Others [72, 73] have translated a 10% violation of factorization into a possible 2–3% enhancement for the total semileptonic width $\Gamma_{\text{SL}}(b \rightarrow u)$. For an endpoint cut that retains only 10% of the total $b \rightarrow u \ell \nu$ rate, the contribution from weak annihilation can thus be magnified to a 20–30% correction to the partial rate. For a q^2 cut at the critical value of $(M_B - M_D)^2$, the correction could still be $\sim 10\%$, increasing in significance as the q^2 cut is raised.

Bigi *et al.* [62] have explored the role of helicity suppression in the decay rate difference between charged and neutral B mesons, still considered one of the best experimental avenues for constraining the contribution from weak annihilation. They find that the width difference due to the term in Eqn 3.63 can be parameterized as

$$\Delta\Gamma_{\text{SL}} \equiv \Gamma_{\text{SL}}(B^-) - \Gamma_{\text{SL}}(B^0) \simeq \frac{G_F^2 |V_{ub}|^2 f_B^2 M_B^3}{8\pi} \left(1 - \frac{m_\ell^2}{M_B^2}\right) \left(v \frac{m_\ell^2}{M_B^2} + 2g\right) \quad (3.68)$$

The first term in the final factor (proportional to v) vanishes in the limit $m_\ell \rightarrow 0$, and reflects the conventional chirality suppression of weak annihilation as it applies to semileptonic decays. On the other hand, it is clear that even in the limit of massless leptons, the second term still survives and the size of the contribution depends on the parameter g . This term captures the non-factorizable contributions of the underlying matrix element. Thus, while the decays $B \rightarrow \ell \nu$ are indeed subject to helicity suppression, the reduction is subtly circumvented when hadrons are present in the final state. Choosing an arbitrary value of $g \sim 1/3$, Bigi finds that the contribution from weak annihilation could be as much as six times the rate for $B \rightarrow \tau \nu$. Current estimates of the branching fraction for this purely leptonic decay are in the neighborhood of 5×10^{-5} , suggesting that weak annihilation could be as large as a 15% effect, relative to a nominal rate of $\mathcal{B}(b \rightarrow u \ell \nu) = 1.75 \times 10^{-3}$.

There is speculation [66] that the non-factorizable contributions from the dimension-6 operator highlighted in Eqn 3.63 do not vanish for neutral B mesons. Whether these contributions should be interpreted physically as manifestations of “weak annihilation” is a matter of definition (or opinion).

We note that lattice QCD also has the potential to shed light on these effects.

3.4.3 Sub-leading Corrections to the Shape Function

From a more distant perspective, almost all of the complications in the analysis of $b \rightarrow u \ell \nu$ arise from problems with approximating an infinite (asymptotic) series by only its first few terms.²⁹ Duality violations can be traced to the neglect of

²⁹Most of the remaining trouble is caused by our ignorance about the few terms that are retained in the truncated series.

terms in Euclidean space that become relevant in the analytic continuation back to Minkowski space; weak annihilation contributions arise from a neglected higher-order term in the OPE of HQET in Λ/m_b . Similarly, in the analysis of the endpoint region we encountered a breakdown of the usual OPE because an infinite number of terms were formally all $\mathcal{O}(1)$; the resummation of the leading singular terms led to the shape or light-cone momentum distribution function $f(k_+)$. The natural next step in understanding the OPE for $b \rightarrow u \ell \nu$ is the investigation now of the *sub-leading* terms in the shape function region. Unfortunately, much less is known about the new terms that arise at next-to-leading order.

We present a lightning review of the appearance of the shape function in the now-standard revision of the OPE for the endpoint region to illustrate where and how new sub-leading contributions to the leading-order result can appear. Recall that hadronic states in the endpoint region are kinematically constrained to have high energy and low invariant mass:³⁰

$$E_X \sim m_b, M_X^2 \sim \Lambda m_b \gg \Lambda^2, \quad (3.69)$$

describing a region where the OPE approach is still valid³¹ but the expansion parameter must be modified. The most singular terms in the modified expansion may be resummed into a non-local operator³²

$$O_0(\omega) = \bar{h}_v \delta(\omega + in \cdot \hat{D}) h_v \quad (3.70)$$

where n^μ is a light-like vector in the direction of the final hadrons, h_v is the HQET heavy quark field introduced previously, and $\hat{D}^\mu \equiv D^\mu/m_b$ is the reduced gauge-covariant derivative. The matrix element of this operator in a B meson defines the light-cone structure function for the meson:

$$f(\omega) = \frac{1}{2M_B} \langle B | O_0(\omega) | B \rangle. \quad (3.71)$$

The rate in the endpoint region is determined by $f(\omega)$,

$$\frac{d\Gamma}{dE_\ell} = \frac{G_F^2 |V_{ub}|^2 m_b^4}{96\pi^3} \int d\omega \Theta(m_b - 2E_\ell - \omega) f(\omega), \quad (3.72)$$

but since the shape function is fundamentally non-perturbative and cannot be determined analytically, the rate in this region is intrinsically model-dependent even at leading order in Λ/m_b . (We've already discussed various efforts to reduce

³⁰Observe that the high-energy, low-mass four-vectors are by definition light-like. Hence the alternative nomenclature describing the shape function as the “light-cone distribution function.”

³¹Note that the OPE completely breaks down in the resonance regime $M_X^2 \sim \Lambda^2$ and becomes basically unsalvageable.

³²We switch notation here and adopt the reduced (dimensionless) variable $\omega \equiv k_+/m_b$.

this basic ignorance with constraints on the moments or input from the photon spectrum in $b \rightarrow s\gamma$.)

Theoretical considerations have now advanced to the next terms in the non-perturbative expansion, suppressed by higher powers of Λ/m_b . Since these so-called “sub-leading twist corrections” are different between $b \rightarrow u\ell\nu$ and $b \rightarrow s\gamma$, they destroy the universality of the shape function across B decays. Hence the remarkable relation in Eqn 3.53 must, for instance, be modified to account for the differing corrections to each process. Recent analyses [67–69] identify five new non-local operators that emerge at the next order in the twist expansion, which reduce to four (unknown) structure functions for B meson decays. Bauer *et al.* evaluate the effects of these new terms on an endpoint analysis, using a standard ansatz for the structure function. They find the corrections to the partial rate are as large as 15% with a lepton energy cut $E > 2.2$ GeV, and conclude that since their analysis is strongly model-dependent, the impact for $|V_{ub}|$ could be equally as large. Leibovich *et al.* have carried out a similar analysis with comparable results.

Neubert [70] has shown that the anomalously large corrections reported by Bauer *et al.* are dominated by the first moment of the sub-leading shape function, which is known in terms of the hadronic parameter λ_2 , making the correction better known than initially believed. Further, while the power correction to the shape function does lead to a $\sim 25\%$ enhancement in the partial rate (a 13% impact for $|V_{ub}|$), it is under good control, since the correction can be absorbed into a redefinition of the leading-order shape function. Residual corrections from sub-leading shape function effects estimated across a range of models are only in the few percent range, so there is no significant loss of predictive power with the neglect of these additional terms. With the first few moments of the structure function known, the correction is largely in hand; hence the sub-leading power corrections do not pose a fundamental limitation on the extraction of $|V_{ub}|$.

More recently, Burrell, Luke, and Williamson [71] have explored the impact of sub-leading power corrections on the hadronic mass spectrum and come to similar conclusions when the extraction of $|V_{ub}|$ is augmented with input from $b \rightarrow s\gamma$. They find that the theoretical uncertainty in $|V_{ub}|$ due to higher-order shape function effects is at the few percent level, substantially less than the corresponding corrections to an endpoint analysis, and sub-dominant when compared to other sources of theoretical and experimental error.

We have now completed our tour of the extensive and mature theoretical program to extract $|V_{ub}|$ from inclusive $b \rightarrow u\ell\nu$ decays.

3.5 Summary and Outlook

Several recent proposals in the phenomenological literature respond to the challenge of extracting $|V_{ub}|$ from realistically obtainable data. No method is particularly stunning, nor fully immune to experimental and theoretical complications. In partial consequence, there is good reason to prefer the pursuit of multiple, competing approaches rather than relying on a single strategy with its array of associated uncertainties. In the end, each strategy promises to deliver a determination of $|V_{ub}|$ with a theoretical error of order 10% or better, but consensus within the theoretical community on quantitative versions of these estimates has grown only slowly. The new precision arises largely from the use of the model-independent framework of HQET and a judicious choice of experimental cuts, but every case displays a subtle interplay between what is theoretically calculable and what is experimentally feasible.

One of the essential tools in the analysis of $b \rightarrow u \ell \nu$ has been the heavy quark expansion, which connects inclusive calculations to the underlying theory of QCD. Inherent in this OPE-based approach, however, are three unavoidable sources of theoretical uncertainty:

1. Unknown (perturbative) terms of higher order in α_s ,
2. Unknown (non-perturbative) terms of higher order in $1/m_b$,
3. Uncertainties in the input parameters α_s , m_b , and the expectation values of various local operators in the OPE.

This list is essentially exhaustive, since it recognizes the parameterization of our ignorance in all parts of the OPE approach. For instance, weak annihilation and sub-leading corrections to the shape function fall into the second category. Violations of duality are perhaps best regarded as additional uncertainties arising from the theory at a lower level, entangled with issues of the adequacy of the OPE program, but they can also be viewed as the failure of an asymptotic expansion of sorts, where corrections of unknown size arise from neglected terms.

Reducing the errors associated with each entry in this list remains the basic theoretical mandate for phenomenological study of $b \rightarrow u \ell \nu$ and progress on $|V_{ub}|$. In a recent presentation, Ligeti [73] outlines a possible program for a response. Among his suggestions are:

- Evaluate the full $\mathcal{O}(\alpha_s^2)$ perturbative corrections to all relevant kinematic distributions (they are currently only known for the total rate and q^2 spectrum).
- Pursue a precision determination of m_b . All rate calculations include a factor of m_b^5 , and sensitivity to m_b increases in the presence of cuts.

- Improve the measurement of the photon energy spectrum in $b \rightarrow s\gamma$, and employ the spectrum directly in E_ℓ and M_X^2 analyses, rather than using it to constrain some intermediate parameterization of the shape function.
- Bring the experimental cuts as close to charm threshold as feasible to include as much $b \rightarrow u\ell\nu$ phase space as possible.
- Constrain the contribution from weak annihilation.

The first few items are specific challenges on the theory front; the last ones are directives for experimental work. Most of these tasks, however, require input and expertise from both arenas, continuing the long tradition of the intertwining of theory and experiment in $b \rightarrow u\ell\nu$.

New analyses at the B factories are implementing advanced experimental techniques to: reduce the contamination from charm backgrounds (for instance, by fully reconstructing the other B decay in the event); improve the efficiency and purity of identifying $B \rightarrow X_u\ell\nu$; and reduce the systematic errors on measurements of both $b \rightarrow u\ell\nu$ and $b \rightarrow s\gamma$. A measurement of $|V_{ub}|$ with experimental errors comparable to the tantalizing promise of a 10% theory error is not far away.

Estimates of the remaining unknown theory errors have already been made. In a recent review [2], Gibbons has attempted to use existing experimental data to derive constraints on the theoretical uncertainties that are not yet fully quantified. By identifying how the unknown corrections to the partial rate vary across phase space, he is able to constrain, for the first time, the uncertainties associated with the three areas discussed in detail above: local quark-hadron duality, weak annihilation, and sub-leading corrections to the shape function. The assembled experimental data (2003) include lepton endpoint [128–130], M_X^2 -only [131, 132], and q^2 - M_X^2 [133] analyses at various of the B experiments—CLEO, BaBar, and Belle. His efforts result in the combined inclusive result:

$$|V_{ub}| = (4.63 \pm 0.28_{\text{stat}} \pm 0.39_{\text{syst}} \pm 0.48_{f_{qM}} \pm 0.32_{\Gamma_{\text{thy}}} \pm 0.11_{\text{QHD}} \pm 0.27_{\text{WA}} \pm 0.31_{\text{SSF}}) \times 10^{-3} \quad (3.73)$$

for a total theory error of 15% and an overall precision of 18%. (The Γ_{thy} term is the contribution to the error from uncertainty on the semileptonic width; f_{qM} , the error in the fraction of the rate above Belle’s combined q^2 - M_X^2 cut; QHD, from violations of local quark-hadron duality; WA, contributions from weak annihilation; and SSF, uncertainty in sub-leading corrections to the shape function.)

The relevance of the work described in this thesis should now be clear. By analyzing semileptonic B decays at CLEO, we attempt to independently bound the contribution of weak annihilation to the analysis of $b \rightarrow u\ell\nu$, and so address directly one of the critical tasks required for a precision measurement of $|V_{ub}|$. It amounts to a refinement of the same kind of data-driven bound Gibbons was able

to achieve by reviewing existing analyses, but by performing the search for weak annihilation directly in data, the constraints are both more informative and more powerful.

In summary, by constraining the contribution of weak annihilation effects to traditional $b \rightarrow u \ell \nu$ decays, we help whittle away at one of many theoretical uncertainties that still limit our knowledge of the CKM element $|V_{ub}|$.