On various aspects of Stein’s method:
Quantitative approximation for stochastic limit theorems

Arne Gouwy
Promotor: Prof. C. Ley
Department of Applied Mathematics, Computer Science and Statistics
Faculty of Sciences
Universiteit Gent

Master thesis submitted to acquire the academic degree of Master of Science in mathematics.

Academic year 2015-2016
On various aspects of Stein’s method:
Quantitative approximation for stochastic limit theorems

Arne Gouwy
Promotor: Prof. C. Ley
Department of Applied Mathematics, Computer Science and Statistics
Faculty of Sciences
Universiteit Gent

Master thesis submitted to acquire the academic degree of Master of Science in mathematics.

Academic year 2015-2016
Acknowledgements

First of all, I would like to thank Prof. C. Ley to help supervise this thesis. I would also like to thank my family and friends for their support and help in proof-reading.

The author gives permission to make this master thesis available for consultation and copy parts of it for personal use. Any other use is subjected to copyright restrictions, in particular to the compulsory citation of the source when results of this master thesis are quoted.

29 May 2016,
Arne Gouwy
Contents

1 Introduction to Stein’s method 1
   1.1 Normal approximation in one dimension .............................. 1
      1.1.1 Dependency .............................................. 7
   1.2 Poisson approximation ............................................. 10
   1.3 Zolotarev type distances and optimal transport ..................... 14
      1.3.1 Relations between the metrics .......................... 16
      1.3.2 Transportation metrics and optimal couplings .............. 18
      1.3.3 Explicit computation on the line ....................... 21
   1.4 A first overview ............................................... 23

2 The one-dimensional density view 26
   2.1 Dual operators ................................................. 26
      2.1.1 Generalities ............................................. 26
      2.1.2 Derivative operators ..................................... 28
   2.2 Lebesgue densities: Stein approximation ........................ 29
   2.3 Discrete distributions ......................................... 40
   2.4 Zero-bias coupling and Lindeberg’s CLT .......................... 44
   2.5 Comparing nested densities .................................... 47

3 Stein’s method in statistics 50
   3.1 Comparing priors in Bayesian statistics ......................... 50
      3.1.1 Normal model ............................................ 51
      3.1.2 Binomial model ......................................... 52
   3.2 Maximal likelihood estimators ................................... 55

4 Gaussian analysis: the language of Malliavin calculus 60
   4.1 Gaussians: motivation and encoding scheme ...................... 60
   4.2 Derivative operators and duals ................................ 63
      4.2.1 Derivative operator ...................................... 64
      4.2.2 Derivatives for general Hilbert space-valued random variables 69
      4.2.3 The divergence operator ................................ 70
      4.2.4 Divergence of Hilbert space-valued variables ............ 73
   4.3 The generator and Ornstein-Uhlenbeck process .................... 74
Introduction

When making an abstraction of some world phenomena based on observations, i.e. modelling, there are many influences to be considered. When a basic model is obtained, there may both be subeffects or noise that are hard to describe. On top of that, the scientist’s methods may only be limited to some idealistic, smooth, tractable modelling tools. Improved tools easily become too difficult to work with or even derive. That is why a central property that a model should have is robustness. This means that small deviations on the assumptions of the model should only give rise to small differences in the inferences, i.e. it should be continuous or have some other smoothness properties. Models based on normal distributions became very popular due to robustness and tractability for example. Then a central question becomes: how can these differences be investigated? This is where stochastic approximation comes in, in the case of probabilistic models. Early methods tried to directly compare probabilities such as

\[ \mathbb{P}[X \leq x] - \mathbb{P}[Y \leq x], \]

where \( X \) and \( Y \) are two random variables on a probability space \( (\Omega, \mathcal{F}, \mathbb{P}) \) and \( x \) runs through an adequate subset of \( \mathbb{R} \).

Exploiting a duality structure, distributions can be compared by:

\[ \mathbb{E}[h(X)] - \mathbb{E}[h(Y)] := \int h d\mathbb{P}_X - \int h d\mathbb{P}_Y, \quad h \in C_b(S). \]

Here, \( X \) and \( Y \) are \( S \)-valued random variables for a topological space \( S \) and \( C_b(S) \) denotes the bounded continuous functions on \( S \). \( \mathbb{P}_X := \mathbb{P} \circ X^{-1} \) denotes the distribution of \( X \). When

\[ \int h d\mathbb{P}_X \approx \int h d\mathbb{P}_Y \]

for a wide enough, yet regular class of measurable \( h \), the distributions of \( X \) and \( Y \) are expected to be close to each other.

How is it possible to actually compare these quantities? Some very successful and well-known methods rely on transformation of the distributions. Generating functions or the Fourier transform became central tools in probability and statistics leading to results such as Lévy’s continuity theorem. Furthermore, when multiple random variables are considered, they are successful in exploiting independence.

However, adapting them for other probability structures, where dependence can not be ignored for example, may take much added effort. Also, it would be convenient to consider other classes of \( h \) and to have quantitative estimates for (2). When considering sequences of random variables namely, it may be of interest to have some idea how fast they approach a certain distribution.

In [25] and [26], Charles Stein developed another method to estimate (2) that allows to investigate slight dependence, generalizes to other classes of \( h \) and offers a
quantitative view. To approximate the normal distribution he exploited the structure of a partial differential equation. Later, his PhD student Louis Chen was able to extend the framework to Poisson approximation. It is remarkable to note the ability of the method to generalize to many probabilistic structures. In this way it even allows to gain more insight in approximation problems, allowing strong results in Gaussian analysis such as the Breuer Major theorem [18].

The main goal of this text is to investigate what unifies Stein’s method in all its appearances in the literature. What makes it work? When is it interesting to use? And how does it translate to other settings? In the first three chapters, we detail the setting of a one-dimensional random variable having a density with respect to a canonical measure. We investigate applications such as central limit theorems and the comparison of Bayesian priors, on which some slight extensions are investigated. In doing this, we try to promote unifying language. For Gaussian analysis\(^1\), a successful symbiosis of Stein’s method with Malliavin calculus now offers a well-worked out framework. That is why we make a considerable effort in analyzing it in chapter 4. In this way, we crystallize the crucial mechanisms in chapter 6. Also, some insight and training into the arguments are needed to master the language. We give a sketch of some interesting approximation results that it leads to in chapter 5.

The appendices serve to establish conventions (B.1) that are used throughout the text, offer some supporting results (B.2 and B.3) and give some additional argumentation (C). It is advisable to take a look at the conventions first.

Now, the method may best be tasted by just diving into it.

\(^1\)The study of functionals of Gaussian processes, diffusion processes for example.
Chapter 1

Introduction to Stein’s method

For random variables $X, Z$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with values in a Polish space\(^1\) $S$, we are interested in comparing their distributions by:

$$
\mathbb{E}[h(X)] - \mathbb{E}[h(Z)] = \int h d\mathbb{P}_X - \int Z d\mathbb{P}_Y = \langle \mathbb{P}_X - \mathbb{P}_Z, h \rangle.
$$

(1.1)

Here, $h$ runs through some function space $\mathcal{C}$. The class should be sufficiently large to determine the behavior of the distributions. On the other hand, it should still be sufficiently regular to efficiently work with (1.1). It is often difficult to evaluate or bound the formula because either the distributions are not both well-known or $\mathcal{C}$ is not suitable for this task. As an example, only the normal distribution is well-known in the central limit theorem. Yet, some useful things on probabilistic approximations may still be inferred.

It will first be investigated how Stein’s method works for normal approximation, meaning that one of the variables is normally distributed. It will then be compared to a similar program for Poisson approximation. Many similarities occur and it is a goal of this text to describe the successful mechanisms. We take the liberty of introducing some language that is more readily generalizable. In the beginning, it makes some arguments more involved, but it will smoothen the transition to more abstraction. Also, it will enable us to state what we think the key notions of Stein’s method are and how they may be generalized. Before we start, the reader is advised to go through the first section of Appendix B for a list of conventions.

1.1 Normal approximation in one dimension

To estimate (1.1), it may be investigated how an operator $D$ acts on $\mathcal{C}$ under the distribution of $Z$. Stein’s method uses differentiation. Consider a standard normal variable $N \sim N(0,1)$ and an absolutely continuous function $g : \mathbb{R} \rightarrow \mathbb{R}$ such that

\(^1\)See definitions B.1.1 and B.1.2 for some conventions.
$z \mapsto g'(z)e^{-z^2/2}$ is integrable on $\mathbb{R}$. Note that we thus make use of the notion of weak derivatives\(^2\).

\[
\mathbb{E}[g'(N)] = \int_{-\infty}^{\infty} g'(z) \frac{e^{-z^2/2}}{\sqrt{2\pi}} dz
\]

\[
= \int_{-\infty}^{0} g'(z)dz \int_{-\infty}^{z} (-y) \frac{e^{-y^2/2}}{\sqrt{2\pi}} dy + \int_{0}^{\infty} g'(z)dz \int_{z}^{\infty} y \frac{e^{-y^2/2}}{\sqrt{2\pi}} dy
\]

\[
= \int_{-\infty}^{0} (-y) \frac{e^{-y^2/2}}{\sqrt{2\pi}} dy \int_{0}^{y} g'(z)dz + \int_{0}^{\infty} y \frac{e^{-y^2/2}}{\sqrt{2\pi}} dy \int_{0}^{y} g'(z)dz
\]

\[
= \mathbb{E}[Ng(N)].
\]

Here the weak condition $g' \in L^1(N)$ is sufficient to compute the dual of differentiation. We have used this condition to use Fubini’s theorem in (1.2): on $(-\infty, 0)$ $(e^{-z^2/2})'$ is positive and it is negative on $(0, \infty)$, so that the function appearing in the double integral is absolutely integrable.

If $X$ is an arbitrary random variable such that for all absolutely continuous $g \in G^{1,1}(N)$ (definition B.1.5) the expectations $\mathbb{E}[g'(X)]$ and $\mathbb{E}[Xg(X)]$ are finite and

\[
\mathbb{E}[g'(X)] = \mathbb{E}[Xg(X)],
\]

then this determines the moments incrementally. Taking $g(x) = 1$ yields $\mathbb{E}[X] = 0$, $g(x) = x$ yields $\mathbb{E}[X^2] = 1$. Going on, we find $\mathbb{E}[X^{k+1}] = k\mathbb{E}[X^k]$. The moments thus coincide with those of a standard normal. It can be shown that the moments determine the distribution uniquely\(^3\), so $X \sim N(0, 1)$. We thus have $X \sim N(0, 1)$ if and only if the functional $g \mapsto \mathbb{E}[g'(X) - Xg(X)]$ vanishes on a wide enough class of functions. This yields a so-called **Stein characterization**. We will give a Stein characterization based on more general conditions in Theorem 1.1.5.

Now consider the linear transformation for $p \in [1, \infty]$:

\[
\delta_N : \text{dom}(\delta_N) \subset L^p(N) \to L^p_\perp(N); g \mapsto g'(x) - xg(x).
\]

Here, $L^p_\perp(N)$ denotes the $h \in L^p(N)$ such that $\mathbb{E}[h(N)] = 0$. Thus, in $L^2(N)$, $L^2_\perp(N)$ is orthogonal to the constants. For the moment, we leave the domain $\text{dom}(\delta_N)$ unspecified. It is sufficient to note that it encompasses $C^1_b(\mathbb{R})$. On this domain, we have thus proven that $\delta_N$ is dual to differentiation with respect to $L^2(N)$. If it is possible to invert $\delta_N$ on $C_\perp := \{h - \mathbb{E}[h(N)] \mid h \in C\}$, and we write $g_h = \delta_N^{-1}(h - \mathbb{E}[h(N)])$:

\[
\mathbb{E}[h(X)] - \mathbb{E}[h(N)] = \mathbb{E}[(h(X) - \mathbb{E}[h(N)])] = \mathbb{E}[\delta_N(g_h)(X)].
\]

---

\(^2\)Definition B.1.5.

\(^3\)This follows from $\mathbb{E}[e^{tN}] < \infty$ for all $t \in \mathbb{R}$ and a series expansion.
The properties of $\delta_N$ only depend on $N$. In this way, the problem of estimating (1.1) may be divided into 2 stages. First, regularity may be inferred for $\delta_N^{-1}$. Then, the problem is retranslated into estimating $g_h \mapsto \mathbb{E}[\delta_N(g_h)(X)]$. We investigate the differential equation:

$$g'(x) - xg(x) = h(x) - \mathbb{E}[h(N)] =: h_\perp(x), \quad \text{for all } x \in \mathbb{R}, \quad (1.7)$$

where $h$ is sufficiently regular and $g$ is solved in the sense of weak derivatives. It is termed the Stein equation for the normal distribution. The reader acquainted with some theory of partial differential equations (PDE) may suspect that inversion on $L_\perp^p(N)$ in terms of Green kernels or potential operators can have nice regularity properties. If the identity were solved, we know by the product rule of absolutely continuous functions, Lemma B.3.2:

$$(e^{-x^2/2}g)'(x) = -xe^{-x^2/2}g(x) + e^{-x^2/2}g'(x) = e^{-x^2/2}h_\perp(x),$$

from which it may be inferred that

$$g(x) = e^{x^2/2} \int_{-\infty}^{x} e^{-y^2/2} (h(y) - \mathbb{E}[h(N)]) \, dy + ce^{x^2/2}, \quad c \in \mathbb{R}. \quad$$

The reader may check that it indeed solves (1.7). Only $c = 0$ yields a solution in $L^p(N)$ for regular $h$ as will follow from Lemmas 1.1.1 and 1.1.3. Otherwise the solution blows up for $x \to \pm \infty$. We define Stein solutions:

$$\delta_N^{-1}(h_\perp)(x) := e^{x^2/2} \int_{-\infty}^{x} e^{-y^2/2} (h(y) - \mathbb{E}[h(N)]) \, dy = \int_{\mathbb{R}} K(x, y)h_\perp(y) \, dz \quad (1.8)$$

Thus, as is usual in PDE theory, the inverse is a kernel operator, with kernel defined by $K(x, y) = e^{x^2/2 - y^2/2} \mathbb{1}_{y \leq x}$. Only on $L_\perp^p(N)$ it is expected to be regular since $\delta_N^{-1}(1)$ diverges for $x \to +\infty$. In contrast, since $h_\perp$ integrates to 0 with respect to $N$:

$$\delta_N^{-1}(h_\perp)(x) = -e^{x^2/2} \int_{x}^{\infty} e^{-y^2/2} (h(y) - \mathbb{E}[h(N)]) \, dy. \quad$$

Note that the apparent loss of regularity in (1.6) due to $\delta_N$ is compensated by integration in $\delta_N^{-1}$. We expect to gain “1 degree of regularity”. To obtain properties on $g_h$ independent of $X$, a Stein approximation is performed. For example:

**Lemma 1.1.1 (W$^{1,\infty}$-bounds for $\delta_N^{-1}$).** If $h \in L^\infty(N)$, then $\delta_N^{-1}(h_\perp)$ is absolutely continuous with:

$$|\delta_N^{-1}(h_\perp)(x)| \leq \|h_\perp\|_\infty \int_{0}^{\infty} e^{-|x|y} e^{-y^2/2} \, dy \leq \sqrt{\frac{\pi}{2}} \|h_\perp\|_\infty, \quad (1.9)$$

$$|\delta_N^{-1}(h_\perp)'(x)| \leq 2\|h_\perp\|_\infty. \quad (1.10)$$
Remark 1.1.2. The extra degree of regularity corresponds to the continuity of
\[ \delta_N^{-1} : L_1^\infty(N) \rightarrow W^{1,\infty}(N). \] (1.11)

Proof. A substitution yields:
\[
\delta_N^{-1}(h_\perp)(x) = \begin{cases} 
  e^{x^2/2} \int_0^\infty e^{-x^2/2} e^{-y^2/2} h_\perp(x - y)dy & \text{if } x \leq 0 \\
  e^{x^2/2} \int_0^\infty e^{-x^2/2} e^{-y^2/2} h_\perp(x + y)dy, & \text{if } x > 0 
\end{cases}
\]
\[ = \int_0^\infty e^{-|x|y} e^{-y^2/2} h_\perp(x + \text{sgn}(x)y)dy. \]

Note the extra damping factor \( e^{-|x|y} \) and push towards tails \( x + \text{sgn}(x)y \). This yields expression (1.9) after noting that \( e^{-|x|y} \leq 1 \). (1.7) then gives:
\[ |\delta_N^{-1}(h_\perp)'(x)| \leq \|h_\perp\|_\infty \int_0^\infty |x|e^{-|x|y}e^{-y^2/2}dy + \|h_\perp\|_\infty. \]

Using \( e^{-y^2/2} \leq 1 \) and computing the integral finally yields (1.9).

For absolutely continuous functions with bounded derivative we get:

<table>
<thead>
<tr>
<th>Lemma 1.1.3 ( W^{2,\infty})-bounds for ( \delta_N^{-1} ). If ( h ) is absolutely continuous with ( h' \in L^\infty(N) ), then:</th>
</tr>
</thead>
<tbody>
<tr>
<td>[</td>
</tr>
<tr>
<td>[</td>
</tr>
<tr>
<td>[</td>
</tr>
</tbody>
</table>

This will be proven in section 2.2 by direct computations. In chapter 4, we will see representations for the Green function that are more apt to showing regularity.

Remark 1.1.4. Now the extra degree of regularity corresponds to the continuity of
\[ \delta_N^{-1} : C^{1,\infty}_1(N) \rightarrow W^{2,\infty}(N). \] (1.15)

The subscript \( \perp \) again indicates that the expectation with respect to \( N \) is zero.

This enables us to give a more general form for the Stein characterization.

Theorem 1.1.5 (Stein characterization for a standard normal). Let \( X \) be a real-valued random variable. Then \( X \) has a standard normal distribution if and only if \( X \) has a first moment and for all \( g \in W^{1,\infty}(N) \) it holds that
\[ \mathbb{E}[g'(X) - Ng(X)] = 0. \]

In this case it remains valid for all \( g \) such that \( g' \in L^1(N) \).
Note that $p = \infty$ makes the function class independent from the exact distribution of $N$.

Proof. For $t \in \mathbb{R}$, we may solve
\[ g'(x) - xg_t(x) = \mathbb{1}_{(-\infty,t]}(x) - \Phi(t), \]
where $\Phi(t) = \int_{-\infty}^{t} \frac{e^{-y^2/2}}{\sqrt{2\pi}} dy$. By Lemma 1.1.1, $g_t$ is bounded with bounded derivative. Thus:
\[ \mathbb{E}[\mathbb{1}_{(-\infty,t]}(X) - \Phi(t)] = \mathbb{E}[g'_t(X) - Ng_t(X)] = 0, \]
or $\mathbb{P}[X \leq t] = \Phi(t)$. This yields the converse direction. The direct implication was proven in (1.3).

Denote $g_h := \delta_N^{-1}(h)$. This characterization is a necessary condition to eliminate (1.1) because $\mathcal{C}$ is measure-determining and:
\[ \mathbb{E}[g_h(X)] = \mathbb{E}[h(X) - \mathbb{E}[h(N)]]. \]
However, it does not yet constitute a method to compare the distributions. For this we combine (1.6) with the added regularity.
\[ |\mathbb{E}[h(X)] - \mathbb{E}[h(N)]| = |\mathbb{E}[g'_h(X) - Xg_h(X)]|. \tag{1.16} \]
When considering weak or distributional convergence of measures, $\mathcal{C}$ may consist of bounded Lipschitz $h$ with Lipschitz constant $L$ by the Portmanteau property, Lemma C.1.2. By Rademacher’s theorem, Theorem B.3.3, it is known that $h$ is absolutely continuous with $||h'||_\infty \leq L$. Thus $g'_h$ and $g''_h$ exist and are bounded even. Then, we do not need to take into account the exact structure of $\mathcal{C}$ anymore, but just consider the larger set $\mathcal{G} \supset \delta_N^{-1}(\mathcal{C})$ of $g \in W^{2,\infty}(N)$ with the bounds of Lemma 1.1.3. We speak of a Stein approximation. Then, the Stein functional needs to be estimated:
\[ S_X : \mathcal{G} \to \mathbb{R}; g \mapsto \mathbb{E}[\delta_N(g)(X)]. \tag{1.17} \]
A central limit theorem may readily be inferred now.

**Lemma 1.1.6.** Let $(X_i)_{i=1,\ldots,n}$ be independent centered random variables with finite third moments. Let $\mathbb{E}[X_i^2] = 1$ and $W = n^{-1/2} \sum_{i=1}^{n} X_i$. Then:
\[ |\mathbb{E}[g'(W) - Wg(W)]| \leq ||g''||_\infty \left( \sum_{i=1}^{n} \frac{\mathbb{E}|X_i|}{n^{3/2}} + \sum_{i=1}^{n} \frac{\mathbb{E}|X_i|^3}{2n^{5/2}} \right), \tag{1.18} \]
\[ \leq ||g''||_\infty \left( \frac{1}{n^{1/2}} + \frac{1}{2n^{3/2}} \sum_{i=1}^{n} \mathbb{E}|X_i|^3 \right). \tag{1.19} \]
Proof. Denote \( W_i := n^{-1/2} \sum_{j \neq i} X_i \) and consider Taylor expansions:

\[
g(W) = g(W_i) + g'(W_i) \frac{X_i}{\sqrt{n}} + \int_{W_i}^{W} g''(t)(W - t)dt.
\]

Note that the last term corresponds to \( g''(\bar{W}_i) \frac{X_i^2}{2n} \) when \( g'' \) is strongly differentiable, where \( \bar{W}_i \) lies in between \( W \) and \( W_i \). The reason for an expansion w.r.t. \( W_i \) is that the independence allows to compare derivatives conveniently:

\[
\mathbb{E}[X_i g(W)] = \mathbb{E}[X_i] \mathbb{E}[g(W_i)] + \frac{1}{\sqrt{n}} \mathbb{E}[X_i^2] \mathbb{E}[g'(W_i)] + \mathbb{E}[X_i F_2(W_i, W)].
\]

We have used the independence to split the expectations. Then this yields the result:

\[
|\mathbb{E}[g'(W) - W g(W)]| = \left| \sum_{i=1}^{n} \frac{1}{n} \mathbb{E}[g'(W) - g'(W_i)] + \sum_{i=1}^{n} \frac{1}{n^{1/2}} \mathbb{E}[X_i F_2(W_i, W)] \right|
\leq \sum_{i=1}^{n} \frac{1}{n} \left| \mathbb{E}[g'(W) - g'(W_i)] \right| + \sum_{i=1}^{n} \frac{\|g''\|_{\infty}}{2n^{3/2}} \mathbb{E}|X_i|^3.
\]

Now apply the Cauchy-Schwarz inequality to obtain (1.19): \( \mathbb{E}|X_i| \leq \mathbb{E}|X_i^2|^{1/2} = 1 \). □

**Theorem 1.1.7** (Central limit theorem). For independent and identically distributed random variables \((X_i)_{i \in \mathbb{N}}\) with \( \mathbb{E}[X_i] = 0 \), \( \mathbb{E}[X_i^2] = 1 \) and finite third moment, \( W^{(n)} := n^{-1/2} \sum_{i=1}^{n} X_i \) converges weakly to a standard normal variable \( N \), for \( n \to \infty \).

**Proof.** Noting the Portemanteau theorem, Lemma C.1.2, it is sufficient that (1.1) converges to 0 for all Lipschitz functions \( h \). This is the case by Lemmas 1.1.6 and 1.1.3. □

**Remark 1.1.8** (General normal distribution). By rescaling, a central limit theorem is obtained for normal random variables with general mean \( \mathbb{E}[X_i] = \mu \) and \( \mathbb{E}[X_i^2] = \sigma^2 \). Actually, all previous results carry over to general Gaussians, such as a Stein characterization for \( Z \sim N(\mu, \sigma^2) \):

\[
X \sim N(\mu, \sigma^2) \text{ if and only if } X \text{ has a first moment and for all } g \in W^{1, \infty}(Z):
\]

\[
\mathbb{E}[\sigma^2 g'(X) + (\mu - X)g(X)] = 0. \tag{1.20}
\]

The Stein solutions are given by, \( x \in \mathbb{R} \):

\[
g_h(x) := e^{\frac{(x-\mu)^2}{2\sigma^2}} \int_{-\infty}^{x} e^{-\frac{(y-\mu)^2}{2\sigma^2}} (h(y) - \mathbb{E}[h(Z)])dy. \tag{1.21}
\]

By rescaling, approximation results such as Lemma 1.1.3 carry over:

\[
g_h(x) = e^{\frac{x^2}{2}} \int_{-\infty}^{x} e^{-y^2/2} (h(\sigma [\mu + y]) - \mathbb{E}[h(\sigma [\mu + N])])dy, \quad \tilde{x} = \frac{x - \mu}{\sigma},
\]

\[
|g(x)| \leq \sigma \|h'\|_{\infty}, \quad |g'(x)| \leq \sqrt{\frac{2}{\pi}} \|h'\|_{\infty}^2, \quad |g''(x)| \leq 2\sigma^{-1} \|h''\|_{\infty}. \tag{1.22}
\]
Remark 1.1.9. The previous results also yield quantitative statements on the speed of convergence. They are related to the Berry-Esseen theorem. It states that for i.i.d. centered random variables $(X_i)_{i \in \mathbb{N}}$ with $\mathbb{E}[X_i^2] = \sigma^2$, we get a bound on the Kolmogorov distance (section 1.3):

$$\left| \mathbb{P}[(\sigma n)^{-1/2}(X_1 + \ldots + X_n) \leq x] - \mathbb{P}[N \leq x] \right| \leq \frac{C \mathbb{E}|X_i|^3}{\sqrt{n\sigma^3}}.$$  \hspace{1cm} (1.23)

For a constant $C$. [16] for example finds $C \leq 0.5129$. Stein’s framework may be used to derive results in this genre. See Chapter 3 of [18] for an example.

We have to comment on some technicalities that were ignored:

Remark 1.1.10 (Convention on a version of $\delta^{-1}_N$). When considering functions that are not necessarily strongly differentiable, but only weakly, the derivatives are only determined up to a set of Lebesgue measure 0. If we evaluate them in a random variable $X$ and take expectations, this poses no issue if $X$ has a density with respect to Lebesgue measure. When this is not the case, then different versions can give different values for $\mathbb{E}[g'(X)]$. When $X \equiv x$, changing $g'$ in $x$ yields another value. However, the property of $g'$ that we are interested in is

$$g'(x) = xg(x) + h(x) - \mathbb{E}[h(N)], \quad \text{for all } x \in \mathbb{R}.$$ 

Thus we only consider the representative of $g'$ that fulfills this identity for all $x \in \mathbb{R}$. We will keep this convention for the remainder of the text. Even more, whenever a function is considered that fulfills a certain equation, we always implicitly fix the version that fulfills it everywhere if this is possible. This coincides with the convention for absolutely continuous functions, where the continuous representative is considered.

After going to a broader problem with a Stein approximation, techniques still need to be investigated to bound $|S_X(g)|$. Taylor expansions worked well for sums of independent random variables. In the literature, other techniques are available, such as size bias couplings, zero bias couplings, Stein pairs and smart interpolation. For the first three, see for example [23]. Zero-bias distributions will be discussed in Chapter 2. Interpolation will be used in chapters 4 and 5.

One of the main strengths of Stein’s method is that it is amenable to more general probability structures. These may include weak forms of dependence. We illustrate this by adjusting the previous method of Taylor approximation:

1.1.1 Dependency

Consider random variables $X_1, X_2, \ldots$ on $(\Omega, \mathcal{F}, \mathbb{P})$.

Definition 1.1.11. We say that $(X_i)_{i=1,\ldots}$ have dependency neighborhoods $(N_i)_{i=1,\ldots}$ where $N_i \subseteq \{1, 2, \ldots\}$ iff $X_i$ is independent of $X_j$ for $j \notin N_i$. 

7
This can be graphically represented by a graph with as nodes the \( X_i \), and an edge connecting \( X_i \) and \( X_j \) if \( i \in N_j \setminus \{i\} \) and \( j \in N_i \setminus \{j\} \). When \( X_i \) and \( X_j \) are unconnected, they are independent. This may help in combinatorical reasoning.

**Theorem 1.1.12.** For random variables \( X_1, \ldots, X_n \) with dependency neighborhoods \( (N_i)_{i \in \{1, \ldots, n\}} \), \( \sigma_n^2 := \text{var}(\sum_i X_i) \), \( W = \sigma_n^{-1} \sum_{i=1}^n X_i \) and \( D_n := \max_i |N_i| \), the Stein functional has the following bounds:

\[
|E[g'(W) - Wg(W)]| \leq \frac{\|g''\|_\infty}{2} \frac{D_n^2}{\sigma_n^3} \sum_{i=1}^n \text{var}(X_i) + \frac{\|g''\|_\infty}{\sigma_n^2} \sqrt{\text{var}(\sum_{i=1}^n \sum_{j \in N_i} X_i X_j)} \quad (1.24)
\]

\[
\leq \frac{\|g''\|_\infty}{2} \frac{D_n^2}{\sigma_n^3} \sum_{i=1}^n \text{var}(X_i)^3 + 2\|g''\|_\infty \frac{D_n^{3/2}}{\sigma_n^2} \sqrt{\sum_{i=1}^n \text{var}(X_i^4)}. \quad (1.25)
\]

**Proof.** The proof proceeds analogous to the independent case. Denoting \( W_i = W - \sum_{j \in N_i} X_i/\sigma_n \), Taylor expansion of \( g(W_i) \) around \( g(W) \) and independence yield:

\[
\frac{1}{\sigma_n} E[X_i g(W)] = \sum_{j \in N_i} \frac{1}{\sigma_n^2} E[X_i X_j g'(W)] + \sum_{j,l \in N_i} \frac{1}{2\sigma_n^3} E[X_i X_j X_k F_2(W, W_i)].
\]

where \( F_2(W, W_i) \) is an error term bounded by \( \|g''\|_\infty \). First note that because \( E[X_i] = 0 \), we have:

\[
\sigma_n^2 = \sum_{i=1}^n \sum_{j \in N_i} E[X_i X_j].
\]

Cauchy-Schwarz then yields:

\[
\left| E[g'(W)] - \sum_{i=1}^n \sum_{j \in N_i} \frac{1}{\sigma_n^2} E[X_i X_j g'(W)] \right| \leq \frac{\|g''\|_\infty}{\sigma_n^2} \left| \sum_{i=1}^n \sum_{j \in N_i} X_i X_j \right| \leq \frac{\|g''\|_\infty}{\sigma_n^2} \sqrt{\text{var}(\sum_{i=1}^n \sum_{j \in N_i} X_i X_j)}. \quad (1.26)
\]

Using \( \text{var}(X) = E[X^2] - (E[X])^2 \):

\[
\text{var}(\sum_{i=1}^n \sum_{j \in N_i} X_i X_j) = -\sigma_n^4 + \sum_{i,k=1}^n \sum_{l \in N_k} E[X_i X_j X_k X_l]
\]

\[
= -\sigma_n^4 + \sum_{i=1}^n \sum_{j,l \in N_i} E[X_i^2 X_j X_k] + \sum_{i \neq k} \sum_{j \in N_i} \sum_{l \in N_k} E[X_i X_j X_k X_l] 
= S_1 + S_2
\]

So we seek ways to estimate \( E[X_i X_j X_k] \) and \( E[X_i X_j X_k X_l] \) carefully to avoid taking in unnecessary factor \( n \) (note that the first includes absolute values and the second
does not). We use that for positive numbers the geometric mean is bounded by the arithmetic mean. This implies for \(a_1, \ldots, a_n \geq 0\) that \(a_1 \cdots a_n \leq n^{-1}(a_1^n + \ldots + a_n^n)\).

\[
\sum_{i=1}^{n} \sum_{j,k \in N_i} \mathbb{E}[X_i X_j X_k] \leq \sum_{i=1}^{n} \sum_{j,k \in N_i} \frac{\mathbb{E}[X_i]^3 + \mathbb{E}[X_j]^3 + \mathbb{E}[X_k]^3}{3} \leq D_n^2 \sum_{i=1}^{n} \mathbb{E}[X_i]^3
\]

\[
S_1 \leq \sum_{i=1}^{n} D_n^2 \frac{\mathbb{E}[X_i^4]}{2} + \frac{1}{2} \sum_{i=1}^{n} \sum_{j \in N_i} \frac{\mathbb{E}[X_i^j] + \mathbb{E}[X_j^k]}{2} = D_n^2 \sum_{i=1}^{n} \mathbb{E}[X_i^4]
\]

\[
S_2 = \sum_{\{i,j\} \neq \{k,l\}} \mathbb{E}[X_i X_j X_k X_l] =: T_1
\]

\[
\sum_{\{i,j,k,l\}} \mathbb{E}[X_i X_j X_k X_l] =: T_2
\]

The first sum is over the subset of indices such that \(\{X_i, X_j\}\) is independent from \(\{X_k, X_l\}\) and the second over the remaining indices. Drawing a graph, we infer for \(T_2\): under a fixed \(i\), there are at most \(D_n\) possibilities for \(j\). If \(X_k\) lies in the dependency neighborhoods of \(X_i\) or \(X_j\) for given \(\{i, j\}\) there are at most \((2D_n - 1)D_n\) possibilities for \(\{k, l\}\). If it does not, then \(X_i\) does, with again at most \((2D_n - 1)D_n\) possibilities. Bounding the geometric mean by the arithmetic yields:

\[
T_2 \leq 4D_n^2 \sum_{i=1}^{n} \mathbb{E}[X_i^4].
\]

The representation for \(\sigma_n^2\) gives:

\[
|T_1| = \sigma_n^4 - \sum_{\{i,j,k,l\}} \mathbb{E}[X_i X_k X_j X_l] \leq \sigma_n^4 + \sum_{\{i,j,k,l\}} \mathbb{E}[X_i^4],
\]

where the last term is the same bound that was obtained for \(T_2\). This implies that \(\text{var}(\sum_{i=1}^{n} \sum_{j \in N_i} X_i X_j) \leq 8D_n^2 \sum_{i=1}^{n} \mathbb{E}[X_i^4]\). Putting all the bounds together gives the statement of the theorem.

\[
\square
\]

Thus, with some careful analysis in Stein’s framework, extended results about sums of random variables are readily obtained. To again obtain a central limit theorem, it is for example sufficient that \(\mathbb{E}[X_i^4] < \infty, \sigma_n \to \infty\) and \((D_n/\sigma_n)_{n \in \mathbb{N}}\) is bounded.

**Remark 1.1.13.** Because there are many variations possible to bound the Stein functional, different bounds may be found in the literature. Small differences may already need higher order moments for example. Also, a bound is often put in a nice looking form. But before reaching it, more structure of the specific problem can sometimes be exploited to improve it. As an illustration in Theorem 1.1.12, the \(\mathbb{E}[X_i X_j X_k]\) or the variance in (1.24) may sometimes be bounded more efficiently. We will be less concerned with finding the optimal conditions or constants (which are in competition
with each other), unless there is some motivation for this text. As another simple illustration, we mention that by expanding \( g(W) \) around \( W \) instead of \( W \) and the extra assumption that for all \( X_j \in N_i, N_j = N_i \) (this just coincides with the independent case by grouping sums together), one can achieve improved bounds only depending on third order moments. In this way, Lemma 1.1.6 improves on \([23]\). 

### 1.2 Poisson approximation

It will now be investigated how a similar program can be carried out for Poisson approximation. Suppose \( M_\lambda \) is Poisson distributed with parameter \( \lambda > 0 \). For \( k \in \mathbb{Z}^+ = \{0, 1, 2, \ldots\} \), it means:

\[
P_\lambda(k) := \mathbb{P}[M = k] = e^{-\lambda} \frac{\lambda^k}{k!}.
\]

(1.27)

On \( \mathbb{Z} \), we can disintegrate with respect to the counting measure \( dl \) in multiple ways. We choose the following, with \( m < k \):

\[
g(k) - g(m) = \int_{[m,k]} \triangle g(l) dl := \sum_{l=m}^{k-1} \triangle g(l) \quad (1.28)
\]

\[
= \int_{[m,k]} \nabla g(l) dl := \sum_{l=m+1}^{k} \nabla g(l),
\]

(1.29)

Where \( g : \mathbb{Z}^+ \to \mathbb{R} \), \( \triangle \) denotes the forward difference operator \( \triangle g(k) := g(k+1) - g(k) \) and \( \nabla \) the backwards difference operator \( \nabla g(k) = g(k) - g(k-1) \) if \( k \geq 1 \) and \( \nabla g(0) = 0 \) (so that \( \nabla \) vanishes on constants). In order to estimate

\[
\mathbb{E}[h(X)] - \mathbb{E}[h(M_\lambda)],
\]

a differential operator that eliminates \( \mathbb{E}[h(N)] \) regularly is sought. We try to compute the dual of \(-\nabla\) with respect to \( P_\lambda \):

\[
\sum_{l=0}^{\infty} -\nabla h(l) g(l) P_\lambda(l) = h(-1) g(0) e^{-\lambda} - \sum_{l=0}^{\infty} h(l) g(l) P_\lambda(l) + \sum_{l=0}^{\infty} h(l) g(l+1) P_\lambda(l+1)
\]

\[
= h(0) g(0) e^{-\lambda} + \sum_{l=0}^{\infty} h(l) \left(-g(l) + \frac{\lambda}{l+1} g(l+1)\right) P_\lambda(l).
\]

Note that boundary conditions arise more directly now. If we want to allow differentiation on constants, a dual is given on \( g : \mathbb{Z} \to \mathbb{R} \) with \( g(0) = 0 \) by:

\[
\delta_{M_\lambda} : \text{dom}(\delta_{M_\lambda}) \subset L^p(M_\lambda) \to L^p_1(M_\lambda); \quad (1.30)
\]

\[
\delta_{M_\lambda}(g)(k) = \frac{\lambda}{k+1} \triangle g(k) + \left( \frac{\lambda}{k+1} - 1 \right) g(k).
\]
Keeping conventions, \( L^p(M_\lambda) \) denotes the \( g : \mathbb{Z}^+ \to \mathbb{R} \) such that \( \sum_{k=0}^{\infty} |g(k)|^p e^{-\lambda k} < \infty \) and the subscript \( \perp \) indicates that \( \mathbb{E}[g(M_\lambda)] = 0 \). This inherent to the construction of a dual operator:

\[
\mathbb{E}[\delta_{M_\lambda}(g)(M_\lambda)] = \mathbb{E}[-\nabla(1)(M_\lambda) \cdot g(M_\lambda)] = 0.
\]

Now consider the Stein equation:

\[
\frac{\lambda}{k + 1} \triangle g(k) + \left( \frac{\lambda}{k + 1} - 1 \right) g(k) = h(k) - \mathbb{E}[h(M)] =: h_\perp(k), \quad k \in \mathbb{Z}^+.
\] (1.31)

For calculations, it is easier to unravel \( \triangle \). However note that the product rule in the discrete case is analogous to the one in Itô calculus and not the continuous case:

\[
\triangle (gh) = \triangle(g)h + g\triangle(h) + \triangle(g)\triangle(h).
\]

To solve the Stein equation, we again inspect a tilt by the solution of the homogeneous equation:

\[
\triangle(P_\lambda g)(k) = \left( \frac{\lambda}{k + 1} P_\lambda(k) \right) \triangle g(k) + \triangle(P_\lambda)(k)g(k) = P_\lambda(k)h_\perp(k).
\]

We choose the solution with zero-boundary conditions:

\[
\delta_{M_\lambda}(h_\perp)^{-1}(k) = \frac{k!}{\lambda^k} \int_{[0,k]} \frac{\lambda^l}{l!} (h(l) - \mathbb{E}[h(M_\lambda)]) dl = -\frac{k!}{\lambda^k} \int_{[k,\infty)} \frac{\lambda^l}{l!} (h(l) - \mathbb{E}[h(M_\lambda)]) dl.
\] (1.32)

(1.33)

One may indeed check that it solves (1.31). In approximation problems for discrete distributions the natural function class \( \mathcal{C} \) consists of \( h = 1_A \) for \( A \subset \mathbb{Z}^+ \) or bounded functions. The solution is then directly seen to be bounded from expression (1.33):

\[
|\delta_{M_\lambda}(h_\perp)^{-1}(k)| \leq \|h_\perp\|_\infty \frac{k!}{\lambda^k} \sum_{l=k}^{\infty} \frac{\lambda^l}{l!} \leq \|h_\perp\|_\infty \sum_{l=0}^{\infty} \frac{k!}{l!} e^\lambda \leq e^\lambda \|h_\perp\|_\infty.
\]

We now come to tilting. Instead of \( -\nabla \), we could also have considered \( c \circ -\nabla \), where \( c \) denotes multiplication with \( c : \mathbb{Z}^+ \to \mathbb{R} \). This yields \( \tilde{\delta}_{M_\lambda} = \delta_{M_\lambda} \circ c \) for the dual on shifted domains:

\[
\mathbb{E}[h(M_\lambda)\tilde{\delta}_{M_\lambda}(\tilde{g})(M_\lambda)] = \mathbb{E}[h(M_\lambda)\delta_{M_\lambda}(c(M_\lambda)\tilde{g})(M_\lambda)] = \mathbb{E}[-(\nabla h)(M_\lambda)c(M_\lambda)\tilde{g}(M_\lambda)].
\]

In this way, we obtain different Stein equations and tilts on the domain. Certain approximation results are more readily seen to be valid in specific tilts. We illustrate this for \( c(k) = k \). We obtain for \( k \in \mathbb{Z}^+ \):

\[
\tilde{\delta}_{M_\lambda}(\tilde{g})(k) = \delta_{M_\lambda}(k\tilde{g})(k) = \lambda \tilde{g}(k + 1) - k\tilde{g}(k) = \lambda \triangle \tilde{g}(k) + (\lambda - k)\tilde{g}(k) = h(k) - \mathbb{E}[h(M_\lambda)].
\] (1.34)
Explicit representation for Stein solutions are:

\[
\tilde{\delta}^{-1}_{M}(h_{\perp})(k) = \frac{(k-1)!}{\lambda^k} \int_{[0,k]} \frac{\lambda}{l!} (h(l) - E[h(M_{\lambda})])dl, \quad k > 0, \tag{1.35}
\]

and 0 for \( k = 0 \).

**Theorem 1.2.1** (Stein characterization for the Poisson distribution). A \( \mathbb{Z}^+ \)-valued random variable \( X \) has a Poisson distribution with parameter \( \lambda > 0 \) if and only if for all bounded \( g : \mathbb{Z}^+ \to \mathbb{R} \):

\[
E[\lambda \triangle g(X) + (\lambda - X)g(X)] = E[\lambda g(X + 1) - Xg(X)] = 0. \tag{1.36}
\]

**Proof.** The ‘only if’-direction is direct from duality:

\[
E[\hat{\delta}_{M,\lambda}(g)(X)] = E[g(X) \cdot (-X\nabla)(1)] = 0.
\]

For the ‘if’-direction, consider the solution \( \tilde{g} \) of (1.34) for \( h = 1_A, A \subset \mathbb{Z}^+ \). It was shown that \( \tilde{g}(k) \leq e^{\lambda}/k, k > 0 \). So it is bounded and:

\[
P[X \in A] - P_{\lambda}(1_A) = E[\lambda \triangle \tilde{g}(X) + (\lambda - X)\tilde{g}(X)] = 0.
\]

This enables the reformulation with \( g_h = \tilde{\delta}^{-1}_{M,\lambda}(h_{\perp})(k) \):

\[
|E[h(X)] - E[h(M)]| = |E[\lambda \triangle g_h(X) + (\lambda - X)g_h(X)]| \tag{1.37}
\]

The goal is now to bound \( S_X(g) \),

\[
S_X : \mathcal{G} \to \mathbb{R}; g \mapsto |E[\lambda \triangle g(X) + (\lambda - X)g(X)]|
\]

after a convenient approximation \( \mathcal{G} \) for the \( g_h \).

A Stein approximation for indicator functions that is often found in the literature is given by:

**Lemma 1.2.2** (Stein approximation for indicator functions). Denote with \( \tilde{g}_A \) the solution of (1.34) for \( h = 1_A \). Then:

\[
\|\tilde{g}_A\|_{\infty} \leq \min\{1, \lambda^{-1/2}\}, \quad \|\triangle \tilde{g}_A\|_{\infty} \leq \frac{1 - e^{-\lambda}}{\lambda} \leq \min\{1, \lambda^{-1}\} \tag{1.38}
\]

The second part is proven in the section on discrete distributions in chapter 2. The first part is lengthy and not really needed here. It suffices to know that \( |\tilde{g}_A(k)| \leq e^{\lambda}/k1_{k>0} \) here (an improved bound follows from the section on discrete distributions). See [3] for a proof. Comparing Lemma 1.2.2 with the earlier bound for \( \delta_{M,\lambda} \), we see that tilting
led to another bound being inferred. \( \tilde{g}_A(k) \) actually decays faster than \( 1/k \). However, this can still directly be read off from (1.35).

The Stein characterization (1.2.1) is a discretized version of the characterization for a general normal distribution:

\[
E[\sigma^2 g'(X) + (\mu - X)g(X)] = 0.
\]

Since \( \lambda \) is both the variance and the mean of the Poisson distribution, we get the same form when we replace the derivative by the difference operator. Actually, it must be said that the distribution is the analog of the exponential distribution when viewed in the Pearson/Ord classification (Chapter 2). Due to the different product rule, the characterization can be put in the form of the normal distribution. This structure can help explain some similarities. It explains why both normal and Poisson distributions approximate sums of random variables:

**Theorem 1.2.3.** Let \( (X_i)_{i=1,...,n} \) be independent random variables such that \( P[X_i = 1] = p_i \) and \( P[X_i = 0] = 1 - p_i \). Denote \( W = \sum_{i=1}^n X_i \), \( \lambda_n := E[W] = \sum_{i=1}^n p_i \). With \( M \sim P(\lambda_n) \):

\[
d_{TV}(W, M) \leq \min \{1, \lambda_n^{-1}\} \sum_{i=1}^n p_i^2
\]

\[
\leq \min \{1, \lambda_n\} \max_{i=1,...,n} p_i.
\]

As is usual, we have denoted the total variation distance by:

\[
d_{TV}(X, Y) = \sup_{A \subseteq Z^+} |P[X \in A] - P[Y \in A]|.
\]

**Proof.** We can again use independence to separate \( X_i \) from \( W_i := W - X_i \):

\[
E[X_i g(W)] = E[g(W) \mid X_i = 1]P[X_i = 1] = E[g(W_i + 1)]p_i.
\]

Thus:

\[
|E[\lambda_n g(W + 1) - W g(W)]| = \left| \sum_{i=1}^n p_i E[g(W + 1) - g(W_i + 1)] \right|
\]

\[
\leq \sum_{i=1}^n \lambda_n \| \triangle g \|_{\infty} E \underbrace{W - W_i}_{=X_i}
\]

\[
= \| \triangle g \|_{\infty} \sum_{i=1}^n p_i^2.
\]

Lemma 1.2.2 and (1.37) yield the result. \( \square \)

If \( \lambda_n \to \lambda \), then \( d_{TV}(M_n, M_\lambda) \to 0 \). Since a binomial distribution is a sum of i.i.d. indicator functions, the triangle inequality for the total variation distance gives the classical result:
Theorem 1.2.4 (Poisson limit theorem). Consider a sequence of random variables \((X_n)_{n \in \mathbb{N}}\), with distribution \(X_n \sim \text{Bin}(n, p_n)\). If \(\lim_n np_n = \lambda > 0\), then \(X_n\) converges in distribution to a Poisson random variable with parameter \(\lambda\).

Again, it is possible to derive limit theorems under slight dependency.

Remark 1.2.5. As stated, we reasoned less directly than we could have, with the operators \(\delta, \nabla\) and \(\triangle\). Direct computations would have led to the same results. For regular one-dimensional distributions having density \(p\) with respect to Lebesgue or counting measure, a Stein characterization is directly found from:

\[
E \left[ \frac{(pg)'}{p}(X) \right] = 0 \text{ for } g \in \mathcal{G} \iff X \sim p(x)dx, \quad (1.39)
\]

\[
E \left[ \frac{\triangle (pg)}{p}(X) \right] = 0 \text{ for } g \in \mathcal{G} \iff X \sim p(l)dl, \quad (1.40)
\]

for a wide enough class \(\mathcal{G}\). At least, this is the case if zero-boundary conditions can properly be taken care of. We discuss this further in the next chapter. We have chosen not to take this approach here for three reasons. First of all, it makes a transition to Malliavin calculus smoother. Also, it enables unified language for Stein’s framework. Finally, it is not directly clear how this density approach can be generalized to multiple dimensions, where a fundamental theorem of calculus (disintegration with respect to a measure) is more difficult to obtain or work with. In infinitely many dimensions, such as for Itō diffusions, it is less obvious to have an explicit density (which led to Feynmann integrals). However, there are notions of differentiability available which in one dimension reduce to absolute continuity. We still gave solutions to the Stein equation in notions that rely strongly on densities however.

1.3 Zolotarev type distances and optimal transport

When studying convergence of measures over a suitable set of functions \(\mathcal{C}\), equicontinuity properties can be used to equivalently study uniform convergence over compact subsets of \(\mathcal{C}\). Consider again a Polish space \(S\) with topology \(\mathcal{B}\), and \(\mu, \nu \in \mathcal{P}(S)\), the space of probability measures on \(S\). Furthermore, let \(X \sim \mu, Z \sim \nu\) on a probability space \((\Omega, \mathcal{F}, \mathbb{P})\). Then weak convergence often gives rise to Zolotarev-type distances:

\[
d_H(\mu, \nu) := d_H(X, Z) = \sup_{h \in \mathcal{H}} \left| \int_S h d\mu - \int_S h d\nu \right| = \sup_{h \in \mathcal{H}} |E[h(X)] - E[h(Z)]|. \quad (1.41)
\]

Here, \(\mathcal{H} \subset \mathcal{C}\) is a sufficiently regular set of functions \(f : S \to \mathbb{R}\) (such as a compact subset of \(\mathcal{C}\)). If \(\mathcal{H}\) is measure-determining or separating, meaning

\[
\int_S h d\mu = 0, \ \forall h \in \mathcal{H} \implies \mu = 0,
\]

...
formula (1.41) defines a metric on the measures with finite values. As can be seen from Lemma 1.1.6, Stein’s method often allows to give quantitative approximations that are uniform over all \( h \) in a suitable \( \mathcal{H} \), whereas general theory on equicontinuity would only yield qualitative statements\(^4\). We will restrict our attention to the following cases:

- \( \mathcal{S} = \mathbb{R}, \mathcal{H} = \{ \mathbb{1}_{(-\infty,x]} \mid x \in \mathbb{R} \} \): We obtain the **Kolmogorov distance**, which in terms of (1.42) is given by:

\[
d_{kol}(X,Z) := \sup_{x \in \mathbb{R}} |\mathbb{P}[X \leq x] - \mathbb{P}[Z \leq x]|. \tag{1.43}
\]

Thus, distribution functions are compared uniformly. It is important in Glivenko-Cantelli type results for empirical measures. It is also connected to the Kolmogorov-Smirnov test.

- \( \mathcal{H} = \{ h \in C_b(S) \mid h \text{ Lipschitz, } \|h\|_{\infty} + \|h\|_{Lip} \leq 1 \} \)\(^5\): This yields a **bounded Wasserstein metric** or the **Fortet-Mourier metric** \( d_{bW} \). It metrizes weak convergence of measures, as will be shown. Alternatively, we may take \( \|h\|_{\infty} \leq 1, \|h'\|_{\infty} \leq 1 \) for \( \mathcal{H} \) to metrize weak convergence. For practical convenience, we often consider the following stronger metric.

- \( \mathcal{H} = \{ h \in C(S) \mid h \text{ Lipschitz, } \|h\|_{Lip} \leq 1 \} =: \text{Lip}_1(S) \): This gives the **Wasserstein metric** \( d_W \) on measures with finite first moment. It metrizes weak convergence on \( \mathcal{P}_1(S) \). Note that \( h \) is not necessarily bounded.

- \( \mathcal{H} = \{ \mathbb{1}_A \mid A \in \mathcal{B} \} \): We obtain the **total variation metric** \( d_{TV} \), which compares probabilities on every Borel set uniformly. Convergence in this metric is a strong concept.

These distances may sometimes be difficult to compute or interpret exactly however. Also, estimating them for two random variables \( X, Z \) may be complicated because:

1. The distribution of \( X \) or \( Z \) is not well-known.

2. It is not clear how to compare the distributions on \( \mathcal{H} \).

For our interests in stochastic approximation, we assume that we at least have “sufficient” information on the distribution of \( Z \). When we know both distributions well, other techniques can be explored to compute the distances. Inspiration may for example be drawn from optimal transport problems. There, one is also interested in finding optimal couplings (joint probability laws) of \((X,Y)\) on a probability space, for the above metrics. Good references include [29] and [22].

This section has two goals. First, to motivate and detail some of the relations between the above distances. Second, to gain some insight when it is relevant or not to

\(^4\)Refer to Theorem 1.3.3.

\(^5\)See B.1.1.
consider Stein’s method. When it is possible to compute the above distances, Stein’s framework may sometimes still be useful to achieve quick out-of-the-box bounds as an alternative. To improve fluency and maintain an overview, proofs are given in the appendix.

1.3.1 Relations between the metrics

The following picture details some of the relations between the metrics:

\[ \begin{align*}
  &d_{TV} \\
  d_K &\quad \|p\|_{\infty}<\|p\|_{\infty} \quad \rightarrow \\
  F \in C(\mathbb{R}) &\quad \rightarrow \\
  d_{bW} &\quad \text{Unif. int.}
\end{align*} \]

A solid line indicates a relation of the form \( d_\alpha \leq C d_\beta \) where \( d_\beta \) is at the head of the arrow and \( d_\alpha \) at the tail. The dashed lines indicate that convergence in \( d_\alpha(\cdot, Z) \) imply convergence for \( d_\beta(\cdot, Z) \), under certain assumptions. The dotted line indicates a non-linear relation. Some results are:

**Proposition 1.3.1.** Consider \( S \)-valued random variables \( X, (X_n)_{n \in \mathbb{N}}, Z \). For the Kolmogorov metric, we assume \( S = \mathbb{R} \). For the Wasserstein metric, we assume finite first moments.

1. \( d_K(X, Z) \leq d_{TV}(X, Z), d_{bW}(X, Z) \leq d_{W}(X, Z), d_{bW}(X, Z) \leq d_{TV}(X, Z) \).

2. For \( n \to \infty \), \( d_K(X_n, Z) \to 0 \) implies \( d_{bW}(X_n, Z) \to 0 \).

3. If \( Z \) has a continuous distribution function, \( d_{bW}(X_n, Z) \to 0 \) implies \( d_K(X_n, Z) \to 0 \) for \( n \to \infty \).

4. If \( (X_n)_{n \in \mathbb{N}} \) have uniformly integrable first moments, meaning
   \[ (\forall \epsilon > 0)(\exists K_\epsilon \subseteq \text{compact } S)(\sup_{n \in \mathbb{N}} \mathbb{E}[X_n 1_{X_n \in K_\epsilon^c}] < \epsilon), \]
   then \( d_{bW}(X_n, Z) \to 0 \) implies \( d_W(X_n, Z) \to 0 \).

5. If \( Z \) is absolutely continuous with respect to Lebesgue measure with density \( f \in L^\infty(\mathbb{R}) \), then:
   \[ d_K(X, Z) \leq d_{bW}(X, Z) + \sqrt{\|f\|_{\infty} d_{bW}(X, Z)} \]
   \[ \leq (\sqrt{2} + \sqrt{\|f\|_{\infty}}) \sqrt{d_{bW}(X, Z)}, \]
   \[ d_K(X, Z) \leq \sqrt{\|f\|_{\infty} d_{W}(X, Z)}. \]
Point 1. is direct when taking into account Lemma 1.3.9. The other points are commented on in Appendix C. It should be noted that bounds on the Zolotarev-type distances may be suboptimal with these relations. For example, \( d_K \) and \( d_W \) can have the same rate of convergence, although the above inequalities would yield \( d_K(X_n, Z) \leq \sqrt{\|f\|_\infty} d_W(X_n, Z) \). In general, if many distributions are considered at once, the guaranteed rates (e.g. \( d_W(X_n, Z) = O(n^{-1/2}) \)) may be crude.

It is important to know that the bounded Wasserstein distance metrizes weak convergence. If \( T \) is a Polish space, then laws \( \mu_n \in \mathcal{P}(T) \) converge weakly to \( \mu \in \mathcal{P}(T) \) iff:

\[
\int h \, d\mu_n \to \int h \, d\mu, \quad \forall h \in C_b(T).
\] (1.46)

Some familiarity with weak convergence is assumed, but important notions will be mentioned. For more information, see [28], [15] or [9]. That \( d_{bW} \) metrizes weak convergence means:

**Proposition 1.3.2.** For a separable metric space \((T, d)\), probability laws \( \mu_n \) converge weakly to a law \( \mu \) on \( T \) if and only if \( d_{bW}(\mu_n, \mu) \to 0 \).

There are actually multiple ways to metrize weak convergence based on equicontinuity properties. Next, a result that holds in general is (see chapters 32–34 of [27] for more information):

**Theorem 1.3.3 (Equicontinuity).** Let \( C \) be a complete metrizable topological vector space or a barreled space, and \( Q \subset C' \). Then \( Q \) is equicontinuous if and only if \( Q \) is weakly bounded, meaning \( \forall f \in C: \sup_{q \in Q} |q(f)| < \infty \). In this case the following ways of convergence are equivalent for \( q_n, q \in Q \) and \( D \subset_{\text{dense}} C \):

1. \( q_n(f) \to q(f), \quad \forall f \in C \).
2. \( q_n(f) \to q(f), \quad \forall f \in D \).
3. \( \sup_{f \in K} |q_n(f) - q(f)| \to 0, \quad \forall K \subset_{\text{compact}} C \).

It is not our intention to delve deep into the framework of topological vector spaces. In the appendix, Proposition C.1.3, a specific procedure is outlined for our purposes. It is sufficient to know that \( C = C_b(S) \) with the supremum norm is a complete metrizable topological vector space and \( \mathcal{P}(S) \subset C' \) (where \( q(f) = E[f(X)] \) for a random variable). The bounded Lipschitz functions with Lipschitz constant at most 1 form a compact subset of \( C_b(L) \), where \( L \subset S \) is compact. We mention the theorem to have some general insight in equicontinuity properties.
In practice, it is often possible to achieve bounds in the stronger Wasserstein than bounded Wasserstein distance. This can be noted in Lemma 1.1.3 or 1.1.6. It complies intuition better because there is no trade-off between bounds on $h$ and $h'$.

In the literature on Stein’s method, stochastic approximation results are often stated in terms of the above distances instead of bounds on (1.1). Combining Lemma 1.1.6 with Lemma 1.1.3 for example yields (due to $\|h'\|_{\infty} \leq 1$):

**Proposition 1.3.4.** Let $(X_i)_{i=1,...,n}$ be independent centered random variables with finite third moments. Let $\mathbb{E}[X_i^2] = 1$ and $W = n^{-1/2} \sum_{i=1}^{n} X_i$. Then:

$$d_W(W, Z) \leq 2 \left( \frac{1}{n^{1/2}} + \frac{1}{2n^{3/2}} \sum_{i=1}^{n} \mathbb{E}|X_i|^3 \right).$$

(1.47)

Theorem 1.2.3 is another example.

### 1.3.2 Transportation metrics and optimal couplings

In the subject of optimal transport, the Zolotarev metrics are often stated in a different form, not depending on $H$. The subject is confronted with finding optimal ways to displace an initial distribution $X$ into another, $Z$, with respect to a cost function. A typical example is distributing pizzas from an enterprise with multiple establishments to clients over a city, while keeping into account trafficking costs. There is a lot of theory about these problems, some of which may aid to know when the metrics can be bounded or computed more directly. Actually, we have used the name Wasserstein distance for a particular case of a whole class of distances on subsets of $\mathcal{P}(S)$. Consider $p \in [1, \infty)$, an $x_0 \in S$ and a metric $d$ that makes $S$ complete for the given topology. The space of **probability measures with finite $p$-th moment** is defined as:

$$\mathcal{P}^p(S) := \{ \mu \in \mathcal{P}(S) : \int d(x, y)^p \mu(dy) < \infty \}.$$

This definition is independent of the choice of $x_0$. Given $\mu$ and $\nu$ in $\mathcal{P}^p(S)$, we say that $\gamma \in \mathcal{P}(S \times S)$ is a **coupling** of $\mu$ and $\nu$ if it has marginals $\mu$ and $\nu$ in this order. This means: $\gamma(B \times S) = \mu(B)$, $\gamma(S \times B) = \nu(B)$, for $B \in \mathcal{B}$. We indicate the set of all couplings of $\mu$ and $\nu$ by $\Pi(\mu, \nu)$. The coupling is **realized** by random variables $X, Y : \Omega \to S$ on a common probability space $(\Omega, \mathcal{A}, \mathbb{P})$, if $(X, Y) \sim \gamma$.

**Definition 1.3.5.** For $p \in [1, \infty)$ and $\mu, \nu \in \mathcal{P}(S)$, we define the $\mathcal{L}^p$ **Wasserstein distance**:

$$\mathcal{W}^p(\mu, \nu) := \inf_{\gamma \in \Pi(\mu, \nu)} \left( \int d(x, y)^p \gamma(dx, dy) \right)^{1/p} = \inf_{\substack{X \sim \mu \\ \ \ Y \sim \nu \\ \ Y \sim \nu}} \mathbb{E}[d(X, Y)^p]^{1/p}.$$

In the last infimum, we consider $X, Y$ defined on a common probability space such that $X \sim \mu$, $Y \sim \nu$. For $p = 1$, it is also called the **Kantorovich distance**.
First note that the expression above is finite on \( \mathcal{P}^p(S) \) by the triangle inequality of \( d \). Also note that there may be multiple metrics for the Polish space \( S \) that turn it into a complete space. Changing the underlying metric actually is a tool that is used in the theory of Markov processes to study stationary measures\(^6\). By coupling arguments, it may be checked that this defines a metric on \( \mathcal{P}^p(S) \), see [29].

The relation to the Wasserstein metric is:

**Theorem 1.3.6 (Rubinstein-Kantorovich duality).** For \( \mu, \nu \in \mathcal{P}(S) \):

\[
W^1(\mu, \nu) = \sup_{h \in \text{Lip}_1(S, d)} \left| \int h d\mu - \int h d\nu \right|.
\]

Again, \( \text{Lip}_1(S, d) \) denotes the Lipschitz functions such that the Lipschitz constant is at most 1. There are similar results for \( p > 1 \). Moreover, \( (\mathcal{P}^p(S), W^p) \) is a complete separable metric space. For all these statements, the reader is referred to [29]. Furthermore, Prohorov’s theorem yields that the infimum is attained in a so-called optimal coupling:

**Lemma 1.3.7 (Existence of optimal couplings).** For \( \mu, \nu \in \mathcal{P}(S) \) and \( p \in [1, \infty) \) there exists a coupling \( \gamma_0 \in \Pi(\mu, \nu) \) such that

\[
W^p(\mu, \nu) = \left( \int d(x, y)^p \gamma_0(dx dy) \right)^{1/p}.
\]

Explicit optimal couplings for distributions on \( \mathbb{R} \) can now be computed. In the case of the \( \mathcal{L}^p \) Wasserstein distances:

**Example 1.3.8 (Coupling by monotone arrangement).** For \( \mu, \nu \in \mathcal{P}(\mathbb{R}) \) let \( F_\mu \) and \( F_\nu \) denote their distribution functions. Further, denote the left-continuous generalized inverse by \( F^{-1}_\mu \):

\[
F^{-1}_\mu(u) = \inf\{c \in \mathbb{R} : F_\mu(c) \geq u\}, \quad u \in (0, 1).
\]

If \( U \sim U(0, 1) \), then \( F^{-1}_\mu(U) \sim \mu \) because \( F^{-1}_\mu(U) \leq c \) if and only if \( U = F(F^{-1}_\mu(U)) \leq F(c) \) (some care has to be taken when the distribution function exhibits jumps). We can couple \( \mu \) and \( \nu \) by \( (F^{-1}_\mu(U), F^{-1}_\nu(U)) \). This is called the **coupling by monotone arrangement** and is optimal with respect to the \( W^p \)-distances for all \( p \geq 1 \). See chapter 3 in [22] for a proof. The reason for the name is that lower-lying parts of equal mass of the distributions are coupled to each other. In particular, if the distribution functions are injective, \( u \)-quantiles of the two distributions are identified with each other. By the Rubinstein-Kantorovich duality, Theorem 1.3.6, this implies the following exact representations for the Wasserstein distance on the line:

\[
d_W(X, Z) = \int_0^1 |F^{-1}_\mu(u) - F^{-1}_\nu(u)| du = \int_{-\infty}^{\infty} |F_\mu(x) - F_\nu(x)| dx. \tag{1.48}
\]

\(^{6}\)These correspond to the distribution of \( Z \) that is known well.
Figure 1.1: Distribution function $F$ with values for the inverse $q_i = F^{-1}(p_i)$

There is also a reformulation for the total variation distance.

Lemma 1.3.9 (Total variation distance). Let $X$ and $Z$ be $S$-valued random variables and $a < b$ real numbers. Then:

$$d_{TV}(X, Z) = \frac{1}{b - a} \sup_{f:S \to [a, b]} |\mathbb{E}[f(X)] - \mathbb{E}[f(Z)]|$$

$$= \inf_{\tilde{X} \sim X, \tilde{Z} \sim Z} \mathbb{P}[\tilde{X} \neq \tilde{Z}].$$

If $S$ is discrete, then:

$$d_{TV}(X, Z) = \frac{1}{2} \sum_{s \in S} |\mathbb{P}[X = s] - \mathbb{P}[Z = s]|.$$ (1.51)

In the supremum, $f$ is supposed to be measurable. See Lemma C.1.7 for a proof of the first equality. For the second we refer to section 1.6 of [10]. (1.49) implies the following estimates:

$$|\mathbb{E}[f(X)] - \mathbb{E}[f(Z)]| \leq (\sup f - \inf f) d_{TV}(X, Z)$$

(1.50) states that the total variation distance can also be seen as the $L^1$-Wasserstein distance of $S$ with the discrete metric (which is not necessarily Polish).

Finally, we have the following transformation rule:

Lemma 1.3.10. Let $X$ and $Y$ be Banach space-valued random variables with finite first moment and let $\sigma > 0$. Then $d_W(\sigma X, \sigma Y) = \sigma d_W(X, Y)$.

Proof. This follows by noting that $\|\sigma^{-1}h(\sigma \cdot)\|_{Lip} = \|h\|_{Lip}$. \qed
1.3.3 Explicit computation on the line

Even with the above properties, it may be difficult to compute or estimate the Wasserstein distance explicitly when there is sufficient information on $X$ and $Z$. In this section, we will take a short digression on some elementary methods to compute $d_W$.

Consider real random variables $X$ and $Z$ with finite first moment and densities $p_1$ and $p_2$ with respect to Lebesgue measure $dx$. Their Wasserstein distance is given by:

$$d_W(X, Z) = \sup_{h \in \text{Lip}_1(\mathbb{R})} \left| \int h(x)(p_1 - p_2)(x)dx \right|$$

(1.52)

Drawing a picture of $\Delta := p_1 - p_2$ may give some insight into how the distributions compare. To find optimal $h$, we would like $h$ to be as large as possible when $\Delta$ is positive and as small as possible when $\Delta$ is negative. Moreover, $h$ can only make its quickest ascent and descent when locally following a straight line with slope 1 or $-1$ respectively. Denote the broken lines with slopes $\pm 1$ and $n$ break-points as follows (see Theorem B.3.3):

$$\mathcal{L}_n(\mathbb{R}) := \left\{ h \in \text{Lip}_1(\mathbb{R}) \right\} \left( \exists c_1 < \ldots < c_n \in \mathbb{R} \right) \left( h' = \pm \sum_{i=0}^{n} (-1)^i \mathbb{1}_{(c_i, c_{i+1})} \right).$$

We set $c_0 := -\infty$, $c_{n+1} := \infty$.

Lemma 1.3.11. Let $n \in \mathbb{N}$. If for $\mu(dx) = p_1(x)dx$ and $\nu(dx) = p_2(x)dx$ there exist $c_1 < \ldots < c_n$ in $\mathbb{R}$ such that $\Delta := p_1 - p_2$ changes sign at each $c_i$, then the supremum in (1.52) can be taken over $h \in \bigcup_{i=1}^{n-1} \mathcal{L}_i(\mathbb{R})$.

If $n = 1$, then:

$$d_W(X, Z) = |\mathbb{E}[X] - \mathbb{E}[Z]|$$

If $n = 2$, then we may restrict $h$ in (1.52) to broken lines of the form $h_a(x) = |x - a|$, $a \in \mathbb{R}$.

Proof. We will try to improve arbitrary $h \in \text{Lip}_1(\mathbb{R})$ to approach the supremum in:

$$\sup_{h \in \text{Lip}_1(\mathbb{R})} \int h(x)(p_1 - p_2)(x)dx.$$ 

(1.53)

Without loss of generality, we may assume that $\Delta \leq 0$ on $(-\infty, c_1)$. On this interval, $h(x) \geq h(c_1) + (x - c_1) =: l(x)$, and so

$$h(x)\Delta(x) \leq l(x)\Delta(x) \quad x \in (-\infty, c_1).$$

Assuming $n \geq 2$, we connect $h(c_1)$ and $h(c_2)$ by a broken line with slopes $\pm 1$ such that it takes maximal values (this corresponds to a triangle above $h$). With $x_1 := c_1 + c_2 + h(c_2) - h(c_1)$ this is $h(x) \leq h(x_1) - |x - x_1| =: l(x)$. Again, $h(x)\Delta(x) \leq l(x)\Delta(x)$

\footnote{\begin{footnotesize}Meaning $\Delta(x)\Delta(y) \leq 0$ for almost every $x \in (c_{i-1}, c_i), y \in (c_i, c_{i+1})$ and $i = 1, \ldots, n$.\end{footnotesize}}
for \( x \in (c_1, c_2) \). Note that so far, the \((x, l(x))\) constitute a broken line with at most one break point. There might be none if \( h \) already was a straight line with slope 1 in between \( c_1 \) and \( c_2 \) (\( x_1 \in \{c_1, c_2\} \)). We may continue this procedure by alternately bounding \( h \) from below and above, each time possibly breaking \( l \) in between subsequent pairs \((c_i, c_{i+1})\). For \( n \geq 1 \) and \( x \geq c_n \) we finish by the bound \( h(x)\Delta(x) \leq l(x)\Delta(x) \) with \( l(x) = h(c_n) + (-1)^{n+1}x \). In the end, we retrieve for \( l \) a broken line in \( \bigcup_{i=1}^{\lceil \frac{n}{2} \rceil} L_i(\mathbb{R}) \) such that:

\[
\int h(x)(p_1 - p_2)(x)\,dx \leq \int l(x)(p_1 - p_2)(x)\,dx.
\]

Thus it suffices to only consider \( h \) as in the claim.

For \( n = 1 \), the result is direct. For \( n = 2 \), the slopes of \( l \) have opposite sign on \(( -\infty, c_1) \) and \(( c_2, \infty) \), thus we have a broken line of the given form. Note that this even holds if \( n \) can be 1 as well.

In the case of two sign changes we get the following nice parametric form:

\[
d_W(X, Z) = \sup_{a \in \mathbb{R}} \left| \int |x - a|(p_1 - p_2)(x)\,dx \right| = \sup_{a \in \mathbb{R}} |\mathbb{E}[X_1 - a] - \mathbb{E}[X_2 - a]|.
\]

We now illustrate how this yields Wasserstein distances between two normal distributions. In the literature on Stein’s method, one usually finds out-of-the-box estimates however.

**Proposition 1.3.12.** For \( N_i \sim N(\mu_i, \sigma_i^2), \sigma_i > 0, i = 1, 2 \) and \( \sigma_1 \neq \sigma_2 \), we have with \( a_0 = \frac{\sigma_2 \mu_1 - \sigma_1 \mu_2}{\sigma_2 - \sigma_1} \):

\[
d_W(N_1, N_2) = |\mathbb{E}[N_1 - a_0] - \mathbb{E}[N_2 - a_0]|. \tag{1.54}
\]

In particular if \( \mu_1 = \mu_2 \):

\[
d_W(N_1, N_2) = \sqrt{\frac{2}{\pi}} |\sigma_2 - \sigma_1|. \tag{1.55}
\]

**Proof.** First note that the difference of the densities exhibits two sign changes. We have

\[
V_i(\mu_i + a) := |\mathbb{E}[N_i - (\mu_i + a)]| = \int_{-\infty}^{a} (a - x) \frac{e^{-x^2/2\sigma_i^2}}{\sqrt{2\pi \sigma_i^2}} \,dx + \int_{a}^{\infty} (x - a) \frac{e^{-x^2/2\sigma_i^2}}{\sqrt{2\pi \sigma_i^2}} \,dx,
\]

and

\[
V_i'(\mu_i + a) = 2\Phi \left( \frac{a}{\sigma_i} \right) - 1.
\]

Thus

\[
(V_1 - V_2)'(a) = 0 \iff \frac{a - \mu_1}{\sigma_1} = \frac{a - \mu_2}{\sigma_2} \iff a = \frac{\sigma_2 \mu_1 - \sigma_1 \mu_2}{\sigma_2 - \sigma_1}.
\]
Furthermore, \((V_1 - V_2)\)' changes sign in \(a_0\) because \(\Phi\) is strictly increasing. We thus find a local optimum. Arguing as in the proof of the previous lemma, it can be seen that the break point for optimal \(a\) should be between the sign changes of the densities. So by compactness, \(a_0\) is the global optimum.

When \(\mu_1 = \mu_2\), we have \(a_0 = \mu\) and it can be computed that:

\[
E|N_i - \mu_i| = 2\int_0^\infty \frac{xe^{-x^2/2\sigma_i^2}}{\sqrt{2\pi}\sigma_i} dx = 2\sigma_i \int_0^\infty \frac{xe^{-x^2/2}}{\sqrt{2\pi}} dx = \sqrt{\frac{2}{\pi}} \sigma_i.
\]

\(\square\)

### 1.4 A first overview

The methods for Normal and Poisson approximation seem to have much in common. Yet, it is not straightforward to crystallize their success after a first encounter. The goal here is to outline an intuitive view on Stein’s method. Later, after having seen improved techniques, we will try to motivate why it works. Now, it should be taken as an intuitive recipe to which the coming descriptions should be compared.

**The goal:** We want to compare two random variables/processes \(X\) and \(Z\) on a metric space of functions \(C\) that is measure-determining, in the following way:

\[
E[h(X)] - E[h(Z)], \quad h \in C.
\] (1.56)

Here, \(C\) may be the continuous functions with supremum-norm or linear spans of specific indicator functions with supremum-norm. We can often restrict ourselves to more regular \(\mathcal{H}\) as discussed in the previous section.

**Difficulty/context:**

A. Only the distribution of \(Z\) is known well.

B. It may be difficult to compare the distributions by (1.56) on \(\mathcal{H}\).

Otherwise we would more readily rely on direct methods from optimal transport and coupling methods, or direct computation.
Stein’s method:

1. Consider an (unbounded) derivative operator \( D : \text{dom}(D) \subset L^2(Z) \to L^2(Z) \) and compute the dual \( \delta_Z \). Then, infer a convenient **Stein approximation** \( \mathcal{G} \supset \delta^{-1}_Z(\mathcal{H}_\perp) \) where regularity is expressed independently of the distribution of \( Z \). Perhaps a convenient tilt needs to be found. For example so that \( \delta^{-1}_Z : H_{\perp}^{k,\infty}(Z) \to W^{k+1,\infty}(Z) \) exists and is continuous. In the process, we want to retrieve a **Stein characterization**:

\[
E[\delta_Z(g)(X)] = 0, \forall g \in \mathcal{G} \iff X \sim Z.
\]

2. Bound the associated **Stein functional** of the new structure \((\delta_Z, \mathcal{G})\):

\[
S_X : \mathcal{G} \to \mathbb{R}; g \mapsto E[\delta_Z(g)(X)].
\]

**Remark 1.4.1.** To estimate (1.56), Stein’s method thus divides the task into two steps. In the first, much information of \( Z \) is separated. In the second, a new problem is obtained where there is still some left-over information of \( Z \), but now the main task is to exploit the structure of \( X \) with respect to \( \delta_Z \). In this way, the distributions do not need to be directly compared to each other on \( \mathcal{H} \). Also, from the new functional, it can become more clear which structures approximate \( Z \) well.

1. To obtain a Stein characterization, \( \text{dom}(D) \) should include constants.

2. For \( p \in [1, \infty] \), \( W^{k,p}(Z) \) currently represents an intuitive space of “weakly differentiable” functions \( g \) with \( k \) derivatives such that all derivatives are in \( L^p(Z) \). \( H^{k,p}(Z) \) are all \( k \)-times weakly differentiable \( h \) such that the highest order derivative is in \( L^p(Z) \). Note that \( W^{k,\infty}(Z) \) only depends on \( Z \) through its state space \( S \) and thus trickles down regularity to \( X \) independent of the distribution. Here, we consider \( k = 0, 1 \). In the case of real-valued random variables we mean Sobolev spaces.

3. Since \( \mathcal{H} \) is measure-determining, \( \mathcal{G} \) is as well. The direct implication of the above Stein characterization thus follows immediately from this. For the converse direction, \( \mathcal{G} \) should not be too large. It can for instance be the closure of \( \delta^{-1}_Z(\mathcal{H}_\perp) \) in \( W^{k+1,\infty}(Z) \).

4. We expect that with a good tilt, \( \delta^{-1}_Z : \mathcal{H}_\perp \subset H^{k,p}(Z) \to \delta^{-1}_Z(\mathcal{H}_\perp) \subset W^{k+1,p}(Z) \) is bicontinuous (continuous with continuous inverse). Tilting means that we consider \( \delta_Z \circ c \) instead of \( \delta_Z \), where \( c \) denotes multiplication by some \( c : S \to \mathbb{R} \).

5. \( X \mapsto S_X \) measures the distance to \( Z \). If \( \|S_X\|_{\mathcal{G}} = \sup_{g \in \mathcal{G}} |S_X(g)| \) is small, then \( X \) is approximately distributed according to \( Z \). This is termed **Stein’s heuristic**. Due to the treatise in section 1.3 it no longer is a heuristic:

\[
\|S_X\|_{\mathcal{G}} = \sup_{g \in \mathcal{G}} |E[\delta_Z(g)(X)]| = \sup_{h \in \mathcal{H}} |E[h(X)] - E[h(Z)]|.
\]
The advantages of Stein’s method as compared to other methods are:

- It is more natural to consider other metric structures on the distributions of interest that are based on duality (measures as functionals on spaces of functions).

- We may consider more complex structures for $X$. As pointed out, a key advantage is that some dependence can be dealt with.

- We retrieve quantitative statements. However, not all metrics have convenient interpretations. Also, depending on how much freedom one wishes to allow, the bounds might be crude.

- The method itself will be seen to allow structural decompositions of a large class of random variables. This can be exploited to infer strong approximation results. See chapter 4.

Stein’s method is strongly embedded in PDE-theory in case continuous densities for $Z$ are considered. Sturm-Liouville spectral decomposition, regularity of PDE’s and semi-group theory can all come into play. This may be with or without stochastic interpretations. We will investigate Malliavin calculus, which is a stochastic counterpart for regularity of PDE’s, such as Hörmander’s theory. Therefore, Stein’s method will actually correspond to a symmetrized operator $\delta \circ D$ instead of $\delta$.

For part 2., coupling techniques, analytical and combinatorial manipulations can all be useful. Coupling methods seem somehow anti-Stein in nature, since they again introduce a direct comparison between two distributions. See the size-bias coupling in section 2.4 for an illustration.
Chapter 2

The one-dimensional density view

When $Z$ is a one-dimensional real random variable having a density function with respect to a canonical measure, it is possible to unravel the above abstractions to “elementary” computations with integrals and derivatives (elementary in the sense that they mostly rely on classical analysis or combinatorics). Because they are more familiar and exhibit a stronger structure, properties may be more readily inferred. The drawback however is the trade-off with generalizability.

In this chapter, the one-dimensional Lebesgue measure and counting measure will be investigated as canonical reference measure. We provide a discussion on how Stein characterizations and approximations may be inferred. We provide a list of Stein equations for some Pearson and Ord type distributions. In the treatment, the Stein kernel takes a central position because it has multiple useful connections to Stein approximations as well as bounding $S_X$. For the latter, it gives rise to a distributional transform called the zero-bias coupling. Finally, a more direct method to compare densities with nested supports will be outlined.

First, we summarize some properties of dual operators because they are central to Stein’s method. In chapter 4, some time will be spent on working out a detailed framework for Gaussian analysis. We choose to already present them to improve the reader’s view on a unifying trend throughout this text.

2.1 Dual operators

In Stein’s method, a Sturm-Liouville type transformation is constructed to eliminate $E[h(Z)]$. It arises from a dual operator with respect to differentiation. We give a brief overview here and refer to [27] or [6] for more information on the vast subject of duality and Sobolev spaces.

2.1.1 Generalities

Let $X, Y$ be Banach spaces. We call a linear operator

$$D : \text{dom}(D) \subset X \rightarrow Y$$


an unbounded operator if and only if \( \text{dom}(D) \) is a linear subspace of \( X \). It is called closed if
\[
\text{graph}(D) := D = \{(x, Dx) \mid x \in \text{dom}(D)\}
\]
is closed in \( X \times Y \) (with the norm \( \|(x, y)\| = \|x\| + \|y\| \)). This is equivalent to: if for \( x_n \in \text{dom}(D), n \in \mathbb{N} \), and \( n \to \infty \):
\[
x_n \to x \in X \text{ and } Tx_n \to y \in Y, \text{ then } x \in \text{dom}(D) \text{ and } Tx = y.
\]
In this case, \( \text{dom}(D) \) is complete for the norm \( \|f\|_D = \sqrt{\|f\|_X^2 + \|Df\|_Y^2} \). \( D \) is called closable if its closure in \( X \times Y \) coincides with the graph of a closed operator \( \bar{D} \). Going from \( D \) to \( \bar{D} \) is called closing the operator. The new operator is usually also denoted by \( D \). Closing may be done as follows: if \( x_n \in \text{dom}(D) \to x \in X \) for \( n \to \infty \) and \( (Dx_n)_{n \in \mathbb{N}} \) is Cauchy in \( Y \), then define \( Dx := \lim_n Dx_n \). To be well-defined, it has to be verified that the same limit is reached for any other approximating sequence. If also \( \tilde{x}_n \in \text{dom}(D) \to x \) and \( (D\tilde{x}_n)_{n \in \mathbb{N}} \) is Cauchy, then it is needed that \( D(x_n - \tilde{x}_n) \to 0 \). Summarized:

**Lemma 2.1.1.** \( D \) is closable iff for any sequence \( (x_n)_{n \in \mathbb{N}} \) in \( X \) that converges to 0 and \( Dx_n \to y \in Y \), it follows that \( y = 0 \).

\( D \) is called densely defined if \( \text{dom}(D) \) is dense in \( X \). The idea behind an unbounded operator is that it can not be defined up to the complete space since values may blow up too fast. Consider differentiation on \((0, 1)\) for example. For \( f_n(x) := \sin(nx) \), we have that \( \|f_n'\|_{L_p'(0,1)} \to \infty \) in \( L_p'(0,1) \) while \( \|f_n\|_{L_p(0,1)} \) remains bounded. We will only be concerned with
\[
D : \text{dom}(D) \subset L^p(\Omega, \mathcal{F}, \mu) \to L^p(\Omega, \mathcal{F}, \mu),
\]
where \((\Omega, \mathcal{F}, \mu)\) is a \( \sigma \)-finite measure space, Sometimes, a whole range of operators \((D_p)_{p \in [1, \infty]}\) is denoted by \( D \). When dealing with notions such as closedness, one has to keep in mind the Banach space theorems such as the closed graph theorem that states that closed operators with \( \text{dom}(D) = X \) are bounded.

To start, we assume that \( \text{dom}(D) \) contains a fixed subspace that is dense in \( L^p(\mu) \), for all \( p < \infty \) and that \( D \) is closeable on these spaces. For \( \Omega \subset \mathbb{R}^n \) open, this subspace may be \( C^\infty_p(\Omega) \). The dual space of \( L^p(\mu) \) can be identified with \( L^{p'}(\mu) \) for \( \frac{1}{p} + \frac{1}{p'} = 1 \) and \( p < \infty \), Theorem B.3.5. A dual operator \( \delta := DT \) may thus be constructed on \( L^{p'}(\mu) \):
\[
\delta : \text{dom}(\delta) \subset L^{p'}(\mu) \to L^{p'}(\mu), \quad (2.1)
\]
that is uniquely defined by
\[
\mathbb{E}[D(f)(Z)g(Z)] = \mathbb{E}[f(Z)\delta(g)(Z)], \quad \forall f \in \text{dom}(D). \quad (2.2)
\]
\(^1\)The graph of \( D \) actually is the same object as \( D \). “Graph” just stresses that it is regarded as a subset of \( X \times Y \).
We use the suggestive notation \( \mathbb{E}[g(Z)] := \int_{\Omega} g(z) \mu(dz) \), although we are not restricting ourselves to probability spaces yet. Also:

\[
\operatorname{dom}(\delta_p') = \{ g \in L^{p'}(\mu) \mid (\exists C > 0)(\forall f \in \operatorname{dom}(D))(\mathbb{E}[(Df)(Z)]g(Z)) \leq C\|f\|_{L^p} \}.
\]

This corresponds to an integration by parts formula. Notice that \( \delta \) is consistent, meaning that for \( g \in L^{p_1'}(\mu) \cap L^{p_2'}(\mu) \), \( p'_i > 1 \), we have that \( \delta_{p_1}(g) = \delta_{p_2}(g) \). This is a consequence of \( \operatorname{dom}(D) \) having a fixed dense subset in all \( L^p(\mu) \)-spaces, \( p < \infty \). Also, as the adjoint of a closed operator it is directly seen that:

**Lemma 2.1.2.** \( \delta_p' \) is closed in \( L^{p'}(\mu) \), \( p' > 1 \).

Furthermore, if \( D \) has a dense image, \( \delta \) is densely defined by the Hahn Banach approximation theorem, Theorem B.3.9. We note that in the above construction for \( L^2(\mu) \) Riesz’ representation theorem, Theorem B.3.4, could have been used instead of Theorem B.3.5. See chapter 4 for worked-out statements in a specific case.

### 2.1.2 Derivative operators

Stein’s method uses derivative operators \( D \) that encode differentiability on the whole space of interest. Again, we give a brief overview of generalities and work out the statements where needed. First recall Definition B.1.5. In one dimension, differentiation is anti-dual to itself with respect to Lebesgue measure under zero-boundary conditions:

\[
\int_a^b f'(x)g(x)dx = -\int_a^b f(x)g'(x)dx.
\]

By extending this continuously in \( L^p(\mathbb{R}) \) spaces, the definition of a weak derivative may be observed. Similarly, the Jacobian matrix of functions \( f : U \subset \mathbb{R}^n \to \mathbb{R}^m \) may be considered for \( D \) in multiple dimensions. This again gives rise to the Sobolev spaces \( W^{k,p}(U) \) in case \( D \) is defined with respect to the Lebesgue measure. Namely, if \( D \) is initially defined as differentiation on \( C_0^\infty(\Omega) \), it is closable on \( L^p(U) \), \( p < \infty \). By closing the domain for \( \|\cdot\|_D = \|\cdot\|_{W^{1,p}} \), the spaces \( W^{k,p}(\Omega) \) are obtained. That \( D \) is closed will be proven in Lemma 4.2.7 for \( \mu \) a Gaussian measure. For distributions on \( U \) that are mutually absolutely continuous with respect to Lebesgue measure, the ‘tilted’ Sobolev spaces \( W^{k,p}(Z) \) are obtained.

Similarly, the \( W_0^{k,p}(U) \) arise as the closure of \( C_0^\infty(a,b) = \operatorname{dom}(D) \) for \( \|\cdot\|_{W^{1,p}} \).

Note that by closing \( D \) with respect to \( \mu \), the properties of \( D \) depend on \( \mu \). Some properties of differentiability, such as chain or product rules, only carry over to \( D \) under some extra assumptions.

Now consider again a general \( \sigma \)-finite measure space \( (\Omega, \mathcal{F}, \mu) \) and let \( \delta_p' \) be the dual of \( D : \operatorname{dom}(D) \subset L^p(\mu) \to L^{p'}(\mu) \) on \( L^{p'}(\mu) \). If (2.2) holds for all \( f \in C_0^\infty(S) \), \( g \in C_0^\infty(S) \), then \( \operatorname{dom}(\delta_p') \) also includes the closure of \( C_0^\infty(S) \) for the norm \( \|g\|_{\delta_p'} := \)
\[
\sqrt{\|g\|_{L^p}^2 + \|\delta(g)\|_{L^{p'}}^2}.
\]
We can also make this consideration for \(p' = 1\). Limits can be taken on both sides of (2.2) to infer this. We therefore similarly denote \(\text{dom}(\delta_1)\) for the \(g \in L^1(\mu)\) such that there exists a \(\delta_1(g) \in L^1(\mu)\) for which (2.2) holds. Actually, we then have for all \(p \in [1, \infty]::
\[
E[(Df)(Z)g(Z)] = E[f(Z)\delta_p'g(Z)], \quad \forall f \in W^{1,p}(Z), g \in \text{dom}(\delta_p').
\]  
(2.3)

For Stein’s method, it is convenient to know which functions are included in the domain of \(\delta\). We come back to this in particular cases.

Sturm-Liouville operators were already mentioned a few times. These actually correspond to \(L = -\delta D = -D^TD\) in \(L^2(\Omega)\), where they are defined. It is a more convenient to study symmetrized operators because they have nicer spectral properties. Symmetry means:
\[
E[Lf(Z)g(Z)] = E[f(Z)Lg(Z)],
\]
for \(f, g\) in ‘suitable domains’. In the case when \(Z\) has a density \(q\) with respect to Lebesgue measure on \((a, b)\), \(a < b\), this is a special case of Green’s formula\(^2\):
\[
\int_a^b [(Lf)(z)g(z) - f(z)(Lg)(z)]q(z)dz = [f(z)g(z)q(z)]_a^b.
\]  
(2.4)

We may obtain a similar formula for the counting measure. Here we see how boundary values come into play. The amazing thing for many Sturm Liouville operators is that the inverse is very regular where it is defined. It is often compact or even a Hilbert-Schmidt operator. This yields an orthogonal spectral decomposition of \(L^2(Z)\), which will be mentioned in chapter 4.

### 2.2 Lebesgue densities: Stein approximation

Suppose \(Z\) is a real random variable with a density \(q\) that it absolutely continuous with respect to Lebesgue measure on \((a, b)\), where \(-\infty \leq a < b \leq \infty\) such that \(q \in W^{1,1}(a, b)\). Note that discontinuities in \(q\) are not allowed, but this can be overcome (considering \(\mathcal{F}_0(q)\) and \(\mathcal{F}_1(q)\) of the following subsection). We have:
\[
\int_a^b q(z)dz = 1.
\]

We work out the duality framework of the previous section. In one dimension it is however possible to use Fubini’s theorem instead of Theorem B.3.5 or Theorem B.3.4 to get more general statements. This is often the approach in the literature. Therefore, we give a statement for extensions of \(C_0(a, b)\) and the more general \(\mathcal{F}_1(q)\).

If we want duality for all \(p \in [1, \infty]\), we could assume that \(q\) is bounded from below and above by a positive constant on every compact subinterval of \((a, b)\) and call this the **local bound condition**.

\(^2\)Which will implicitly be proven in the next section.
We now discuss some general results and heuristics. The similarity in calculations and formula’s suggests an underlying unifying theme.

A direct Stein equation

We first construct a dual operator for the negative derivative operator \( D : C_0^\infty(a, b) \rightarrow L^p(q); f \mapsto -f' \). Then by Fubini’s theorem it holds for all \( g \in C_0^\infty(a, b) \) that:

\[
\mathbb{E}[Df(Z)g(Z)] = \lim_{y \to a+} \int_{x}^{z} -f'(z)g(z)q(z)dz
\]

\[
= \lim_{x \to b-} \int_{y}^{x} -f'(z) \left((qg)(y) + \int_{y}^{z} (qg)'(z')dz'\right)
\]

\[
= \lim_{x \to b-} \left[-f(z)g(z)q(z)\right]_{y}^{x} + \int_{a}^{b} f(z)(qg)'(z)dz = \mathbb{E} \left[f(Z) \frac{(qg)'(Z)}{q} \right].
\]

We wrote the boundary values in the third line to emphasize this extra term if \( g \) would not have compact support (refer to \( F \)). Note that \( \frac{(qg)'}{q} \in L^p(q) \) for all \( p \in [1, \infty] \) under the local bound condition. \( p = 1 \) does not require additional assumptions on \( q \).

We thus find:

\[
\tilde{Z}_{Z,q'} : \text{dom}(\tilde{Z}_{Z,q'}) \subset L^p(q) \rightarrow L^p(q); g \mapsto \frac{(qg)'}{q}. \tag{2.5}
\]

The Fubini argument may also used to verify that \( \text{dom}(\tilde{Z}_{1,1}) \) includes

\[
\mathcal{F}_1(q) := \{ g \in L^1(q) | \text{gg is absolutely continuous, } (qg)' \in L^1(a,b), \lim_{x \to a+} q(x)g(x) = 0 \}. \tag{2.6}
\]

It is dense in \( L^1(q) \) since it contains \( C_0^\infty(a,b) \). To extend a Stein characterization to the largest possible class of functions, only duality with respect to constants is needed. In one dimension, Fubini’s theorem can again be used to verify that this is the case for \( g \in \mathcal{F}_0(q) \).

\[
\mathcal{F}_0(q) := \{ g : \mathbb{R} \rightarrow \mathbb{R} | \text{gg is absolutely continuous, } (qg)' \in L^1(a, b), \lim_{x \to a+} q(x)g(x) = 0 \}. \tag{2.7}
\]

Namely, we have that \( \mathbb{E}[\tilde{Z}(g)(Z)] = 0 \). Thus, for duality with respect to constants, no global integrability assumptions on \( g \) are needed.

An inversion formula for the Stein equation:

\[
(qg)'(x) = q(x) \cdot (h(x) - \mathbb{E}[h(Z)]), \quad (qg)(a) = (qg)(b) = 0, \tag{2.8}
\]

takes the form of:

\[
\tilde{g}_h(x) = \tilde{Z}^{-1}(h - \mathbb{E}[h(Z)])(x) = \frac{1}{q(x)} \int_{a}^{x} q(y)(h(y) - \mathbb{E}[h(Z)])dy \tag{2.9}
\]

\[
= -\frac{1}{q(x)} \int_{x}^{b} q(y)(h(y) - \mathbb{E}[h(Z)])dy. \tag{2.10}
\]
The zero-boundary conditions of (2.8) imply that these solutions are again included in \( \mathcal{F}_1(q) \) if \( \mathcal{H} \subset L^1(q) \) and the \( x \mapsto \int_a^x q(y)(h(y) - \mathbb{E}[h(Z)])dy \) are also integrable. Anticipating coming results, this is the case for bounded \( \mathcal{H} \) if \( Z \) has a first moment and Lipschitz functions if \( Z \) has a first moment.

Some heuristics for regularity results will now be investigated. Denote the distribution function \( F(x) = \int_a^x p(y)dy \) and the median of \( Z \) by \( m \).

\[
\begin{align*}
|h(x)| &\leq \|h - \mathbb{E}[h(Z)]\|_\infty \frac{F(x) \wedge (1 - F(x))}{q(x)}, \\
\left| \frac{1}{q(x)} \int_a^x q(y)(h(y) - \mathbb{E}[h(Z)])dy \right| &\leq \|h - \mathbb{E}[h(Z)]\|_\infty \frac{1}{q(x)} \int_a^x q(y)dy, \\
\left| -\frac{1}{q(x)} \int_x^b q(y)(h(y) - \mathbb{E}[h(Z)])dy \right| &\leq \|h - \mathbb{E}[h(Z)]\|_\infty \frac{1}{q(x)} \int_x^b q(y)dy,
\end{align*}
\]

lead to
\[
|g_h(x)| \leq \|h - \mathbb{E}[h(Z)]\|_\infty \frac{F(x) \wedge (1 - F(x))}{q(x)}. \tag{2.11}
\]

Note that
\[
M(x) = q(x)^{-1} \int_a^x q(y)\text{sgn}(m - y)dy = \delta_Z^{-1}(\text{sgn}(m - \cdot))(x) \tag{2.12}
\]
\[
= q(x)^{-1}(F(x) \wedge (1 - F(x))). \tag{2.13}
\]

So the bound is optimal. It allows to infer a Stein characterization:

**Theorem 2.2.1** (Stein characterization). Let \( Z \sim q(z)dz \) with \( q \in W^{1,1}(a,b) \) and finite first moment. If \( X \) is a real-valued random variable on \( (\Omega, \mathcal{F}, \mathbb{P}) \), then \( X \sim Z \) if and only if for all \( g \in \mathcal{F}_1(q) \) it follows that \( (qg)' / q \in L^1(X) \) and
\[
\mathbb{E}\left[\frac{(qg)'}{q}(X)\right] = 0. \tag{2.14}
\]

Here, \( (qg)' / q \) is extended by 0 outside \((a,b)\).

**Remark 2.2.2.** The space \( \mathcal{F}_1(q) \) is often not insightful. We will therefore discuss a tilt that sometimes yields more convenient Stein approximations. The class can also be enlarged to \( \mathcal{F}_0(q) \) for which a first moment is not required or reduced to another characterizing class to more easily test \( X \sim Z \).

**Proof.** Suppose \( X \sim Z \). Then (2.14) follows by duality, (2.3):
\[
\mathbb{E}\left[\frac{(qg)'}{q}(X)\right] = \mathbb{E}\left[\delta_Z'(g)(Z)\right] = \mathbb{E}[D(1)(Z)g(Z)] = 0.
\]

Conversely, suppose (2.14) holds true. We want to evaluate in \( \tilde{g}_h \), where \( h = 1_A \), \( A \) being a Borel subset of \((a,b)\). We have already noted zero-boundary conditions. We only need to check that \( \tilde{g}_h \) itself is integrable with respect to \( Z \). Because \( |\tilde{g}_h(x)| \leq
\[ \| h - E[h(Z)] \|_\infty M(x), \]
it only remains to verify that \( M \in L^1(q) \). Choose \( z \in (a, b) \). By Fubini’s theorem:
\[
\int_a^z M(y)q(y)dy \leq \int_a^z F(y)dy = \int_a^z (z - y)q(y)dy < \infty.
\]
Carrying out a similar inspection for \((z, b)\) gives that \( \tilde{g}_h \in F_1(q) \). Then:
\[
0 = E \left[ \frac{(g \tilde{g}_h)'}{q}(X) \right] = E \left[ \delta_Z (\delta_Z^{-1}(1_A - P[Z \in A])) \right] = P[X \in A] - P[Z \in A].
\]

To infer regularity for specific classes \( H \) such as \( \{1_{(-\infty,z]} \mid z \in (a, b)\} \) another representation can perhaps be more convenient:
\[
q(x) \tilde{g}_h(x) = \int_a^x q(y)h(y)dy - \int_a^x q(y)dy \left( \int_a^x q(y')h(y')dy' + \int_x^b q(y')h(y')dy' \right)
\]
\[
= (1 - F(x)) \int_a^x q(y)h(y)dy - F(x) \int_x^b q(y)h(y)dy \quad (2.15)
\]
Denoting by \( g_z \) the solution for \( h = 1_{(-\infty,z]} \), \( g_z = -g_{-z} \) may be used so that one of the above integrals can always be chosen to disappear.

**First order Stein equation**

Now assume that the distribution has a first moment. Denote the mean by \( \mu \). Then we apply integration by parts (Fubini’s theorem) to obtain an efficient representation:
\[
h(y) - E[h(N)] = \int_a^b \frac{(h(y) - h(t))q(t)dt}{= f(y)h'(z)dz}
\]
\[
= \int_a^y \int_t^y h'(z)dq(t)dt - \int_y^b \int_y^t h'(z)dq(t)dt
\]
\[
= \int_y^b h'(z)F(z)dz - \int_y^b h'(z)(1 - F(z))dz. \quad (2.16)
\]
For the Stein solution (2.9), we may then write:
\[
\tilde{\delta}_Z^{-1}(h_\perp)(x) = q(x)^{-1} \int_a^x q(y) \left( \int_y^b h'(z)F(z)dz - \int_y^b h'(z)(1 - F(z))dz \right)
\]
\[
= q(x)^{-1} \int_a^b dz \int_a^x dyq(y) \left( 1_{x < y}h'(z)F(z) - 1_{x < y}h'(z)(1 - F(z)) \right).
\]
Note that we could not split the integral and that the conditions of Fubini’s theorem are indeed satisfied since the distribution has a first moment. The inner integral equates:
\[
= \begin{cases} 
\sqrt{2\pi}(F(x) - F(z))h'(z)F(z) - \sqrt{2\pi}F(z)h'(z)(1 - F(z)), & z \leq x \\
-\sqrt{2\pi}F(x)h'(z)(1 - F(z)), & z > x \end{cases} \quad (2.17)
\]
\[
= \begin{cases} 
-\sqrt{2\pi}(1 - F(x))F(z)h'(z), & z \leq x \\
-\sqrt{2\pi}F(x)(1 - F(z))h'(z), & z > x \end{cases}. \quad (2.18)
\]
\]

32
Yielding for $\tilde{\delta}_Z^{-1}(h_{\perp})(x)$:

$$-q(x)^{-1} \left( (1 - F(x)) \int_a^x F(z)h'(z)dz + F(x) \int_x^b (1 - F(z))h'(z)dz \right). \quad (2.19)$$

Now everything within brackets is positive, except possibly $h'$. The largest value is achieved for $h'(z) = -\|h'\|_\infty$, for all $z \in (a, b)$. By replacing $h'$ by $-\|h'\|_\infty$, we can follow the procedure backwards to see:

$$\left| \tilde{\delta}_Z^{-1}(h_{\perp})(x) \right| \leq \tilde{\delta}_Z^{-1}(y \mapsto \|h'\|_\infty (\mu - y))(x)$$

$$= q(x)^{-1} \int_a^x q(y)\|h'\|_\infty (\mu - y)dy =: \|h'\|_\infty \mathcal{T}(x). \quad (2.21)$$

This defines a function $\mathcal{T} : (a, b) \to \mathbb{R}$ called the **Stein kernel**. We have proven:

**Lemma 2.2.3** (Stein kernel bound). For a random variable $Z$ with density $q$ with respect to Lebesgue measure on $(a, b)$ and having a first moment $\mu$, we have for Lipschitz $h : (a, b) \to \mathbb{R}$:

$$|g_h(x)| = \left| \frac{1}{q(x)} \int_a^x (h(y) - \mathbb{E}[h(Z)])q(y)dy \right| \leq \|h'\|_\infty \mathcal{T}(x). \quad (2.22)$$

Note that $\mathcal{T}$ is positive on $(a, b)$:

$$q(x)\mathcal{T}(x) = \int_a^x (\mu - y)q(y)dy = \int_x^b (y - \mu)q(y)dy.$$

Also, if $Z$ has a second moment Fubini’s theorem can again be used to see that $\mathcal{T} \in \mathcal{F}_1(q)$. Thus duality holds: for all $f \in W^{1,\infty}(q)$:

$$\mathbb{E}[(Z - \mu)f(Z)] = \mathbb{E}\left[\mathcal{T}(Z)f'(Z)\right]. \quad (2.23)$$

We now come back to how in one dimension Fubini’s theorem may give more general conditions for a duality formula. We can often split $(a, b)$ in regions where derivatives have a fixed sign. Without the second moment condition, it is sufficient that $\mathbb{E}[\mathcal{T}(Z)f'(Z)] < \infty$ for (2.23) to hold:

$$\int_{\mu}^b q(z)\mathcal{T}(z)f'(z)dz = \int_{\mu}^b dz f'(z) \int_z^b (y - \mu)q(y)dy$$

$$= f(\mu) \int_{\mu}^b (\mu - y)q(y)dy + \int_{\mu}^b f(y)(y - \mu)q(y)dy, \quad (2.24)$$

$$\int_a^\mu q(z)\mathcal{T}(z)f'(z)dz = \int_a^\mu dz f'(z) \int_z^\mu (\mu - y)q(y)dy$$

$$= f(\mu) \int_a^\mu (\mu - y)q(y)dy + \int_a^\mu f(y)(\mu - \mu)q(y)dy,$$

33
because on \((\mu,b)\), \(|T'| = -T'\) and the integrability condition on \(|f'(z)T'(y)|\) is fulfilled since:
\[
\int_{\mu}^{b}dz \int_{z}^{b}|f'(z)(y - \mu)q(y)|dy = \int_{\mu}^{b}q(z)T(z)|f'(z)|dz \leq \mathbb{E}[T(Z)|f'(Z)|].
\]
Similarly, \(|T'| = T'\) on \((a,\mu)\). Then sum up these equalities. Such arguments are less clear in multiple dimensions.

Furthermore, it offers a useful tilt \(g \mapsto Tg\). We define the first order Stein equation:
\[
\delta Z(g)(x) := \frac{(qTg)'}{q}(x) = T(x)g'(x) + (\mu - x)g(x) = h(x) - \mathbb{E}[h(Z)], \tag{2.26}
\]
\[
(qTg)(a) = (qTg)(b) = 0.
\]

For this operator we infer a general Stein characterization. First we inspect the derivative of the solution \(g_h\) of (2.26). By using the equation:
\[
|T(x)g'(x)| \leq \|h - \mathbb{E}[h(Z)]\|_{\infty} \left(1 + \frac{1}{T(x)q(x)} \int_{a}^{x}(\mu - x)q(y)dy\right) \\
\leq 2\|h - \mathbb{E}[h(Z)]\|_{\infty}. \tag{2.27}
\]
The last inequality follows by enlarging \((\mu - x) \leq (\mu - y)\) within the integral.

Now (2.11) gives
\[
|g_h(x)| \leq \|h - \mathbb{E}[h(Z)]\|_{\infty} \frac{M(x)}{T(x)}. \tag{2.28}
\]
Again by enlarging or reducing \(\mu - x\) to \(\mu - y\) respectively:
\[
\left(\frac{F}{qT}\right)'(x) = \frac{1}{T(x)} - \frac{(\mu - x)F(x)}{q(x)T(x)^2} \geq 0,
\]
\[
\left(\frac{1-F}{qT}\right)'(x) = -\frac{1}{T(x)} - \frac{(\mu - x)(1-F)(x)}{q(x)T(x)^2} \leq 0.
\]
Under symmetry of \(q\) around \(m\), \(M/T\) reaches its maximum in the median of the distribution. In particular, it is bounded. If the mean is known better, it may also be bounded by \((F/qT)(\mu) \lor ((1-F)/qT)(\mu)\). We then have:

Lemma 2.2.4. If \(h : (a,b) \rightarrow \mathbb{R}\) is bounded, we have under the above conditions:

1. Denoting the mean by \(\mu\) and the median by \(m\):
\[
|g_h(x)| \leq \|h - \mathbb{E}[h(Z)]\|_{\infty} \left(\frac{1}{\mu - x} \lor ((F/qT)(m) \lor ((1-F)/qT)(m))\right)
\]
\[
|g_h(x)| \leq \|h - \mathbb{E}[h(Z)]\|_{\infty}((F/qT)(\mu) \lor ((1-F)/qT)(\mu)).
\]
2. \(|T(x)g_h(x)| \leq 2\|h - \mathbb{E}[h(Z)]\|_{\infty}.
\]

34
The bound with \((\mu - x)\) can again be inferred by enlarging \((\mu - x)M(x)\) to \(T(x)\).

**Theorem 2.2.5** (One-dimensional Stein characterization). Consider \(Z \sim q(z)dz\) with finite first moment and \(q \in W^{1,1}(a, b)\). Let \(X\) be a real random variable. Then \(X\) has density \(q\) if and only if for all absolutely continuous \(g : (a, b) \to \mathbb{R}\) such that \(|T(x)g'(x)| + |(\mu - x)g(x)|\) is bounded and \(\lim_{x \to a^+, x \to b^-} qTg(x) = 0:\)

\[
E[T(X)g'(X) + (\mu - X)g(X)] = 0. 
\] (2.29)

Here, \(g\) and \(g'\) are extended by 0 outside \((a, b)\).

**Proof.** If \(h\) is bounded, then with \(g_h\) solving (2.26), \(Tg'\) and \((\mu - X)g_h(X)\) are bounded by the previous lemma. By letting \(h\) run through the indicator functions \(1_A, A \subset \mathbb{R}\) Borel, we see that (2.29) imply that \(X \sim Z\), as in the proof of the previous Stein characterization. On the other hand, under the above conditions on \(g\), \(Tg \in \mathcal{F}_1(q)\). Thus we can use the canonical Stein characterization, Theorem 2.14, to see that (2.29) holds. \(\square\)

Note that we have shown a Stein approximation for bounded \(h\), although this characterization may be better adapted to Lipschitz \(h\). We will soon illustrate how regularity for Lipschitz functions may be inferred for the normal distribution. It seems however that exploiting the particular structure of the distribution leads to better results than general reasoning. Other methods for inspecting regularity can be found in [30] or using semi-group theory as in chapters 4 and 5.

**Pearson family of distributions**

We will now focus on the Pearson family of distributions. They cover some of the classical continuous distributions, and were originally introduced with the idea of constructing probability models with a given mean, variance, skewness and kurtosis. They consist of the distributions with densities \(q\) that arise from the following differential equation with \(c_i, d_i \in \mathbb{R}:\)

\[
q'(x) = \frac{c_1 x + c_0}{d_2 x^2 + d_1 x + d_0} q(x) =: \frac{p_1(x)}{p_2(x)} q(x). 
\] (2.30)

Depending on the degree and roots of \(p_2\), it defines some structurally different solutions. Here, we consider a density \(q\) that is defined on \((a, b)\) where \(a < b\) may be \(\pm \infty\). Then it may hold for some \(c \in (a, b)\) and \(C > 0:\)

\[
q(x) = C \exp \left( \int_c^x \frac{c_1 y + c_0}{d_2 y^2 + d_1 y + d_0} dy \right).
\]

If the degree of \(p_2\) is two, denote the discriminant \(\Delta := d_1^2 - 4d_0d_2\). After appropriate rescaling, the following distinctions for probability densities on some interval \((a, b)\) can be made:
1. \(\text{deg}(p_2) = 0\): Then \(q(x) = Ce^{-(x-\mu)^2/2\sigma^2}\), which includes the normal densities.

2. \(\text{deg}(p_2) = 1\): \(q(x) = Cx^{\alpha-1}\exp(-\beta x)\), which includes the densities of the Gamma distribution.

3. \(\text{deg}(p_2) = 2, \Delta = 0\): \(q(x) = Cx^{-\alpha}\exp(-\beta/x)\).

4. \(\text{deg}(p_2) = 2, \Delta < 0\): \(q(x) = C(1 + x^2)^{-\alpha}\exp(\beta \arctan(x))\), which includes the densities of the t-distributions.

5. \(\text{deg}(p_2) = 2, \Delta > 0\): \(q(x) = Cx^{\alpha-1}(1 - x)^{\beta-1}\), which includes the densities of the Beta distributions.

\(C, \alpha, \beta\) are some suitable real constants such that \(q\) defines a density function on \((a,b)\).

We can now use the previous frameworks to devise Stein equations and assess regularity of solutions. In [24] a specific view/tilt is proposed that enables a link to an interpretation in terms of Markov processes. It coincides with the approach taken before for the first order equation. There, polynomials \(s\) and \(\tau\) of degree respectively at most 2 and 1 are put central, such that:

\[
(s(x)q(x))' = \tau(x)q(x).
\] (2.31)

In this way, \(s = p_2, \tau = p_1 + s'\). We will also assume that \((a,b)\) takes up the whole interval in between \(\pm \infty\) or the first roots of \(s\) that it can meet. By possibly changing the sign of \(s\) and \(\tau\) and imposing zero boundary conditions we then assume:

\[s(x) > 0\text{ for }a < x < b \text{ and } \lim_{x \to a+} s(x)q(x) = 0.\] (2.32)

This excludes the Cauchy distribution for example. To achieve a density function we require \(q \geq 0, \int_a^b q(x)dx = 1\). So \(\tau\) can not be a constant, because otherwise

\[
\tau = \int_a^b \tau q(x)dx = \int_a^b (sq)'(x)dx = 0.
\]

Also, \(\tau\) must be decreasing since it is linear and

\[0 \geq -s(x)q(x) = \int_x^b \tau(z)q(z)dz.
\]

We take \(Z \sim q(x)dx\). Note that (2.32) now implies that \(Z\) has a first moment (\(\tau\) can be integrated over any subinterval of \((a,b)\)). Finally, we compute:

\[
\mathbb{E}[\tau(Z)] = \int_a^b \tau(z)q(z)dz = \int_a^b (sq)'(z)dz = [sq]_a^b = 0.
\]

Thus \(\tau(z) = a(\mathbb{E}[Z] - z)\) for some \(a > 0\). For the Stein kernel we thus conclude:

\[
\mathcal{K}(x) = \frac{a}{q(x)} \int_a^x q(z)\tau(z)dz = as(x), \quad x \in (a,b).
\] (2.33)
These may have quadratic growth. So, using the Stein characterization (2.29) we see that a distribution with a first moment and density in $W^{1,1}(a,b)$ has a polynomial Stein kernel of degree at most 2 if and only if it has a Pearson distribution. After the tilt $g \mapsto sg$, we consider the first order Stein operators for the Pearson family of distributions defined by:

$$\delta_Z(g)(x) = \left(\frac{(qsg)'(x)}{q(x)}\right) = s(x)g'(x) + a(\mu - x)g(x),$$

(2.34)

with Stein solutions on $\mathcal{H}_+$:

$$g_h(x) := \delta_Z^{-1}(h - \mathbb{E}[h(Z)])(x) = \frac{1}{s(x)q(x)} \int_a^b \frac{q(y)(h(y) - \mathbb{E}[h(Z)])dy}{s(x)q(x)}.$$

(2.35)

So they coincide with the first order operators from the previous subsection. Here we outline some specific distributions:

1. **Normal distribution**: $q'(x)/q(x) = \frac{a - x}{\sigma^2}$, thus $a = 1$ and $T(x) = s(x) = \sigma^2$. They are characterized by constant Stein kernels. The Stein equation is given by:

$$\sigma^2g'(x) + (\mu - x)g(x) = h(x) - \mathbb{E}[h(Z)].$$

(2.36)

Note that Lemma 2.2.4 coincides with Lemma 1.1.1.

2. **Gamma distribution**: $G(\alpha, \beta)$: $q(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$ for $x \in (0, \infty)$ and $\alpha, \beta > 0$. $q'(x)/q(x) = (\alpha - 1 - \beta x)/x$. After computing $s(x) = x$, $\tau(x) = (\alpha - 1 - \beta x) + 1$ we see $T(x) = \beta^{-1}x$. The Stein equation is given by

$$xg'(x) + (\alpha - \beta x)g(x) = h(x) - \mathbb{E}[h(Z)].$$

(2.37)

3. **Beta distribution**: $B(\alpha, \beta)$: $q(x) = C x^{\alpha-1}(1 - x)^{\beta-1}$ for $x \in (0, 1)$.

$$\frac{q'(x)}{q(x)} = \frac{\alpha - 1 + (2 - \alpha - \beta)x}{x(1-x)}.$$  

We again compute $s(x) = x(1-x)$, $\tau(x) = \alpha + (\alpha + \beta)x$; $T(x) = (\alpha + \beta)^{-1}x(1-x)$. The Stein equation becomes:

$$x(1-x)g'(x) + (\alpha + (\alpha + \beta)x)g(x) = h(x) - \mathbb{E}[h(Z)].$$

(2.38)

4. **Student’s t-distribution**:

$$q(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi\Gamma(\nu/2)}} \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2},$$

where $x \in \mathbb{R}$ and $\nu \in \{1,2,\ldots\}$ denotes the degrees of freedom. We compute $q'(x)/q(x) = -(\nu + 1)x/(\nu + x^2)$. Thus $s(x) = (\nu + x^2)$, $\tau(x) = -(\nu - 1)x$. We
exclude \( \nu = 1 \), the Cauchy distribution, to suffice the above conditions and see that the Stein kernel \( \mathcal{T}(x) = (\nu - 1)^{-1}(\nu + x^2) \). The Stein equation becomes:

\[
(\nu + x^2)g'(x) - (\nu - 1) x g(x) = h(x) - \mathbb{E}[h(Z)].
\] (2.39)

Furthermore, Lemma 2.2.4 gives \( \|g'\|_\infty \leq 2\nu^{-1}\|h - \mathbb{E}[h(Z)]\|_\infty \).

That \( \tau \) and \( s \) are polynomials of degree 1 and at most 2 was only exploited to easily compute the Stein kernel. As a remark, these allow the existence and uniqueness of strong Itô solutions of the stochastic differential equation

\[
dX_t = a(\mu - X_t)dt + \sqrt{s(X_t)}dB_t,
\] (2.40)

where \((B_t)_{t \in [0, \infty)}\) is a Brownian motion on \([0, \infty)\). The noise \( \sqrt{s(X_t)}dB_t \) is in this case not strong enough to distort the drift \( a(\mu - X_t)dt \) towards the mean. This is convenient to represent solutions with the help of semi-group theory, see chapters 4 and 5. \(-\delta \circ D\) namely coincides with the generator of the diffusion.

**Lipschitz regularity for the normal distribution**

We end this section by iterating the earlier attempts to infer regularity of Stein solutions for the normal distribution, Lemma 1.12. The Lipschitz representation is not convenient for general bounds on derivatives however, unless the mean of the distribution coincides with the median and the density function is symmetric around this point. Still much fiddling around is needed to find the right representations for a useful bound in this case (here, the right representations will just be presented). Therefore, interpolation techniques based on semi-group theory are preferable. We however illustrate the density view and start with the general notation from before.

Using (2.16) and (2.19), we obtain from the Stein equation:

\[
\mathcal{T}(x)g'(x) = (\mu - x)g(x) + h(x) - \mathbb{E}[h(Z)]
\] (2.41)

\[
= \left[ \frac{(\mu - x)(1 - F(x))}{\mathcal{T}(x)q(x)} + 1 \right] \int_a^x h'(z)F(z)dz + \left[ \frac{(\mu - x)F(x)}{\mathcal{T}(x)q(x)} - 1 \right] \int_x^b h'(z)(1 - F(z))dz.
\] (2.42)

Integration by parts, or Fubini’s theorem gives:

\[
\int_a^x F(z)dz = - (\mu - x)F(x) + \mathcal{T}(x)q(x),
\] (2.43)

\[
\int_x^b (1 - F(z))dz = (\mu - x)(1 - F(x)) + \mathcal{T}(x)q(x).
\] (2.44)
This enables us to write (2.42) as:

\[
\mathcal{T}(x)g'(x) = \frac{1}{\mathcal{T}(x)q(x)} \int_x^b (1 - F(z))dz \int_x^b h'(z')F(z')dz' - \frac{1}{\mathcal{T}(x)q(x)} \int_x^b F(z)dz \int_x^b h'(z')(1 - F(z'))dz'.
\]

We bound this by:

\[
|\mathcal{T}(x)g'(x)| \leq 2\|h'\|_\infty \frac{1}{\mathcal{T}(x)q(x)} \int_x^b F(z)dz \int_x^b (1 - F(z'))dz'.
\]

For the standard normal distribution, this gives:

\[
\frac{|g'(x)|}{2\|h'\|_\infty} \leq \sqrt{2\pi} e^{x^2/2} \int_x^\infty \Phi(z)dz \int_x^\infty (1 - \Phi(z'))dz',
\]

where \(\Phi(t)\) again denotes the distribution function of standard normal. Note that this expression is symmetric around \(x = 0\). Therefore, we will only investigate bounds for \(x \geq 0\). By symmetry, we may rewrite it as:

\[
\frac{|g'(x)|}{2\|h'\|_\infty} \leq \sqrt{2\pi} e^{x^2/2} \left( \frac{1}{2\pi} - \left[ \int_0^x \Phi(z)dz \right]^2 \right). \quad (2.45)
\]

The term within parentheses is decreasing in \(x\) towards 0. Differentiation of the total bound yields up to a factor \(\sqrt{2\pi} e^{x^2/2}\):

\[
x \left( \frac{1}{2\pi} - \left[ \int_0^x \Phi(z)dz \right]^2 \right) - 2\Phi(x) \int_0^x \Phi(z)dz \leq x \left( \frac{1}{2\pi} - \left[ \int_0^x \Phi(z)dz \right]^2 \right) - \frac{x}{2} \leq 0,
\]

because \(\Phi(x) \geq 1/2\). Thus (2.45) attains its maximum, \((2\pi)^{-1/2}\), in \(x = 0\). This yields the claimed inequality:

\[
|g'(x)| \leq \sqrt{\frac{2}{\pi}} \|h'\|_\infty. \quad (2.46)
\]

Next, we again use Stein’s equation and the representations to infer:

\[
g''(x) = g(x) + xg'(x) + h'(x) = (1 + x^2)g(x) + x(h(x) - \mathbb{E}[h(N)]) + h'(x).
\]

\[
|g''(x)| \leq |h'(x)| + \left| -\sqrt{2\pi} e^{x^2/2}(1 + x^2)(1 - \Phi(x)) + x \int_{-\infty}^x h'(z)\Phi(z)dz \right| + \left| -\sqrt{2\pi} e^{x^2/2}(1 + x^2)\Phi(x) - x \int_x^\infty h'(z)(1 - \Phi(z))dz \right|.
\]

39
Eliminating \( h' \) as before and using again (2.43), (2.44) to both infer the sign of the terms and rewrite the equation:

\[
\frac{|g''(x)|}{\|h'\|_\infty} \leq 1 + \left[ \sqrt{2\pi e^{x^2/2}} (1 + x^2)(1 - \Phi(x)) - x \right] \left[ x\Phi(x) + \frac{e^{-x^2}}{\sqrt{2\pi}} \right] \\
+ \left[ \sqrt{2\pi e^{x^2/2}} (1 + x^2)\Phi(x) + x \right] \left[ -x(1 - \Phi(x)) + \frac{e^{-x^2}}{\sqrt{2\pi}} \right]
\]

\[= 2.\]

### 2.3 Discrete distributions

A similar program can be carried out for distributions that are absolutely continuous with respect to the counting measure on, say, \(|a, b| := \{a, a+1, \ldots, b\} \subset \mathbb{Z}\). If \(|a, b| \cap \{\pm\infty\} \neq \emptyset\), unbounded sides are considered correspondingly. We state some results, but the argumentation will be shorter due to the analogy with Lebesgue densities. Some care has to be taken with the various difference operators. There also arise shifts from \(x\) to \(x + 1\) because they are atoms of the counting measure. Let \(Z\) be a random variable with values in \(|a, b|\) such that it has a positive density \(q\) with respect to the counting measure on \(|a, b|\).

**A direct Stein equation**

As an example, we consider the dual operator of \(-\nabla\):

\[
\tilde{\delta}_Z : \text{dom}(\tilde{\delta}_Z) \subset L^2(\mathbb{Z}) \to L^2(\mathbb{Z}); g \mapsto \frac{\Delta(qg)}{q}.
\]  

(2.47)

The Stein equation

\[
\frac{\Delta(qg)}{q} = h - \mathbb{E}[h(Z)] =: h_{\perp},
\]

(2.48)

with bounded \(h\), is solved with zero-boundary conditions by:

\[
g_h(k) = \frac{1}{q(k)} \int_{[a,k)} h_{\perp}(l)q(l)dl = -\frac{1}{q(k)} \int_{[k,b]} h_{\perp}(l)q(l)dl.
\]

(2.49)

Here, zero-boundary conditions mean that \(\lim_{k \to a} g_h(k)q(k) = 0\) and if \(b = \infty\) similarly for \(b\). We can again bound it by \(|g_h(k)q(k)| \leq \|h_{\perp}\|_\infty (F(k - 1) \land (1 - F(k - 1)))\) and explore a bound for \(\Delta g_h\). The first bound is “almost” an admissible solution (because the probability mass below the median is not necessarily 1/2 now). In this way, a Stein characterization is achieved:

**Proposition 2.3.1.** Let \(X\) be a random variable with values in \(\mathbb{Z}\) Then:

\[
\mathbb{E}\left[ \frac{\Delta(qg)}{q}(X) \right] = 0, \quad \forall g \in \mathcal{F}_0(q) \iff X \sim \mathcal{Z},
\]

(2.50)
where the functions are extended by 0 outside \(|a, b|\) and

\[
\mathcal{F}_0(q) = \{g : |a, b| \to \mathbb{R} \mid \Delta(qg) \in L^1(|a, b|), qg(a) = 0\}.
\]

Here, the integrability assumption is with respect to the counting measure. The indicator functions \(g = 1_j, j \in |a, b| \setminus \{a\}\), are also sufficient if \(X\) is \(|a, b|\)-valued. Also, it is implicitly understood that the terms should be integrable with respect to \(X\).

To bound differences of solutions for \(h = 1_A\), it may again be convenient to consider (tilts of):

\[
g_h(k) = \frac{1}{q(k)} \left( (1 - F(k - 1)) \int_{(a,k]} h(l)q(l)dl - F(k - 1) \int_{[k,b]} h(l)q(l)dl \right). \tag{2.51}
\]

**A first order Stein equation**

In order to better comply results from the literature connected with a so-called generator method, we consider a different Stein equation than in section 1.2. We consider the Stein kernel for the direct equation from before, with \(\mu\) the mean of \(Z\) (we assume that \(Z\) has a first moment or a second moment for integrability of \(\mathcal{T}\)):

\[
\mathcal{T}(k) = \frac{1}{q(k)} \int_{(a,k]} (\mu - l)q(l)dl = -\frac{1}{q(k)} \int_{[k,b]} (\mu - l)q(l)dl. \tag{2.52}
\]

Then we consider the **first order Stein equation**:

\[
\delta_Z(g)(k) := \mathcal{T}(k) \nabla g(k) + (\mu - k)g(k) = \frac{\nabla(\mathcal{T} q_+g)(k)}{q(k)} = h(k) - \mathbb{E}[h(Z)], \tag{2.53}
\]

where we set \(\mathcal{T}_+(k) = \mathcal{T}(k + 1)\) and \(q_+(k) = q(k + 1)\) to obtain a preferable result. The solutions may be read off:

\[
g_h(k) = \frac{1}{(q\mathcal{T})(k + 1)} \int_{[a,k]} h_\perp(l)q(l)dl = -\frac{1}{(q\mathcal{T})(k + 1)} \int_{[k,b]} h_\perp(l)q(l)dl. \tag{2.54}
\]

Let \(m\) be the median of \(Z\). Then we have for non-constant \(h\):

\[
\frac{|g_h(k)|}{\|h - \mathbb{E}[h(Z)]\|_\infty} \leq \frac{1}{|\mu - k|} \wedge \left( \frac{F([m]) \vee (1 - F([m]))}{(q\mathcal{T})([m] + 1)} \right), \tag{2.55}
\]

\[
\frac{|g_h(k)|}{\|h - \mathbb{E}[h(Z)]\|_\infty} \leq \frac{F([\mu]) \vee (1 - F([\mu]))}{(q\mathcal{T})([\mu] + 1)}, \tag{2.56}
\]

where \([\cdot]\) denotes the floor function. For the difference operator, it may be computed that:

\[
\frac{|\nabla g_h(k)|}{\|h - \mathbb{E}[h(Z)]\|_\infty} \leq \begin{cases} \frac{1}{\mu - a}, & \text{if } k = a \\ \frac{2}{q(k)}, & \text{if } k \neq a \end{cases}, \tag{2.57}
\]

by using the first order Stein equation. This yields a Stein characterization:
Proposition 2.3.2. Let $X$ be a random variable with values in $\mathbb{Z}$. Then:

$$
\mathbb{E} \left[ T(X) \nabla g(X) + (\mu - X)g(X) \right] = 0, \quad \forall g \in \mathcal{G} \iff X \sim \mathcal{Z},
$$

(2.58)

where $\mathcal{G} = \{ g : [a, b] \to \mathbb{R} \mid \| \nabla g \|_{\infty} + \| (\mu - k) g \|_{\infty} < \infty \}$ and functions are extended by 0 outside $[a, b]$.

It may be difficult to compute the Stein kernel however. This is not the case for Ord’s family of distributions, which is analogous to Pearson’s family.

Ord’s family of distributions

Ord’s family of discrete distributions have densities that satisfy for $c_i, d_i \in \mathbb{R}$:

$$
\frac{\Delta q(k)}{q(k)} = \frac{c_1 k + c_0}{d_2 k^2 + d_1 k + d_0} =: \frac{p_1(k)}{p_2(k)}, \quad k \in [a, b],
$$

(2.59)

so that $p_2(k) \neq 0$ for $k \in [a, b]$. Distinctions are again made based on the behavior of the denominator. It can be reformulated to:

$$
\Delta(sq)(k) = \tau(k)q(k), \quad k \in [a, b].
$$

(2.60)

It may be seen that $s(k) = p_2(k - 1)$, $\tau(k) = p_1(k) + \Delta s(k)$. However, it is often more convenient to infer these quantities from:

$$
\frac{q(k + 1)}{q(k)} = \frac{s(k) + \tau(k)}{s(k + 1)}.
$$

(2.61)

We assume again that the distribution has a second moment, mean $\mu$,

$$
s(a) = 0 \text{ if } a \text{ is finite, and } s(x) > 0, \quad a < k \leq b.
$$

(2.62)

Then it may be inferred analogously that $\tau$ is a positive multiple of $(\mu - k)$ and $s$ is a multiple of $T$. We can use the reasoning of the first order Stein equation now. Some examples are:

1. **Poisson distribution:** $q(k) = e^{-\lambda} \lambda^k / k!$ for $k \in \{0, 1, 2, \ldots \}$ and $\lambda > 0$ gives $s(k) = k$, $\tau(k) = \lambda - k$. So it is the analog of the exponential distribution. The Stein equation now takes a slightly different form than before:

$$
k\nabla g(k) + (\lambda - k)g(k) = h(k) - \mathbb{E}[h(Z)].
$$

(2.63)

2. **Binomial distribution:** $q(k) = \binom{n}{k} p^k (1 - p)^{n-k}$, with $n \in \mathbb{N}$, $p \in (0, 1)$ and $k \in \{0, 1, \ldots, n\}$. Using (2.61), it may be computed that $s(k) = (1 - p)k$, $\tau(k) = pn - k$. So it resembles the Poisson distribution, but restricted to a bounded domain and with a different form for the parameters. The Stein equation becomes:

$$
(1 - p)k\nabla g(k) + (pn - k)g(k) = h(k) - \mathbb{E}[h(Z)].
$$

(2.64)
3. **Negative binomial distribution**: \( q(k) = \binom{k+r-1}{k} (1-p)^r p^k \), where \( r > 0, p \in (0, 1) \) and \( k \in \{0, 1, \ldots\} \). Then \( s(k) = k, \tau(k) = pr - (1-p)k \), again having a similar structure as the Poisson distribution, but with the parameters attaining different values. The Stein equation namely is:

\[
k \nabla g(k) + [pr - (1-p)k] g(k) = h(k) - \mathbb{E}[h(Z)]. \tag{2.65}
\]

Note that it can not coincide with the Poisson or binomial characterization. The first could be reached with \( p \to 0, pr/(1-p) \to \lambda > 0 \) however. There is corresponding distributional convergence.

4. **Hypergeometric distribution**:

\[
q(k) = \binom{n}{k} \binom{\beta}{n-k} \binom{\alpha}{n-k}^{-1},
\]

where \( n \in \mathbb{Z}^+, \alpha \geq n, \beta > n, k \in \{0, \ldots, n\} \). Then \( s(k) = k(\beta - n + k), \tau(k) = \alpha n - (\alpha + \beta)k \), yielding the Stein equation:

\[
k(\beta - n + k) \nabla g(k) + [\alpha n - (\alpha + \beta)k] g(k) = h(k) - \mathbb{E}[h(Z)]. \tag{2.66}
\]

This resembles the Beta distribution.

In the continuous case, it was mentioned that the characterizations coincide with the generators of Itô-diffusions. The reason for listing the slightly different Stein equation (2.53) is that that the operator also coincides with the generator of a Markov process after symmetrization \((-\delta \circ D)\). It is a birth-death process in this case. See [24] for more information.

**Regularity for Poisson approximation**

This small subsection has as a goal a proof of the second part of Lemma 1.2.2. We use a tilt of (2.51), with \( k > 0 \):

\[
g_A(k) = \frac{(k-1)!}{\lambda^k} \left( (1 - F(k-1)) \int_{[n,k]} \mathbb{1}_A \frac{\lambda^l}{l!} dl - F(k-1) \int_{[k]} \mathbb{1}_A \frac{\lambda^l}{l!} dl \right). \tag{2.67}
\]

Suppose \( A = \{j\}, j \in \mathbb{Z}^+ \) and denote the solution \( g_j \) correspondingly. Then:

\[
g_j(k) = \begin{cases} 
-F(k-1) \frac{(k-1)!}{\lambda^k} \frac{\lambda^j}{j!} = -\frac{\lambda^j}{j!} \sum_{l=0}^{k-1} \frac{\lambda^{l+1}(k-1)!}{(l+1)(l-k)!} & \text{if } k \leq j \\
(1-F(k-1)) \frac{(k-1)!}{\lambda^k} \frac{\lambda^j}{j!} = \frac{\lambda^j}{j!} \sum_{l=0}^{\infty} \frac{\lambda^{l}(k-1)!}{(l+k)!} & \text{if } k > j 
\end{cases}
\]

So \( \triangle g_j(k) \leq 0 \) for \( k \neq j \). For \( k = j \):

\[
\triangle g_j(j) = e^{-\lambda} \left( \frac{\lambda}{j} \sum_{l=0}^{j-1} \frac{\lambda^l}{l!} + \sum_{l=j+1}^{\infty} \frac{\lambda^l}{l!} \right) \leq e^{-\lambda}(e^\lambda - 1).
\]
We may then write by linearity of solutions:

\[ \Delta g_A(k) = \sum_{j \in A} \Delta g_j(k). \]

All terms on the right hand side are negative, except for possibly one (\( \Delta g_k(k) \) if present) for which the bound of interest holds. So we obtain \( \Delta g_A(k) \leq \frac{1-e^{-\lambda}}{\lambda} \). The bound on the absolute value is then obtained from \(-\Delta g_A(k) = \Delta g_{A'}(k) \leq \frac{1-e^{-\lambda}}{\lambda} \).

### 2.4 Zero-bias coupling and Lindeberg’s CLT

To bound the Stein functional \( S_X \), it is also possible to use distributional transforms\(^3\). Size-bias couplings, zero-bias-couplings and Stein-pairs have been studied in the literature, see [23] for a summary. We will discuss and demonstrate the zero-bias coupling on Lindeberg’s central limit theorem.

Consider a random variable \( X \) with \( \mu := \mathbb{E}[X], \text{var}(X) = \sigma^2 \). Investigating the Stein equation (1.1.8) of the normal distributions, it would be convenient to have a random variable \( X^z \) that has the following property:

\[ \mathbb{E}[(X - \mu)g(X)] = \sigma^2 \mathbb{E}[g'(X^z)], \]

for all Lipschitz \( g : \mathbb{R} \to \mathbb{R} \). At this point the reader needs to be cautious, keeping Remark 1.1.10 in mind. If this property holds, then \( X^z \) is said to have the zero-bias distribution of \( X \). If \( X^z \) and \( X \) exist on the same probability space, then \( X^z \) is said to be a zero-bias coupling of \( X \). In this case, it would imply the following bound for the Stein functional of the standard normal distribution:

\[ |\mathbb{E}[g'(X) - Xg(X)]| = |\mathbb{E}[g'(X)] - \mathbb{E}[g'(X^z)]|. \]

To couple \((X, X^z)\) in the most efficient way, inspiration can be drawn from optimal transport problems, section 1.3. For example, the monotone coupling on the line gives the optimal coupling between the two random variables for the Wasserstein distance. For unknown distributions, coupling may be difficult. Coupling methods are somehow “anti-Stein” in nature since Stein’s method is often used because it is difficult to couple \( X \) and \( Z \) efficiently. Nevertheless, distributional transforms and couplings can yield useful results.

Of course, the above formula coincides with the duality relation of the Stein kernel. Distributional existence and uniqueness holds:

**Proposition 2.4.1** (Zero-bias distribution). Let \( X, Y \) be real random variables on \((\Omega, \mathcal{F}, \mathbb{P})\) with \( \mu = \mathbb{E}[X], \text{var}(X) = \sigma^2 < \infty \) and Stein kernel \( T \).

\(^3\)Transformations of probability spaces are widely applied, such as in Girsanov’s theorem.
1. If $W$ is a random variable that is absolutely continuous with respect to Lebesgue measure with density
\[ q^2(x) = \sigma^{-2}T(x), \tag{2.68} \]
then it has the size-bias distribution of $W$.

2. If $E[(X - \mu)g(X)] = \sigma^2E[g'(Y)]$ holds for $g$ in a class of functions $G$ such that the $g'$ are bounded and measure determining, then $X^z$ must have the distribution $p^x(x)dx$.

**Proof.** It was noted in section 2.1 that the Stein kernel and thus $p^x$ is positive. Because $X$ has a second moment, duality can be seen to hold ($\delta$ is dual to minus differentiation):
\[ E[(X - \mu)g(X)] = E[T(X)g'(X)]. \]

First take $g(x) = (x - \mu)$ to see that $p^x$ integrates to 1. Thus it defines a probability distribution on $\mathbb{R}$. So, if $X^z$ has distribution $p^x(x)dx$, it has the zero-bias distribution of $X$.

We are now able to show the central limit theorem (CLT) of Lindeberg-Feller. It can for example be used in a construction of Brownian motion. This proof is due to Goldstein, [12]. It is a CLT for standardized sums of random variables that are independent, but not necessarily identically distributed. Consider a triangular array of real random variables $X = (X_{i,n})_{n \in \mathbb{N}, i=1,...,n}$ with $E[X_{i,n}] = 0$ and $\text{var}(X_{i,n}) = \sigma^2_{i,n}$. Triangular just means that at every time step $n$, we are considering $n$ random variables. Assume that it is such that $W_n = \sum_{i=1}^n X_{i,n}$ has $\text{var}(W_n) = 1$. Then it is said that the Lindeberg condition holds iff for all $\epsilon > 0$ and $n \to \infty$:
\[ \sum_{i=1}^n E[X_{i,n}^2 \mathbb{1}_{|X_{i,n}| > \epsilon}] \to 0. \tag{LC} \]

It is related to the condition that every term should become small enough uniformly. Actually, when $\max_{1 \leq i \leq n} \sigma^2_{i,n} \to 0$, (LC) is also necessary for the following theorem, see [15]. When it does not, $X_{i,n} \sim \delta_{i,1}N(0,1)$ shows that it is not necessary.

**Theorem 2.4.2** (CLT of Lindeberg-Feller). If (LC) holds, then $W_n \to N(0,1)$ in distribution.

(LC) neatly translates into a property of the zero-bias distribution. Consider a possibly enlarged probability space where $X$ and $X^z_{i,n}$ are jointly defined, where $X^z_{i,n}$ has the zero-bias distribution of $X_{i,n}$ and is independent of the other variables (this can be done by considering product measures). Also suppose that there are random variables $I_n, n \in \mathbb{N}$, independent of $X$ and the $X^z_{i,n}$ such that $P[I_n = i] = \sigma^2_{i,n}$ for
Then $W_n^* := \sum_{j \neq i} X_{j,n} + X_{i,n}^*$ has the zero-bias distribution of $W_n$: take $g : \mathbb{R} \to \mathbb{R}$ with bounded derivative, then by independence (condition the expectation on the remaining variables):

$$
\mathbb{E}[W_n g(W_n)] = \sum_{i=1}^{n} \mathbb{E}[X_{i,n} g(W_n - X_{i,n} + X_{i,n}^*)] = \sum_{i=1}^{n} \sigma_{i,n}^2 \mathbb{E}[g'(W_n - X_{i,n} + X_{i,n}^*)] = \mathbb{E}[g'(W_n - X_{i,n} + X_{i,n}^*)] = \mathbb{E}[g'(W_n^*)].
$$

**Lemma 2.4.3.** $(LC)$ holds if and only if $X_{i,n}^* \to 0$ in probability for $n \to \infty$.

**Proof.** We unravel definitions, with $\epsilon > 0$:

$$
P[|X_{i,n}^*| > \epsilon] = \sum_{i=1}^{n} \sigma_{i,n}^2 P[|X_{i,n}^*| > \epsilon] = \sum_{i=1}^{n} \mathbb{E}[X_{i,n}(X_{i,n} - \epsilon \text{sgn}(X_{i,n})) 1_{|X_{i,n}| > \epsilon}].
$$

Because

$$
\frac{x^2}{2} 1_{|x| > 2\epsilon} \leq (x^2 - \epsilon|x|) 1_{|x| > \epsilon} \leq x^2 1_{|x| > \epsilon},
$$

the equivalence follows from:

$$
\sum_{i=1}^{n} \mathbb{E} \left[ \frac{X_{i,n}^2}{2} 1_{X_{i,n} > 2\epsilon} \right] \leq P[|X_{i,n}^*| > \epsilon] \leq \sum_{i=1}^{n} \mathbb{E} \left[ X_{i,n}^2 1_{X_{i,n} > \epsilon} \right].
$$

With this translation the CLT follows quickly after noting that $|W_n^* - W_n| = |X_{i,n}^* - X_{i,n}|$:

**Proof of Theorem 2.4.2.** By (1.16), Lemma 1.12 and the definition of the zero-bias distribution, it is sufficient to show that for $n \to \infty$:

$$
|\mathbb{E}[g'(W_n)] - g'(W_n^*)| \to 0,
$$

for all twice continuously differentiable $g$ with bounded derivatives. Compute when $g'' \neq 0$:

$$
|\mathbb{E}[g'(W_n)] - g'(W_n^*)| \leq \int_0^{2\|g''\|_{\infty}} P[|g'(W_n) - g'(W_n^*)| > t] \, dt \\
\leq \int_0^{2\|g''\|_{\infty}} P[|W_n - W_n^*| > t/\|g''\|_{\infty}] \, dt.
$$

(2.69)
Then by dominated convergence, it is sufficient to show that $|W_n - W_n| = |X_{I_{n,n}} - X_{I_{n,n}}|$ converges to 0 in probability. We already have that $X_{I_{n,n}} \to 0$ in probability by the previous Lemma. Then we investigate:

$$
P[X_{I_{n,n}} > \epsilon] \leq \frac{\text{var}(X_{I_{n,n}})}{\epsilon^2} = 1 \epsilon^2 \sum_{i=1}^n \sigma_{i,n}^4 \leq \frac{\max_{1 \leq i \leq n} \sigma_{i,n}^2}{\epsilon^2},$$

where we have used that $\sum_{i=1}^n \sigma_{i,n}^2 = 1$. For any $\delta > 0$ we infer:

$$
\sigma_{i,n}^2 = \mathbb{E}[X_{i,n}^2 \mathbb{1}_{|X_{i,n}| \leq \delta}] + \mathbb{E}[X_{i,n}^2 \mathbb{1}_{|X_{i,n}| > \delta}].
$$

The first term is bounded by $\delta^2$ and the second converges to 0 for $n \to \infty$, uniformly in $i$ by $(LC)$. Thus, $\max_{1 \leq i \leq n} \sigma_{i,n}^2$ can be made arbitrary small by taking $n$ sufficiently large. We can hence conclude the argument because $X_{I_{n,n}} \to 0$ in probability.

2.5 Comparing nested densities

Consider real-valued random variables $X_1, X_2$ having densities $q_i, i = 1, 2$ respectively with respect to Lebesgue measure and support $I_i := \{q_i > 0\}$. To avoid too much technicalities, we assume:

1. $I_i, i = 1, 2$ are (possibly unbounded) intervals with interior $I_i^0 = (a_i, b_i)$ and $I_2 \subset I_1$.

2. $q_i|_{I_i} \in W^{1,1}(I_i^0)$.

3. $X_1$ and $X_2$ have first moments $\mu_i, i = 1, 2$ respectively.

4. $\pi_0$ defined from $q_2 = \pi_0 q_1$ is assumed to be absolutely continuous on $I_2^0$.

1. expresses that $q_1$ and $q_2$ are nested densities, i.e. having nested supports. 2. is the regularity condition in which the density view was discussed. We extend $\pi_0$ by 0 outside $I_2^0$. Furthermore, we denote the respective direct Stein operators on $L^1(q_i)$ by $\tilde{\delta}_i, i = 1, 2$, $\tilde{\delta}_i(g) = (q_i g)' / q_i$. We have for $g \in \text{dom}(\tilde{\delta}_1) \cap \text{dom}(\tilde{\delta}_2)$:

$$
\tilde{\delta}_2(g) = \frac{(g \pi_0 q_1)'}{\pi_0 q_1} = \tilde{\delta}_1(g) + \frac{\pi_0'}{\pi_0} g = \tilde{\delta}_1(g) + (\log \pi_0)' g. \quad (2.70)
$$

Note that now the index refers to the variable and not the $L^p'$-space. (2.70) can also be extended to absolutely continuous $g$ due to the explicit formula for $\tilde{\delta}$ by the product rule (Fubini’s theorem).

Let $T_i$ denote the Stein kernel of $X_i$ and $M_i = (q_i)^{-1}(F_i \land (1 - F_i))$, where $F_i$ is the distribution function of $X_i$. Recall the definition (2.7) of $\mathcal{F}_0(q)$. For a function class $\mathcal{H}, \mathcal{H}_\perp$ denotes the centered class for $X_1$. We state the main result from [17] in our context:
Lemma 2.5.1. 1. Assume that $\mathbb{E}|T_1(X_1)\pi'_0(X_1)| < \infty$. Let $\mathcal{H}$ denote the Lipschitz functions $\mathbb{R} \to \mathbb{R}$ with Lipschitz constant at most 1. If $\tilde{\delta}_1^{-1}(\mathcal{H}_\perp) \subset \mathcal{F}_0(q_2)$, we have:

$$\mathbb{E}|\pi'_0(X_1)T_1(X_1)| \leq d_W(X_1, X_2) \leq \mathbb{E}||\pi'_0(X_1)T_1(X_1)|.$$

(2.71)

2. Let $\mathcal{H}' := \{1_A | A \subset \mathbb{R} \text{ Borel}\}$. If $\tilde{\delta}_1^{-1}(\mathcal{H}'_\perp) \subset \mathcal{F}_0(q_2)$, we have:

$$d_{TV}(X_1, X_2) \leq \mathbb{E}||\pi'_0(X_1)|M_1(X_1)|.$$

(2.72)

Proof. We first prove 1. Let $h \in \mathcal{H}$ and denote $g_h = \tilde{\delta}_1^{-1}(h - \mathbb{E}[h(X_1)])$. Starting as usual we infer:

$$\mathbb{E}[h(X_2)] - \mathbb{E}[h(X_1)] = \mathbb{E}[\tilde{\delta}_1(g_h)(X_2)] = \mathbb{E}[\tilde{\delta}_2(g_h)(X_2)] - \mathbb{E}[(\log \pi_0)'(X_2)g_h(X_2)]$$

$$= -\mathbb{E}[(\log \pi_0)'(X_2)g_h(X_2)].$$

(2.73)

We have used that $g_h \in \mathcal{F}_0(q_2)$, so that the expectation of $\tilde{\delta}_2(g_h)$ under $X_2$ vanishes. Then we can use Lemma 2.2.3 to infer the upper bound:

$$|\mathbb{E}[h(X_2)] - \mathbb{E}[h(X_1)]| \leq \mathbb{E}||(\log \pi_0)'(X_2)|T_1(X_2)| = \mathbb{E}||\pi_0'(X_1)|T_1(X_1)|.$$

(2.74)

Because $\mathbb{E}|T_1(X_1)\pi'_0(X_1)| < \infty$, duality also applies as in (2.24):

$$d_W(X_1, X_2) \geq |\mathbb{E}[X_2 - \mathbb{E}[X_1]]| = \left|\mathbb{E} \left[ \tilde{\delta}_1(T_1)(X_2) \right] \right|$$

$$= \left|\mathbb{E} \left[ \tilde{\delta}_1(T_1)(X_1)\pi_0(X_1) \right] \right| = \left|\mathbb{E} \left[ T_1(X_1)\pi'_0(X_1) \right] \right|.$$ 

For 2., we can repeat (2.73) with $h = 1_A$, but then bound by (2.11):

$$|\mathbb{E}[h(X_2)] - \mathbb{E}[h(X_1)]| \leq \mathbb{E}||(\log \pi_0)'(X_2)|M_1(X_2)| = \mathbb{E}||\pi'_0(X_1)|M_1(X_1)|.$$

\[
\square
\]

The upper bound is promising since we compare the Stein solutions to their best upper bound. The lower bound usually is more directly computed as $|\mu_1 - \mu_2|$. However, in the above form, computations for the upper bound can be reused for the lower bound. Also recall that $M_1/T_1$ can be bounded as in Lemma 2.2.4. So, for part 2., there is also a bound in terms of the Stein kernel.

The assumption on the domain of $\tilde{\delta}_2$ can be verified in the following way:

Lemma 2.5.2. Suppose that:

1. $\lim_{x \to a_2, x \to b_2} q_2(x)T_1(x) = 0$.

2. $\pi'_0 q_1 T_1 = q'_2 T_1 - \frac{q'_1}{q_1} T_1 q_2 \in L^1(\mathcal{I}_1)$.

Then $\tilde{\delta}_1^{-1}(\mathcal{H}_\perp) \subset \mathcal{F}_0(q_2)$ where $\mathcal{H}$ are the Lipschitz functions with Lipschitz constant at most 1.
Proof. Because for \(g_h := \tilde{\delta}^{-1}(h - \mathbb{E}[h(Z)])\), \(h \in \mathcal{H}\), \(|q_2(x)g_h(x)| \leq q_2(x)\mathcal{T}_1(x)\), the zero-boundary conditions are fulfilled by assumption 1. Note that \(q_1\mathcal{T}_1\) is bounded (inspect the sign of the derivative to see that it reaches its maximum in \(\mu_1\)). So by Fubini’s theorem:

\[
(q_2g_h)' = (\pi_0q_1g_h)' = \pi_0'q_1g_h + \pi_0q_1(h - \mathbb{E}[h(X_1)]),
\]

because \(|\pi_0'q_1g_h| \leq \pi_0'q_1\mathcal{T}_1 \in L^1(\mathcal{I}_1)\) and the second term also is in \(L^1(\mathcal{I}_1)\) because the integral of its absolute value equals \(\mathbb{E}|h(X_2) - \mathbb{E}[h(X_1)]| < \infty\).

An analogous argument yields:

**Lemma 2.5.3.** Suppose that:

1. \(\lim_{x \to a^+, x \to b^-} q_2(x)M_1(x) = 0\).
2. \(\pi_0'q_1M_1 = q_2'\mathcal{M}_1 - \frac{q_1'}{q_1}M_1q_2 \in L^1(\mathcal{I}_1)\).

Then \(\tilde{\delta}^{-1}(\mathcal{H}'_x) \subset \mathcal{F}_0(q_2)\) where \(\mathcal{H}'\) are the indicator functions of Borel \(A \subset \mathbb{R}\).

The lower bound \(|\mu_2 - \mu_1|\) is ‘a’ choice for \(h\). In this context, recall Lemma 1.3.11. Using the lemma, it may sometimes be verified when it is optimal.

For the Pearson family of distributions, the Stein kernel was readily computed. This is not always possible. In this case, we can also use the results of the previous section to obtain:

\[
d_W(X_1, X_2) \leq \|\pi_0'\|_\infty \text{var}(X_1).
\]

(2.75)

In many cases, \(\pi_0'\) is not bounded however.
Chapter 3

Stein’s method in statistics

This chapter outlines two examples of how the methods of the previous section can be used in statistics. First, different priors are compared in Bayesian statistics by the theory on nested densities of last chapter. Next, it is illustrated how Stein’s method may be used to obtain bounds on the Wasserstein distance for the maximal likelihood estimator of a one-dimensional parameter.

3.1 Comparing priors in Bayesian statistics

In Bayesian statistics, probabilities express a belief on the occurrence of events. When formulating probabilistic parametric models to study random variables \(X_i, i = 1, \ldots, n\), the parameters \(\theta\) of the model are treated as random as well. As an illustration of the results on nested densities, we are only concerned with one-dimensional distributions on \(\theta\) that are absolutely continuous with respect to Lebesgue measure. We let \(\theta, X_1, \ldots, X_n \in \mathbb{R}\). A prior belief on the distribution of \(\theta\) is expressed in a density \(\pi\) on \((a,b) \subset \mathbb{R}\). If the proposed model for the distribution of the \((X_i)_{i=1,\ldots,n}\) is given by \(f(x|\theta)dx\), then the posterior distribution of \(\theta\) given the data \(x \in \mathbb{R}^n\) is computed using Bayes’ rule:

\[
q_2(\theta|x) = \kappa_2(x)\pi(\theta)f(x|\theta).
\]

where \(\kappa_2(x)\) normalizes the density. We would like to investigate how different choices for a prior density influence the model for large \(n\). In particular, do they approach each other? Here, we will compare the prior \(\pi\) to the possibly improper uniform prior on \((a,b)\) for \(\theta\). This means we assume that we obtain a proper distribution from:

\[
q_1(\theta|x) = \kappa_1(x)f(x|\theta).
\]

And so:

\[
q_2(\theta|x) = \pi_0(\theta|x)q_1(\theta|x) = \kappa_3(x)\pi(\theta)q_1(\theta|x),
\]

Comparing to other priors can be done analogously or using the triangle inequality.

Further note that with \(\Theta_i \sim q_i(\theta|x)d\theta\):
\[ 1 = \int q_2(\theta|x) d\theta = \kappa_3(x) E[\pi(\Theta_1)]. \]

Our goal is to estimate the distance between \( q_1 \) and \( q_2 \) under consistent data \( x \in \mathbb{R}^n \), namely \( d_W(q_1(\cdot|x) d\theta, q_2(\cdot|x) d\theta) \). In order to give probabilistic interpretations, we now assume the \( \Theta_i \) to follow the respective posterior distributions given \( X := (X_i)_{i=1,\ldots,n} \). However, \( X \) is assumed to follow a fixed probability distribution that does not necessarily arise from one of the conditional models (because they might give inconsistent marginals). We have just motivated the forms of the \( q_i \) from a Bayesian perspective.

Suppose the conditions of Lemma 2.5.1, point 1, are met. Specializing (2.71) to this setting we obtain with \( T_1 \) the Stein kernel of \( \Theta_1 \):

\[
\frac{|E[\pi'(\Theta_1)|T_1(\Theta_1|x)]|}{E[\pi(\Theta_1)]} \leq d_W(P_1, P_2) \leq \frac{E[|\pi'(\Theta_1)|T_1(\Theta_1|x)]}{E[\pi(\Theta_1)]}. \tag{3.4}
\]

But as is usual in Bayesian statistics, evaluation problems arise due to intractability of normalizing constants. So we reformulate:

\[
|E[\rho(\Theta_2)|T_1(\Theta_2|x)]| \leq d_W(P_1, P_2) \leq E[|\rho(\Theta_2)|T_1(\Theta_2|x)], \tag{3.5}
\]

with \( \rho(\theta) = (\log \pi)' = \pi'(\theta)/\pi(\theta) \).

### 3.1.1 Normal model

Consider a normal model, where data \((X_i)_{i=1,\ldots,n}\) stem from i.i.d. \( N(\theta, \sigma^2) \)-distributed random variables. We assume the parameter \( \theta \) to have a differentiable prior density \( \pi > 0 \), giving rise to the density function \( q_2 \). Take \( \Theta_2 \sim q_2(\theta|x) d\theta \). We assume that the data allow to consider a well-defined posterior model based on the uniform prior \( \tilde{\pi} \equiv 1 \). The posterior \( q_1 \) based on a uniform prior is then as a function of \( \theta \) proportional to:

\[
q_1(\theta|x) \propto f(x; \theta) = \exp \left( -\frac{1}{2} \sum_{i=1}^{n} \frac{(x_i - \theta)^2}{\sigma^2} \right) \propto \exp \left( -\frac{1}{2} \frac{(\theta - \bar{x})^2}{\sigma^2/n} \right),
\]

where \( \bar{x} := n^{-1} \sum_{i=1}^{n} x_i \). Take \( \Theta_1 \sim q_1(\theta|x) d\theta \sim N(\bar{x}, \sigma^2/n) \). If the mild assumptions of Lemma 2.5.1 are fulfilled, the following bounds can be inferred:

\[
\frac{\sigma^2}{n} E[|\rho(\Theta_2)|] \leq d_W(\Theta_1, \Theta_2) \leq \frac{\sigma^2}{n} E[|\rho(\Theta_2)|]. \tag{3.6}
\]

If for example \( E[|\rho_0(\Theta_2)||Y_1, \ldots, Y_n] \) is \( o_p(n) \) (stochastically smaller than \( n \)), then with probability arbitrarily close to 1, we eventually get posterior models which are close to each other for the Wasserstein distance.

This illustrates how different prior models models may be compared qualitatively. We now come to an illustration of explicit bounds.
3.1.2 Binomial model

Here we study posteriors arising under the study of a binomial model, i.e. for $X$ distributed with density with respect to the counting measure:

$$f(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}, \ x = 0, \ldots, n.$$ 

Priors for $\theta$ are often chosen as Beta distributions $B(\theta|\alpha, \beta), \alpha > 0, \beta > 0$ because they allow straightforward computation of the posteriors. They are so called conjugate priors. Using (3.1), they give rise to posterior distributions with density:

$$q_{\alpha,\beta}(\theta|x) = B(\theta|x + \alpha, n - x + \beta). \ (3.7)$$

When it is desirable to express little knowledge about the parameter before any data is observed, one wants the prior to be as uninformative as possible. Therefore, some effort has been put in finding such priors. We focus our attention particularly on three proposals: the Bayes, Jeffreys and Haldane priors. They correspond to $B(\theta|1,1)$, $B(\theta|1/2,1/2)$ and the improper “$B(\theta|0,0)$” respectively. The Bayes prior expresses a uniform distribution on $(0,1)$ and the Haldane prior a distribution that does not “regularize” (3.7), i.e. the parameters are the ones that directly stem from the model $f$. Naïvely these would seem good as indicating no prior knowledge. They however give rise to different models under different parametrizations $\psi \mapsto \theta(\psi)$. Jeffreys prior is a density that is chosen to be independent of the parametrization. Given the model $f(x|\theta)$, it is chosen proportional to $\sqrt{I(\theta)}$. Here, $I(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial \theta}\log f(X;\theta)\right)^2\right|\theta]$ is the Fisher information of the proposed model. For clarity, the posterior distributions based on Bayes’, Jeffreys’ and Haldane’s prior are explicitly stated respectively as:

$$q_B(\theta|x) = B(\theta|x + 1, n - x + 1) = q_1(\theta|x),$$
$$q_J(\theta|x) = B(\theta|x + 1/2, n - x + 1/2),$$
$$q_H(\theta|x) = B(\theta|x, n - x).$$

Take $\Theta_i \sim q_i(\theta|x)d\theta$, $i = B, J, H$. These expressions already yield much information on weak convergence and intuition about distance. Lemma 1.3.11 actually enables to compute the Wasserstein distances explicitly or approximate it well numerically, along the same line of reasoning for Proposition 1.3.12. Comparing two of the three densities, it can be seen that they have two crossing points that arise from an equation $\theta(1-\theta) = C$ for some constant $C$ (because the quotient tends to zero or infinity at the border of $(0,1)$ and a density can not be larger than the other on $(0,1)$). Note that we use that both parameters of the beta distribution are regularized by the same increment. Thus the Wasserstein distance between any two of the above distributions is of the form

$$d_W(\Theta_i, \Theta_j) = |\mathbb{E}[\Theta_i - a] - \mathbb{E}[\Theta_j - a]|,$$
where \( a \in (0, 1) \) is determined by the equation

\[
F_i(a) = F_j(a).
\]

Here, \( F \) denotes the distribution function of the respective density. When both beta distributions are symmetric around \( \theta = 1/2 \), \( a = 1/2 \).

In this subsection we are however interested in how explicit bounds in terms of the parameters can be derived quickly using the above framework. We thus estimate Wasserstein distances on Beta distributions with (3.5).

First of all, recall from section 2.2 that the Stein kernel of a \( B(\theta|\alpha, \beta) \) distribution is given by \( T(\theta) = \theta(1-\theta) \). The upper bound in (3.5) for a random variable \( \Theta \) arising from a posterior distribution with general prior \( \pi(\theta) \) becomes (given that the assumptions are met):

\[
d_W(\Theta_B, \Theta) \leq \frac{1}{n+2} E[|\rho(\Theta)|\Theta(1-\Theta)]. \tag{3.8}
\]

We first compare the general conjugate prior to the Bayesian prior with density \( q_B \). Recall the following moments of \( \Theta \sim B(\theta|\alpha, \beta) \):

\[
E[\Theta] = \frac{\alpha}{\alpha + \beta}, \quad \text{Var}[\Theta] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}. \tag{3.9}
\]

With \( \Theta_{\alpha,\beta} \sim q_{\alpha,\beta}(\theta|x)\,d\theta \), we obtain (note that the conditions of Lemma 2.5.2 are met, with \( 0 < x < n \) for the Haldane prior):

\[
d_W(\Theta_B, \Theta_{\alpha,\beta}) \geq |E[\Theta_B] - E[\Theta_{\alpha,\beta}]| = \left| \frac{x+1}{n+2} - \frac{x + \alpha}{n + \alpha + \beta} \right|
\]
\[
= \frac{1}{n + \alpha + \beta} \left| (\alpha + \beta - 2) \frac{x+1}{n+2} - (\alpha - 1) \right|.
\]

\[
\rho_{\alpha,\beta}(\theta) := \frac{B'(\theta|\alpha, \beta)}{B(\theta|\alpha, \beta)} = \frac{(\alpha - 1)(1 - \theta) - (\beta - 1)\theta}{\theta(1 - \theta)},
\]

\[
T_1(\theta)\rho_{\alpha,\beta}(\theta) = \frac{1}{n+2} [(\alpha - 1)(1 - \theta) - (\beta - 1)\theta].
\]
Which yields for the estimate (3.8):

\[
d_W(\Theta_B, \Theta_{\alpha,\beta}) \leq \frac{1}{n+2} \mathbb{E}[(\alpha - 1) - (\alpha + \beta - 2)\mathbb{E}[\Theta_{\alpha,\beta}]] + \frac{|\alpha + \beta - 2|}{n+2} \mathbb{E}[\Theta_{\alpha,\beta} - \mathbb{E}[\Theta_{\alpha,\beta}]] \\
\leq \frac{1}{n+2} (\alpha - 1) - (\alpha + \beta - 2) \frac{(x + \alpha)}{n + \alpha + \beta} + \frac{|\alpha + \beta - 2|}{n+2} \sqrt{\text{Var}[\Theta_{\alpha,\beta}]} \\
\leq \frac{1}{n + \alpha + \beta} (\alpha + \beta - 2) \frac{x + 1}{n + 2} - (\alpha - 1) + \frac{|\alpha + \beta - 2|}{(n + \alpha + \beta)(n+2)} \sqrt{\frac{(x + \alpha)(n - x + \beta)}{(n + \alpha + \beta + 1)}}.
\]

The second term is almost surely of order \( O(n^{-3/2}) \) if we regard \( X \) as a sum of \( n \) i.i.d. variables. The first term vanishes for \( \alpha + \beta \neq 2 \) if

\[
x = \frac{(\alpha - \beta) + (\alpha - 1)n}{\alpha + \beta - 2}.
\]

Suppose the marginal distribution of \( X \) arises as a sum of \( n \) i.i.d. Bernoulli distributed random variables with parameter \( \theta_0 \), i.e. it is binomially distributed. Then, if \( \theta_0 \) coincides with \( (\alpha - 1)/(\alpha + \beta - 2) \), the central limit theorem implies that with probability arbitrarily close to 1 convergence will be \( O(n^{-3/2}) \). In this case we namely have \( X = n(\alpha - 1)/(\alpha + \beta - 2) + \sqrt{n}Y \) with \( Y \) tending in distribution to a normal distribution.

[17] treats Jefferys’ prior \( \alpha = \beta = 1/2 \) in this way, and retrieves (3.10). Comparing the Bayesian and Haldane prior (under \( 0 < x < n \) to suffice the conditions of Lemma 3.4) yields (3.11). Comparing the Jeffreys and Haldane prior directly can be done by the triangle inequality or by the transformations \( x \mapsto (x - 1)/2, n \mapsto (n - 1), \alpha = \beta = 1/2 \), giving (3.12). Note that the lower bound has been rewritten so that it emphasizes some symmetry.

\[
\frac{1}{n+2} \left| \frac{1}{2} - \frac{x + 1/2}{n+1} \right| \leq d_W(P_B, P_J) \leq \frac{1}{n+2} \left| \frac{1}{2} - \frac{x + 1/2}{n+1} \right| + \frac{1}{n+2} \sqrt{\frac{(x+1/2)(n-x+1/2)}{(n+1)^2(n+2)}} ,
\]

(3.10) \[
\frac{1}{n+2} \left| \frac{1}{2} - \frac{x}{n} \right| \leq d_W(P_B, P_H) \leq \frac{2}{n+2} \left( \frac{1}{2} - \frac{x}{n} \right) + \sqrt{\frac{x(n-x)}{n^2(n+1)}} ,
\]

(3.11) \[
\frac{1}{n+2} \left| \frac{1}{2} - \frac{x}{n} \right| \leq d_W(P_J, P_H) \leq \frac{1}{n+2} \left( \frac{1}{2} - \frac{x}{n} \right) + \sqrt{\frac{x(n-x)}{n^2(n+1)}} .
\]

(3.12)
3.2 Maximal likelihood estimators

In [1] and [2], Stein’s method is used to assess how fast the maximal likelihood estimator (MLE) of a model based on i.i.d. data can converge to a normal distribution. Suppose $X = (X_i)_{i=1,...,n}$ arises as a sample of i.i.d. random variables with joint distribution $f(x|\theta) \times_{i=1}^n \nu(dx)$ with respect to a canonical measure $\nu$ on a Polish state space $S$. Here $\theta$ denotes a parameter that takes its values in a parameter space $(a,b) \subset \mathbb{R}$. We suppose that $f$ is positive on $S$ for all choices of $\theta \in (a,b)$. When the density function $f$ is regarded as a function of $\theta$, this is stressed by calling it the likelihood function $L(\theta|x) = f(x|\theta)$. It is usually more convenient to study its natural logarithm, the log-likelihood function $l(\theta;x) = \log L(\theta;x)$ (we extend the definition of log to $\log(0) := -\infty$).

Here we argue from a frequentist perspective, as opposed to the previous section. It is assumed that there is a fixed parameter $\theta_0$ that describes the model from which $X$ arises, but it is unknown. Given some observed values $X_i = x_i$, one would like to estimate $\theta_0 \in (a,b)$ with the maximal likelihood estimator $\hat{\theta}_n(X)$. When this is possible, it is defined as a value that maximizes the (log-)likelihood function. We assume that first, second, ..., $j$-th derivatives of $f$ with respect to $\theta$ exist for some $j \geq 3$ and $x \in S^n$. They will be denoted as $l', l'', \ldots, l^{(j)}$ ($f$ does not become zero). Since we have assumed the parameter space to be open, $l'(\hat{\theta}(x), x) = 0$ if the estimator exists. Under some regularity assumptions, it is well-known that $\hat{\theta}(x)$ is consistent and even asymptotically normally distributed after standardization. A clear discussion of stringent and less stringent conditions for the fore-mentioned properties, and more, can be found in chapter 5 of [28] in the context of $M$- and $Z$-estimators. We will always assume that a unique maximal likelihood estimator exists with probability 1.

Furthermore it is assumed that:

1. The parameter is identifiable, i.e. if $\theta \neq \theta'$, then $f(\cdot|\theta) \neq f(\cdot|\theta')$.

2. $\frac{\partial^k}{\partial \theta^k} \int_S f(x|\theta) \times_{i=1}^n \nu(dx) = \int S \frac{\partial^k}{\partial \theta^k} f(x|\theta) \times_{i=1}^n \nu(dx)$ for $k = 1, 2, 3$.

3. For all $\hat{\theta} \in (a,b)$ and $x \in S^n$ there is a $\epsilon_{\hat{\theta}} > 0$ and a measurable function $M_{\hat{\theta}} : S \rightarrow [0, \infty]$ such that $E_{\hat{\theta}}[M_{\hat{\theta}}(X)] < \infty$ and:

$$\left| \frac{\partial^3}{\partial \theta^3} \log f(x|\theta) \right| \leq M_{\hat{\theta}}(x), \quad x \in S^n, \theta \in (\hat{\theta} - \epsilon_{\hat{\theta}}, \hat{\theta} + \epsilon_{\hat{\theta}}).$$

$E_{\hat{\theta}}$ denotes the expectation when $X \sim f(x|\hat{\theta}) \times_{i=1}^n \nu(dx)$.

4. The Fisher information $\mathcal{I}(\theta_0) := \text{var}[l'(\theta_0; x)^2] \neq 0$.

These are conditions that allow direct Taylor expansion to yield information on the behavior of the MLE $\hat{\theta}_n$. As argued in [28], they are more stringent than needed for central limit theorems to hold. 2. implies that $E[l'(\theta; X)] = 0$, for all $\theta \in (a,b)$ and $\mathcal{I}(\theta_0) = E[l''(\theta_0; X)]$.

Stein’s method comes into consideration by the following lemma:
Lemma 3.2.1. For i.i.d. random variables with \( \mathbb{E}[Y_i] = 0 \) and \( \mathbb{E}[Y_i^2] = \sigma^2 > 0 \), \( W := n^{-1/2} \sum_{i=1}^{n} Y_i \), \( K \sim N(0, \sigma^2) \), the following bound on the Wasserstein distance is valid:

\[
|\mathbb{E}[h(W)] - \mathbb{E}[h(K)]| \leq \sigma \|h'\|_{\infty} \sqrt{n} \left( 2 + \frac{\mathbb{E}[|Y_i|^3]}{\sigma^3} \right),
\]

(3.13)

This is just a consequence of Lemmas 1.1.6 and 1.3.10 (the extra factor \( \sigma \) is due to scaling).

In [1], some bounds on the bounded Wasserstein distance between \( W/\sigma \) and a standard normal are stated for general models. An \( O(n^{-1/2}) \)-decay is obtained.

To illustrate how to achieve explicit bounds, we discuss the situation where there exists a one-to-one function \( r : (a,b) \to \mathbb{R} \in C^2(a,b) \) such that the maximal likelihood estimator is the solution of:

\[
r(\hat{\theta}_n(X)) = \frac{1}{n} \sum_{i=1}^{n} t(X_i).
\]

(3.14)

Here, \( t : S \to \mathbb{R} \) are integrable functions such that the right hand side is in the range of \( r \) (such that the MLE exists). Such formulas are common if \( \theta \) may be multidimensional. For one dimension, \( \theta \in (a,b) \), suppose that \( X_i \) stems from a 1-dimensional exponential family with density

\[
f_{i}(x_i|\theta) = c(\theta)h(x_i)\exp^{Q(\theta)t(x_i)}, \quad x_i \in S,
\]

with respect to \( \nu \). Here, \( Q : (a,b) \to \mathbb{R} \) is supposed to be one-to-one and twice continuously differentiable. The density is assumed to be non-zero on \( S \), for all choices of \( \theta \). First suppose \( Q(\theta) = \theta \). By computing the zero points of the score function \( l' \), it can be seen that the MLE satisfies (chapter 4 of [28]):

\[
\frac{1}{n} \sum_{i=1}^{n} t(X_i) = \mathbb{E}_{\hat{\theta}_n(X)}[t(X_1)],
\]

and so for general \( Q \):

\[
\frac{1}{n} \sum_{i=1}^{n} t(X_i) = \mathbb{E}_{Q(\hat{\theta}_n(X))}[t(X_1)].
\]

Then there should still be some regularity so that the right hand side is invertible with probability 1 (often, it tends to 1 for \( n \to \infty \) however). Examples include:

1. The normal distribution \( f_i(x_i|\mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right), x_i \in \mathbb{R} \), where \( \sigma^2 > 0 \) is known. Then \( \mu \in \mathbb{R} \) is estimated by:

\[
\hat{\theta}_n(X) = \frac{1}{n} \sum_{i=1}^{n} X_i.
\]

2. The normal distribution, where \( \mu \) is known. Then \( \sigma^2 \in (0, \infty) \) is estimated by:

\[
\hat{\theta}_n(X) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2.
\]
3. The Weibull distribution with density \( f_i(x_i|\alpha, \sigma) = \frac{\alpha}{\sigma} (x_i/\sigma)^{\alpha-1} \exp(-(x_i/\sigma)^\alpha) \), \( x_i > 0 \), where \( \alpha > 0 \) is known. The MLE for \( \sigma \) becomes:

\[
\left( \hat{\theta}_n(X) \right)^\alpha = \frac{1}{n} \sum_{i=1}^{n} X_i^\alpha.
\]

4. The Laplace scale model, \( f_i(x_i|\sigma) = \frac{1}{2\sigma} \exp(-|x_i|/\sigma), \ x_i \in \mathbb{R} \). The MLE for \( \sigma > 0 \) becomes:

\[
\hat{\theta}_n(X) = \frac{1}{n} \sum_{i=1}^{n} |X_i|.
\]

Then a bound from [2] reads:

**Proposition 3.2.2.** Consider \( X = (X_i)_{i=1, \ldots, n} \) consisting of i.i.d \( S \)-valued random variables \( X_i \) from the distribution \( f_i(x_i|\theta_0)\mu(dx_i) \). Assume conditions 1. – 4. above and that the MLE \( \hat{\theta}(X) \) exists and is unique. Suppose \( \hat{\theta}_n(X) \) is the solution of (3.14) with \( r : (a, b) \to \mathbb{R} \) one-to-one, twice continuously differentiable. Also assume that \( r'(\theta) \neq 0 \) for all \( \theta \in (a, b) \) and \( t : S \to \mathbb{R} \) is measurable. For any \( \epsilon > 0 \) such that \( (\theta_0 - \epsilon, \theta_0 + \epsilon) \subset (a, b) \) and Lipschitz \( h \):

\[
\left| \mathbb{E}[h(\sqrt{nI}(\theta_0)(\hat{\theta}_n(X)) - \theta_0)] - \mathbb{E}[h(Z)] \right| \leq \frac{\|h\|_{\infty}}{\sqrt{n}} \left( 2 + \frac{I(\theta_0)^{3/2}}{|r'(\theta_0)|^{3/2}} \mathbb{E}|t(X_i) - r(\theta_0)|^3 \right) \\
+ 1_{r' \neq 0} \mathbb{E} \left[ (\hat{\theta}_n(X) - \theta_0)^2 \right] \left( \frac{2\|h\|_{\infty}}{\epsilon^2} + \frac{\|h\|_{\infty} \sqrt{nI(\theta_0)}}{2|r'(\theta_0)|} \sup_{\theta : |\theta - \theta_0| \leq \epsilon} |r''(\theta)| \right).
\]

**Proof.** Under conditions 1. – 4., it is known that a central limit theorem holds for \( n \to \infty \) (see chapter 5 of [28]):

\[
\sqrt{nI(\theta_0)}(\hat{\theta}_n(X) - \theta_0) \to N(0, 1), \quad \text{in distribution.} \quad (3.15)
\]

Since \( r'(\theta_0) \neq 0 \), Taylor approximation yields:

\[
\frac{\sqrt{nI(\theta_0)}}{r'(\theta_0)} (r(\hat{\theta}_n(X)) - r(\theta_0)) = \sqrt{nI(\theta_0)}(\hat{\theta}(X) - \theta_0) + \frac{\sqrt{nI(\theta_0)}}{2r'(\theta_0)} r''(\theta^*_n)(\hat{\theta}(X) - \theta_0)^2,
\]

for some \( \theta^*_n \in (\theta_0, \hat{\theta}(X)) \). The first term on the right hand side converges to \( N(0, 1) \) in distribution, while the second converges to 0. This last statement holds true because \( (\hat{\theta}(X) - \theta_0) \) converges to 0 in distribution and thus probability. Then \( r''(\theta^*_n) \) approaches \( r''(\theta_0) \) so that it is bounded in probability. Thus, as a product of terms that are bounded in probability and one that converges to 0 in probability, it converges to 0 in probability. We have:

\[
\frac{\sqrt{nI(\theta_0)}}{r''(\theta_0)}(r(\hat{\theta}_n(X)) - r(\theta_0)) \to N(0, 1), \quad \text{in distribution}
\]
\[
\left| \mathbb{E}[h\left(\sqrt{n\mathcal{I}(\theta_0)}(\hat{\theta}_n(X)) - \theta_0\right)] - \mathbb{E}[h(Z)] \right| \\
\leq \left| \mathbb{E}\left[h\left(\sqrt{n\mathcal{I}(\theta_0)/r'(\theta_0)}(r(\hat{\theta}_n(X)) - r(\theta_0))\right)\right] - \mathbb{E}[h(Z)] \right| \\
+ \left| \mathbb{E}\left[h\left(\sqrt{n\mathcal{I}(\theta_0)/r'(\theta_0)}(r(\hat{\theta}_n(X)) - r(\theta_0))\right) - h\left(\sqrt{n\mathcal{I}(\theta_0)}(\hat{\theta}_n(X))\right)\right] \right|.
\]

(3.17)

For the first term convergence was already noted but for a bound we reformulate:

\[
\sqrt{n\mathcal{I}(\theta_0)}(\hat{\theta}_n(X)) - \theta_0 = \sqrt{n\mathcal{I}(\theta_0)/r'(\theta_0)} \left(\frac{1}{n} \sum_{i=1}^{n} t(X_i) - q(\theta_0)\right).
\]

Because it is a sum of i.i.d. random variables, we can apply Lemma 3.2.1. The variance of the variables themselves is 1 after taking in the factor before the brackets, because the total term converges to a \(N(0, 1)\) distribution. So the first term of the right hand side can be bounded by:

\[
\frac{||H'||_{\infty}}{\sqrt{n}} \left(2 + \frac{\mathcal{I}(\theta_0)^{3/2}}{|r'(\theta_0)|^{3/2}} \mathbb{E}|t(X_i) - r(\theta_0)|^3\right).
\]

So, comparing a standard normal to the substituted quantity with \(r\) yields explicit bounds. When \(r'\) is not a constant, or \(r'' \neq 0\), this leads to an error \(A\) of which the expected value remains to be estimated. In order to have bounded terms, it is split into two cases by the law of total probability:

\[
|\mathbb{E}[A]| \leq \mathbb{E}\left[|A| \mid |\hat{\theta}_n(X) - \theta_0| > \varepsilon\right] \mathbb{P}\left[|\hat{\theta}_n(X) - \theta_0| > \varepsilon\right] \\
+ \mathbb{E}\left[|A| \mid |\hat{\theta}_n(X) - \theta_0| \leq \varepsilon\right] \mathbb{P}\left[|\hat{\theta}_n(X) - \theta_0| \leq \varepsilon\right].
\]

Then Markov’s inequality, |\(A\)| \(\leq 2||h||_{\infty}\) and \(\mathbb{P}\left[|\hat{\theta}_n(X) - \theta_0| \leq \varepsilon\right] \leq 1\) yield:

\[
|\mathbb{E}[A]| \leq 2||h||_{\infty} \frac{\mathbb{E}\left[|\hat{\theta}_n(X) - \theta_0|^2\right]}{\varepsilon^2} + \mathbb{E}\left[|A| \mid |\hat{\theta}_n(X) - \theta_0| \leq \varepsilon\right].
\]

Using the Lipschitz property of \(h\) and (3.16) subsequently gives:

\[
|A| \leq ||h'||_{\infty} \frac{\sqrt{n\mathcal{I}(\theta_0)}}{2r'(\theta_0)} r''(\theta_\ast_n)(\hat{\theta}(X) - \theta_0)^2.
\]

And thus:

\[
\mathbb{E}\left[|A| \mid |\hat{\theta}_n(X) - \theta_0| \leq \varepsilon\right] \leq ||h'||_{\infty} \frac{\sqrt{n\mathcal{I}(\theta_0)}}{2r'(\theta_0)} \sup_{\theta : |\theta - \theta_0| \leq \varepsilon} |r''(\theta)| \mathbb{E}\left[(\hat{\theta}_n(X) - \theta_0)^2\right].
\]
This is because $\mathbb{E}[Y^2 \mid |Y| \leq \epsilon] \leq \mathbb{E}[Y^2]$ for any random variable:

$$
\mathbb{E}[Y^2] = \mathbb{E}[Y^2 \mid |Y| \leq \epsilon](1 - \mathbb{P}[|Y| > \epsilon]) + \mathbb{E}[Y^2 \mid |Y| > \epsilon]\mathbb{P}[|Y| > \epsilon] \\
= \mathbb{E}[Y^2 \mid |Y| \leq \epsilon] + (\mathbb{E}[Y^2 \mid |Y| > \epsilon] - \mathbb{E}[Y^2 \mid |Y| \leq \epsilon])\mathbb{P}[|Y| > \epsilon] \geq \epsilon.
$$

Putting all bounds together yields the result of the proposition. \hfill \square

If the $t(X_i)$ have a finite third moment, then $\epsilon > 0$ can be chosen to see that the bounded Wasserstein distance:

$$d_{bW}(\sqrt{nI(\theta_0)}(\hat{\theta}_n(X)) - \theta_0), Z) = O(n^{-1/2}).$$

Indeed,

$$\mathbb{E}[(\hat{\theta}_n(X) - \theta_0)^2] = \text{var}[\hat{\theta}_n(X)] + \text{bias}^2[\hat{\theta}_n(X)],$$

where the MLE is asymptotically efficient under 1.-4. (see again [28]):

$$n\text{var}[\hat{\theta}_n(X)] \rightarrow I(\theta_0)^{-1},$$

and the bias is of order $n^{-1/2}$ by (3.15).

That the Fisher information is non-zero, is clearly needed. However, to overcome this, or to overcome a singular Fisher matrix in the multidimensional case ($\theta \in \mathbb{R}^d$), one could investigate higher order derivatives of the likelihood function. This is done in [5] for asymptotic results of the MLE, where the rank of the Fisher matrix is one degree less than the dimension of the parameter space. Stein’s method as in [1] could then perhaps be adapted to this setting to investigate the decay on the bounded Wasserstein distance between the MLE and a standard normal. Lastly, in [2], the above result gets specialized to exponential families and a simulation study on exponentially distributed variables is performed. For very large values of $n$, the bounds seem to get close to the quantity of interest.
Chapter 4

Gaussian analysis: the language of Malliavin calculus

Malliavin calculus describes differential operators on multi- or infinite-dimensional Gaussian spaces. It allows to give probabilistic interpretations to objects of interest, such as solutions of PDE. Also, it matches the needs of Stein’s framework. We namely want to work out the recipe from sections 1.4 and 2.1. For Gaussian distributions, such a theory works remarkably well. A drawback is that only processes that are measurable with respect to the Gaussian family can be considered. The main references are [20] and [8] for Malliavin calculus and [18] for the interaction with Stein’s method. The first two offer an extensive description plus a nice overview over different kinds of applications. The latter specializes to the use for Stein’s method.

We will work out the language of Malliavin calculus, while trying to pertain a picture of the underlying PDE structure. The idea of the framework is a specialized version of the one behind section 2.1, connected with certain properties of PDE. Developing the language takes up a considerable amount of the thesis, so that we can express ourselves in it and analyze a unifying view. It is difficult to deploy without having some familiarity with the arguments behind the theory. We highlight useful connections and indicate the ambiguities that are related to Malliavin calculus.

4.1 Gaussians: motivation and encoding scheme

In probability theory, Gaussian distributions are often put central due to convenient approximation theorems and computation properties. An advantage that Gaussian families have over other families, is that they can be encoded well by characteristic functions. Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and real random variables $(W_t)_{t \in T}$ indexed by an index set $T$.

**Definition 4.1.1.** $(W_t)_{t \in T}$ is called a Gaussian family or Gaussian process if and only if for any $W_{t_1}, \ldots, W_{t_k}$ there exists a $\mu \in \mathbb{R}^k$ and a symmetric positive semi-
definite\(^1\) matrix \(\Sigma \in \mathbb{R}^{k \times k}\) such that:

\[
\mathbb{E} \left[ e^{it^T(W_{n_1}, \ldots, W_{n_k})} \right] = e^{it^T \Sigma t}, \quad t \in \mathbb{R}^k.
\] (4.1)

By the theorem of Daniell-Kolmogorov, Theorem B.3.1, it is sufficient to specify a mean function \(\mu(t) := \mathbb{E}[W_t]\) and a covariance function \(\Sigma(s, t) := \text{cov}(W_s, W_t)\) to construct a probability space with on it a Gaussian process that has these means and covariances. The distribution of the process is uniquely determined. Sometimes some considerations have to be made in order to restrict to a more convenient \(\sigma\)-algebra\(^2\).

**Example 4.1.2.** \(d\)-dimensional normal distributions are a special case of interest, but also a Brownian motion \((B_t)_{t \in [0, T]},\) with \(T \in [0, \infty]\). Furthermore, solutions of stochastic differential equations (SDE)\(^3\) may also be considered:

\[
dX_t = b(t, X_t)dt + \sigma(t, X_t)dB_t, \quad X_0 = \chi, \quad t \in [0, T).
\] (4.2)

We have denoted real random variables \(X_t\) and \(\chi\) with \(t \in [0, T)\), measurable \(b : [0, T) \times \mathbb{R} \to \mathbb{R}\) and \(\sigma : [0, T) \times \mathbb{R} \to \mathbb{R}\).

An encoding scheme that is fundamental for the coming abstractions is based on a measure-theoretic construction:

**Example 4.1.3** (Gaussian measure). Consider a measure space \((S, \mathcal{B}, \nu)\) where \((S, \mathcal{B})\) is a Polish space and \(\nu\) is a positive, \(\sigma\)-finite, non-atomic measure on the Borel sets \(\mathcal{B}\). Then \(H = L^2(S, \mathcal{B}, \nu) := L^2(\nu)\) is a real separable Hilbert space for the inner product \(\langle f, g \rangle := \int_S f g d\nu\). A family of random variables \(W\) on \((\Omega, \mathcal{F}, \mathbb{P})\) is called a real Gaussian random measure over \((S, \mathcal{B})\) with control \(\nu\) if

\[
W = \{W(B) \mid B \in \mathcal{B}, \nu(B) < \infty\}
\] (4.3)

is a centered Gaussian family with covariance function \(\mathbb{E}[W(B)W(C)] = \nu(B \cap C)\). This gives an isometry which we can extend by standard measure theoretic induction. Define elementary functions as

\[
\xi := \left\{ \sum_{i=1}^n a_i \mathbb{1}_{A_i} \mid n \in \mathbb{N}, a_i \in \mathbb{R}, A_i \in \mathcal{B} \text{ disjoint}, \nu(A_i) < \infty \right\}.
\]

For \(g := \sum_{i=1}^n a_i \mathbb{1}_{A_i} \in \xi\), define \(W(g) = \sum_{i=1}^n a_i W(A_i)\). With \(h \in \xi\), the isometry \(\mathbb{E}[W(g)W(h)] = \nu(gh) = \int_S gh d\nu\) remains valid.

\(\xi\) is dense in \(L^2(S, \mathcal{B}, \nu)\) (see perhaps section 1.1.2 of [20]), we can take the completion of \(\{W(h) \mid h \in \xi\}\) in \(L^2(\nu)\) and obtain a Gaussian process \(\{W(h) \mid h \in L^2(\nu)\}\). Then, (4.3) again holds true. These are so-called Wiener-Itô integrals. For multiple Wiener-Itô integrals,

\(^1\)Meaning that \(\Sigma\) is symmetric and \(x^T \Sigma x \geq 0\) for any \(x \in \mathbb{R}^k\).

\(^2\)For Brownian motion, one wants to restrict the \(\sigma\)-algebra of continuous functions starting in 0.


61
refer to section 1.1.2 in [20]. Choosing $S = [0, \infty)$ and $\nu$ the Lebesgue measure, we obtain $W_t := W(1_{[0,t)})$, $t \in S$. Because these are centered with covariance function $\mathbb{E}[W_sW_t] = s \wedge t$, we see that by the theorem of Kolmogorov-Chentsov ([14]), we may find a continuous version, i.e. a Brownian motion.

We mention this procedure, because an analogous construction can be performed for a Poisson process, based on intensity measures\(^4\). In this way, a similar theory might perhaps be constructed for jump processes.

Actually, an abstraction allows to use the essential information more efficiently:

**Definition 4.1.4.** Consider a real separable Hilbert space $H$ with inner product $\langle \cdot, \cdot \rangle_H$ and norm $\|\cdot\|_H$. A family of real random variables $W = \{W(h) \mid h \in H\}$ on $(\Omega, \mathcal{F}, \mathbb{P})$ is called an **isonormal Gaussian process** over $H$ if and only if it is a centered Gaussian family with covariance function $\mathbb{E}[W(g)W(h)] = \langle g, h \rangle_H$, for all $g, h \in H$.

**Remark 4.1.5.** 1. Given a Hilbert space $H$, it is again possible to construct an isonormal Gaussian process over $H$ by the extension theorem of Daniell-Kolmogorov.

2. $H \rightarrow L^2(\Omega, \mathcal{F}, \mathbb{P})$; $h \mapsto W(h)$ is linear. For $a, b \in \mathbb{R}$ and $g, h \in H$:

$$
\mathbb{E}[(W(ag + bh) - aW(g) - bW(h))^2] = \|ag + bh\|^2_H + a^2\|g\|^2_H + b^2\|h\|^2_H
$$

$$
- 2b\langle ag + bh, h \rangle_H - 2a\langle ag + bh, g \rangle_H + 2ab\langle g, h \rangle_H = 0.
$$

3. Because the mapping $h \mapsto W(h)$ is linear, it would actually be sufficient to assume that every $W(h)$ is a centered Gaussian random variable:

$$
\mathbb{E}[e^{it^T(W(h_1), \ldots, W(h_k))}] = \mathbb{E}[e^{iW(t^T(h_1), \ldots, h_k))}] = e^{-\|t^T(h_1, \ldots, h_k)\|^2_H/2}, \quad t \in \mathbb{R}^k.
$$

This abstraction is a method to encode the dimensions of the process efficiently and in a non-canonical way. By encoding, we mean the isomorphism $h \mapsto W(h)$ that maps $H$ onto the closed subspace of the considered Gaussian family in $L^2(\Omega, \mathcal{F}, \mathbb{P})$. For a non-degenerate $d$-dimensional normal distribution $W = (W_1, \ldots, W_d)$, the Hilbert space can be $\mathbb{R}^d$. We could choose the canonical base $e_1 = (1,0,\ldots,0), \ldots, e_d = (0, \ldots, 0,1)$ to describe the process $W$. The base $(\sqrt{2}^{-1}(e_1 + e_2), \sqrt{2}^{-1}(e_1 - e_2), e_3, \ldots, e_d)$ would however also work. Other Gaussian processes, such as Gaussian free fields and isonormal processes derived from covariances (see section 2.1 of [18]) can also be encoded by an isonormal Gaussian process. Actually, since all infinite-dimensional separable Hilbert spaces are isomorphic to each other, all the infinite-dimensional isonormal Gaussian processes are isomorphic as well. To solve a specific problem, it is however important that the encoding scheme should be well adapted.

\(^4\)For more information, see chapter 1 of [11].
4.2 Derivative operators and duals

We now try to work out the recipe from sections 1.4 and 2.1. Fix a real separable Hilbert space $H$ with associated isonormal Gaussian process $W = \{W(h) \mid h \in H\}$ on a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For the rest of this chapter assume that $\mathcal{F}$ is the completion with respect to $\mathbb{P}$ of the sigma-algebra generated by $W$. That is, $\mathcal{F} = \sigma(W(h) \mid h \in H)^\mathbb{P}$. We know for one-dimensional normal random variables $N$ how to construct derivatives of functions $f(N)$ and how to compute the dual of $D$. This was discussed in sections 1.1 and 2.2. Now we want to extend this construction to multiple dimensions. As was argued in section 2.1, derivatives should be defined on constants. Because the Gaussian density function is fast decreasing, we can however consider more functions to have regular properties on.

Let $C_p(\mathbb{R}^n)$ denote the functions $f : \mathbb{R}^n \to \mathbb{R}$ that are infinitely differentiable with $f$ and all derivatives having at most polynomial growth. That is, $\exists k \in \mathbb{N}$ such that for all $x \in \mathbb{R}^n |f(x)| \leq C(1 + |x|^{2k})$. For a Gaussian kernel, if $\Sigma \in \mathbb{R}^{n \times n}$ is positive definite, then

$$\lim_{x \to \infty} f(x)e^{-x^T\Sigma^{-1}x/2} = 0.$$ 

So zero-boundary conditions hold. Now consider the following spaces:

**Definition 4.2.1.** 1. The space of smooth random variables $S_p$, the subspace of $L^2(\Omega, \mathcal{F}, \mathbb{P})$ consisting of random variables of the form:

$$F = f(W(h_1), \ldots, W(h_n)), \quad (4.4)$$

where $n \in \mathbb{N}$, $f \in C_p(\mathbb{R}^n)$ and $h_1, \ldots, h_n \in H$.

2. $S_b$, the subspace of random variables $(4.4)$, where now $f \in C_b(\mathbb{R}^n)$.

3. $S_0$, the subspace of random variables $(4.4)$ such that $f \in C_0^\infty(\mathbb{R}^n)$.

**Remark 4.2.2.** If $F$ fulfills $(4.4)$, the representation is not unique. Given $h_1, \ldots, h_n \in H$, orthonormal vectors $e_1, \ldots, e_k \in H$, where $k \leq n$ and $h_1 = ae_1$, $a \in \mathbb{R}$, can be obtained by the Gramm-Schmidt orthonormalization procedure and $g \in C_p(\mathbb{R}^k)$ such that:

$$F = f(W(h_1), \ldots, W(h_n)) = g(W(e_1), \ldots, W(e_k)) \quad \text{a.s.} \quad (4.5)$$

When $h_1, \ldots, h_n$ are orthonormal, we will call this a standard form for $F$.

This is where we need that $\mathcal{F}$ is generated by $W$:

**Lemma 4.2.3.** $S_0$ is dense in $L^p(\Omega, \mathcal{F}, \mathbb{P}), \ p \in [1, \infty)$.

**Proof.** First of all, consider $\psi : \mathbb{R}^n \to [0, 1] \subseteq C_0^\infty(\mathbb{R}^n)$ such that $\psi = 1$ on $[-1, 1]^n$ and $\psi(x) = 0$ for $|x| \geq n$. For $F \in S_p$ in standard form, we can consider densities to infer (using dominated convergence):

$$\psi(W(h_1)/n, \ldots, W(h_n)/n) \cdot F \to F \text{ in } L^p(\Omega, \mathcal{F}, \mathbb{P}), \quad n \to \infty.$$
So $S_0$ is dense in $S_p$ for $L^p(\Omega, F, \mathbb{P})$. Now we will show that $S_p$ is dense in $L^p(\Omega, F, \mathbb{P})$ with the Hahn-Banach theorem, Theorem B.3.9. Let $p' \in \mathbb{R}$ fulfill $p^{-1} + p'^{-1} = 1$. Take $F \in L^{p'}(\Omega, F, \mathbb{P})$ and suppose that $E[FG] = 0$ for all $G \in S_p$. For any $h \in H$ we then have by the Hölder inequalities and dominated convergence:

$$E[Fe^{iW(h)}] = \sum_{k=0}^{\infty} \frac{i^k}{k!} E[FW(h)^k] = 0.$$  \hspace{1cm} (4.6)

Thus by linearity for any $n \in \mathbb{N}$, $t_i \in \mathbb{R}$, $h_i \in H$:

$$E[Fe^{i(t_1W(h_1) + \ldots + t_nW(h_n))}] = 0.$$  \hspace{1cm} (4.7)

This means that the measure $\gamma$ defined on $\sigma(W(h_1), \ldots, W(h_n))$ by

$$\gamma(B) = E[F1_B]$$

has a vanishing Fourier transform. Since it is known that finite measures with Fourier transform 0 are themselves 0 (by similar arguments as in section 9.5 of [9]), we have that

$$E[F1_B] = 0, \quad \forall B \in \sigma(W(h_1), \ldots, W(h_n)).$$

Because these sigma algebra’s generate $F$, it also holds for all $B \in F$. So we have $F = 0$ a.s. The theorem of Hahn-Banach concludes the proof.

**4.2.1 Derivative operator**

We consider the space $L^p(\Omega; H^{\otimes k})$ of random variables $u : \Omega \rightarrow H^{\otimes k}$ such that $\|u\|_{L^p(\Omega; H^{\otimes k})} := E[\|u\|_{H^{\otimes k}}^p]^{1/p} < \infty$, where $H^{\otimes k}$ denotes the Hilbert space of $k$-fold tensor products. Refer to appendix B for a list of notations and properties regarding tensor products. Within $L^p(\Omega; H^{\otimes k})$, we may find the closed subspace of $L^p$-integrable $H^{\otimes k}$-valued random variables, denoted by $L^p(\Omega; H^{\otimes k})$.

**Definition 4.2.4.** For $F \in S_p$ with representation (4.4), we define the **derivative operator** on $S_p$:

$$D : S_p \subset L^p(\Omega) \rightarrow L^p(\Omega; H); F \mapsto \sum_{i=1}^{n} \frac{\partial}{\partial x_i} f(W(h_1), \ldots, W(h_n))h_i.$$  \hspace{1cm} (4.7)

Similarly, we define the **$k$-th iterated derivative operator** on $S_p$:

$$D^k : S_p \subset L^p(\Omega) \rightarrow L^p(\Omega; H^{\otimes k});$$

$$F \mapsto \sum_{i_1, \ldots, i_k=1}^{n} \frac{\partial^k f}{\partial x_{i_1} \ldots \partial x_{i_k}} (W(h_1), \ldots, W(h_n))h_{i_1} \otimes \ldots \otimes h_{i_k}. \hspace{1cm} (4.8)$$

$5_x \mapsto e^{t|x-x^2/2} \in L^1(\mathbb{R})$ for any $t \in \mathbb{R}$.
Remark 4.2.5. 1. $D$ does not follow the convention of chapter 2. For convenience, $-D$ was studied there. Also, we only explicitly state $p \in [1, \infty]$ where otherwise confusion could occur.

2. $D$ and $D^k$ are well-defined. That is, they do not depend on the a.s. representation of $F$. Suppose two representations are given. By the Gramm-Schmidt orthonormalization procedure, interchanging variables and possibly incorporating dummy variables such that $y \mapsto f(x, y)$ is constant, we may assume that they are of the form:

$$f(W(e_1), \ldots, W(e_n)) = g(W(e_1), \ldots W(e_n)) \quad a.s.,$$

where $f, g \in C^\infty_p(\mathbb{R}^n)$, $e_1, \ldots, e_n \in H$ orthonormal. Because the distribution of $(W(e_1), \ldots, W(e_n))$ is mutually absolutely continuous with respect to Lebesgue measure, $f$ and $g$ are a.s. equal and by continuity everywhere.

3. For $k = 1, 2$, these definitions coincide with those of Jacobians and Hessians respectively, but without referring to a canonical base. This can be seen by considering an orthonormal base. Moreover, if $h \in H$, $\langle DF, h \rangle_H$ represents a directional derivative:

$$\langle DF, h \rangle_H = \lim_{\epsilon \to 0} \frac{f(W(h) + \epsilon h_1, \ldots, W(h_n) + \epsilon h_n) - F}{\epsilon}. \quad (4.9)$$

So $H$ is just a way to encode the dimensions in a canonical way, as stated.

4. $D^k F$ takes values in the symmetric tensor products because the derivatives of smooth random variables do not depend on the order of partial differentiation.

By considering $S_p$, the dual of $D$ can conveniently be computed with formula (1.3) as a special case. We start with:

Lemma 4.2.6. For smooth random variables $F, G$ and $h \in H$, we have an integration by parts rule:

$$\mathbb{E}[\langle DF, h \rangle_H] = \mathbb{E}[FW(h)], \quad (4.10)$$

$$\mathbb{E}[G\langle DF, h \rangle_H] = \mathbb{E}[-F\langle DG, h \rangle_H] + \mathbb{E}[FGW(h)]. \quad (4.11)$$

Proof. Suppose $F$ is in standard form (4.5). By linearity we may assume that $h = e_1$. Then it follows from one-dimensional integration by parts, or (1.3) after using Fubini’s theorem to split the dimensions of the integral:

$$\mathbb{E}[\langle DF, h \rangle_H] = \mathbb{E}[\partial_1 f(W(e_1), \ldots W(e_n))] = \mathbb{E}[f(W(e_1), \ldots W(e_n))W(e_1)].$$

The second formula is seen to be valid after bringing $F$ and $G$ in a standard form with shared $e_1, \ldots, e_n$ as in point 2. of the previous remark and using the regular product rule $\partial_i(fg) = g\partial_if + f\partial_ig$.  

65
Now we may illustrate the fundamental closability of differentiation:

**Proposition 4.2.7.** $D^k : S_p \subset L^p(\Omega) \to L^p(\Omega; H^{\otimes k})$ is closable for $p \in [1, \infty)$, $k \in \mathbb{N}$.

**Proof.** The condition of Lemma 2.1.1 needs to be checked. It is sufficient to only show the ‘difficult’ case $p = 1$. Take $k = 1$, $F \in S_b$ and $G_h = Fe^{-\epsilon W(h)^2}$, $h \in H$. Suppose $F_n \in S_p \to 0$ in $L^p(\Omega)$ and $DF_n \to \eta$ in $L^p(\Omega; H)$. Inspecting formula (4.11):

$$\mathbb{E}[G(\eta, h)_H] = \lim_{n \to \infty} \mathbb{E}[G(DF_n, h)_H]$$

$$= \lim_{n \to \infty} (\mathbb{E}[-F_n(DG, h)_H] + \mathbb{E}[F_n GW(h)]) = 0.$$

We have used that $\langle DG, h \rangle_H$ and $GW(h)$ are bounded. By letting $\epsilon \to 0$, dominated convergence yields that $\mathbb{E}[F(\eta, h)_H] = 0$. Note that we can not use the Hölder inequalities since $S_b$ is not necessarily dense in $L^\infty(\Omega)$. Fix an orthonormal base $(e_n)_{n \in \mathbb{N}}$ for $H$ ($N$ countable). Taking $B \subset \mathbb{R}^k$ Borel and $\psi$ as in the proof of Lemma 4.2.3, we use a standard Dirac approximation $\psi = n^d(\int_{\mathbb{R}^k} \psi^{-1}(n \cdot))$. Then the convoluted sequence $\rho_n := 1_B * \psi_n \in S_b$ has a subsequence that converges pointwisely a.e. to $1_B$, while $\|\rho_n\|_\infty \leq 1$. $F = \rho_n(W(e_1), \ldots, W(e_k))$, and dominated convergence for $n \to \infty$ then yield $\mathbb{E}[1_B(W(e_1), \ldots, W(e_k))(\eta, h)_H] = 0$. We have used here that $(W(e_1), \ldots, W(e_n))$ has a density with respect to Lebesgue measure. Because the $\sigma(W(e_1), \ldots, W(e_k))$ generate $\mathcal{F}$, $(\eta, h)_H = 0$ a.s. for all $h \in H$. Finally, we only need countably many $e_k$ to infer $$\mathbb{P}[\langle \eta, e_k \rangle_H = 0, \forall k \in \mathbb{N}] = 1,$$

and $\eta = 0$ a.s. For $k > 1$, the integration by parts rule (4.2.6) may be iterated and a similar argument yields the result. \hfill \Box

The iterated integration by part formula can be investigated by multilinearity and going to a standard form ($(e_i)_{i \in \mathbb{N}}$ being orthonormal):

$$\mathbb{E}[\langle D^k f(W(e_1), \ldots, W(e_n)), e_{i_1} \otimes \ldots \otimes e_{i_k} \rangle_{H^{\otimes k}}]$$

$$= (2\pi)^{-n/2} \int_{\mathbb{R}^n} \partial_{i_1, \ldots, i_k} f(x_1, \ldots, x_n) e^{-(x_1^2 + \ldots + x_n^2)/2} \prod_{j=1}^n dx_j$$

$$= \mathbb{E}[f(W(e_1), \ldots, W(e_n)) H_{i_1, \ldots, i_k} ((W(e_{i_1}), \ldots, W(e_{i_k}))],$$

where $H_{i_1, \ldots, i_k}(x_1, \ldots, x_k) = (-1)^k e^{-(x_1^2 + \ldots + x_k^2)/2} \partial_{i_1, \ldots, i_k} e^{-(x_1^2 + \ldots + x_k^2)/2}$ is a multidimensional version of a Hermite polynomial. We will come back to this.

Following the procedure of section 2.1, $D^k$ may be closed by the norm:

$$\|F\|_{k,p} := \left( \mathbb{E}[\|F\|^p] + \sum_{j=1}^k \mathbb{E}[\|D^j F\|_{H^{\otimes j}}^p] \right)^{1/p}.$$  (4.13)

We obtain a possibly infinite-dimensional Banach space of Sobolev-type that will be denoted by the symbol $\mathbb{D}^{k,p} := \text{dom}(D^k)$, with the above norm ($D^k$ now being the closed operator). Also define $\mathbb{D}^{\infty,p} := \cap_{k \geq 1} \mathbb{D}^{k,p}$. In the case $p = 2$, $\mathbb{D}^{k,2}$ is a Hilbert space with inner product:

66
\[ \langle F, G \rangle_{k,2} := \mathbb{E}[FG] + \sum_{j=1}^{k} \mathbb{E}[(D^jF, D^jG)_{H^{\otimes j}}]. \]  

(4.14)

\[ \mathbb{D}^{k,p} \] may be canonically embedded in \((L^p(\Omega), \ldots, L^p(\Omega; H^{\otimes k}))\) by \( F \mapsto (F, DF, \ldots, D^kF) \). Thus it makes sense to speak of \( \mathbb{D}^{k,p} \cap \mathbb{D}^{j,q} \) in \( L^{p\wedge q}(\Omega) \). As point 2. of the following Lemma points out, the intersection is also well-defined in \((L^{p\wedge q}(\Omega), \ldots, L^{p\wedge q}(\Omega; H^{\otimes (k\wedge j)}))\).

Some basic properties are:

**Remark 4.2.8.**

1. Monotonicity: \( \|F\|_{k,p} \leq \|F\|_{j,q} \) for \( F \in \mathcal{S}_p \) and \( p \leq q, k \leq j \).

2. Compatibility: For \( F \in \mathbb{D}^{k,p} \cap \mathbb{D}^{j,q} \subset L^1(\Omega) \), \( D^iF \) is well-defined for \( i \leq (k \wedge j) \), i.e. equal for the completions of \( \|\cdot\|_{k,p} \) and \( \|\cdot\|_{j,q} \). This is a consequence of the closability of the operators and the monotonicity of norms \( \|\cdot\|_{k,p} \). In particular \( \mathbb{D}^{j,q} \subset \mathbb{D}^{k,p} \) for \( p \leq q, k \leq j \).

3. The continuity can be used to extend formula’s such as (4.11) to \( F \in \mathbb{D}^{1,p} \), \( G \in \mathcal{S}_p \), by Hölder’s inequalities.

We iterate again that some care has to be taken with tensor products due to the equality under multilinear combinations of components. Also, the norm \( \|\cdot\|_{H^{\otimes k}} \) is in fact quite strong. It restricts the random variables that can be considered, but will yield strong properties.

We mention some computation rules, but the proofs do not offer much new insight. They may be found in [20].

**Theorem 4.2.9 ((Extended) chain rule).** Let \( p \in [1, \infty) \). For a continuously differentiable function with bounded derivatives \( \varphi : \mathbb{R}^n \to \mathbb{R} \) and \( F = (F_1, \ldots, F_n) \) with \( F_i \in \mathbb{D}^{1,p} \), \( \varphi(F) \in \mathbb{D}^{1,p} \) we have:

\[ D\varphi(F) = \sum_{i=1}^{n} \partial_i \varphi(F)DF_i. \]  

(4.15)

If \( p > 1 \), then it is also true if \( \varphi \) is Lipschitz and the law of \( F \) is absolutely continuous with respect to Lebesgue measure.

These assumptions actually are stringent, as [18] notes. For instance, \( D(W(h)^m) = mW(h)^{m-1}h \) and \( De^W(h) = e^W(h)h, m \in \mathbb{N}, h \in H \), by approximation by \( \mathcal{S}_p \). The Theorem may for example yield for \( h_1, \ldots, h_n \in H \), \( D(\max_{1 \leq i \leq n} W(h_i)) = h_{I_0} \), where \( I_0 := \arg\max_{1 \leq i \leq n} W(h_i) \) [18].

As expected, the kernel of \( D \) is given by the constants.

**Proposition 4.2.10.** Let \( F \in \mathbb{D}^{1,1} \). If \( DF = 0 \), then \( F = \mathbb{E}[F] \).

The proof again has to rely on approximation arguments however. A computation rule for higher order derivatives of products is:
**Proposition 4.2.11** (Leibniz rule). Let \( k \in \mathbb{N} \) and \( p, q, r \in [1, \infty) \) such that \( r^{-1} = p^{-1} + q^{-1} \). If \( F \in \mathbb{D}^{k,p} \) and \( G \in \mathbb{D}^{k,q} \), then \( FG \in \mathbb{D}^{k,r} \) and:

\[
D^k(FG) = \sum_{i=0}^{k} \binom{k}{i} D^i F \otimes D^{k-i} G, \tag{4.16}
\]

where the tilde denotes the canonical symmetrization (see the appendix).

We prove this theorem to illustrate an argument with tensor products. Multi-index notation for derivatives will be convenient\(^6\). We can also unambiguously denote \((|\alpha|!)^{-1} \sum_{\sigma} h_{\sigma(\alpha)}\) for the symmetrization of \( h_{i_1} \otimes \ldots \otimes h_{i_{|\alpha|}}\), where \((i_1, \ldots, i_{|\alpha|})\) consists of \( \alpha_1 \) ones, \( \alpha_2 \) twos and so on.

**Proof.** By continuity (using Hölder’s inequality), we may consider \( F, G \in \mathcal{S}_p \) in mutual standard form \( F = f(W(e_1), \ldots, W(e_n)) \), \( G = g(W(e_1), \ldots, W(e_n)) \), \( f, g \in C_p^\infty(\mathbb{R}^n) \), \( e_1, \ldots, e_n \in H \) orthonormal. In the following, we will drop explicit evaluation in \( (W(e_1), \ldots, W(e_k)) \). By the explicit representation (4.8), the usual Leibniz rule (noting symmetry when interchanging the order of differentiation):

\[
D^k(FG) = \sum_{i_1, \ldots, i_k=1}^{n} \partial_{i_1, \ldots, i_k} (fg) e_{i_1} \otimes \ldots \otimes e_{i_k} = \sum_{|\alpha|=k} \partial^\alpha (fg) |\alpha|! \frac{1}{|\alpha|!} \sum_{\sigma} \epsilon_{\sigma(\alpha)}
\]

\[
= \sum_{|\alpha|-k} \sum_{\beta \leq \alpha} \frac{1}{\beta! (\alpha - \beta)!} (\partial^\beta f) (\partial^{\alpha-\beta} g) \sum_{\sigma} \epsilon_{\sigma(\alpha)}.
\]

We have replaced \( e_{i_1} \otimes \ldots \otimes e_{i_k} \) by its symmetrization. \( \partial^\alpha (fg) \) occurs \( |\alpha|! / \alpha! \) times: there are \( |\alpha|! \) permutations of \((i_1, \ldots, i_n)\), but then each term in the sum is counted \( \alpha! \) times (for example, \( \partial_{11} \) occurs only once if \( k = 2 \)). On the other hand:

\[
\sum_{i=0}^{k} \binom{k}{i} D^i F \otimes D^{k-i} G
\]

\[
= \sum_{i=0}^{k} \sum_{|\beta|=i} \sum_{|\gamma|=k-i} \frac{1}{\beta! \gamma!} (\partial^\beta f) (\partial^\gamma g) \left[ \binom{k}{i} \frac{1}{i!} \sum_{\sigma_1} \left( \sum_{\sigma_2} \epsilon_{\sigma(1)} \right) \otimes \left( \sum_{\sigma_2} \epsilon_{\sigma(2)} \right) \right],
\]

\[
\sum_{\sigma} \sum_{\sigma_1} \sum_{\sigma_2} \epsilon_{\sigma(1)} \otimes \epsilon_{\sigma(2)} = |\beta||\gamma|! \sum_{\sigma'} \epsilon_{\sigma'(\beta+\gamma)}.
\]

So the term between square brackets coincides with \( \sum_{\sigma} \epsilon_{\sigma(\beta+\gamma)} \). We infer equality. \( \square \)

Note that in the definition of \( D^k \), derivatives are always a.s. symmetric. Namely, if \( F_n \in \mathcal{S}_p \) approximates \( F \in \mathbb{D}^{k,p} \), then we find a subsequence such that \( D^k F_{n_k} \to D^k F \) in \( H^{\otimes k} \) a.s. Thus:

\[
\langle F, h_{\sigma(1)} \otimes \ldots \otimes h_{\sigma(n)} \rangle_{H^{\otimes k}} = \langle F, h_1 \otimes \ldots \otimes h_n \rangle_{H^{\otimes k}},
\]

\(^6\)See Definition B.1.4.
by taking limit expressions with $F_{n_k}$ and using its symmetry. As usual, $\sigma$ is a permutation of $\{1, \ldots, n\}$ and $h_i \in H$.

Furthermore, an element $f \in H \otimes 2$ can be interpreted as a bounded linear operator:

$$f : H \to H; h \mapsto \langle f, h \rangle_H.$$  \hspace{1cm} (4.17)

$\langle f, h \rangle_H$ denotes $f \otimes_1 h$ or $\sum_{i \in \mathbb{N}} \langle f, h \otimes e_i \rangle_{H \otimes 2} e_i$, with an orthonormal basis $(e_i)_{i \in \mathbb{N}}$ for $H$ ($N$ countable). In the appendix, it was noted that this is well-defined by symmetry. The tensor product norm is strong. Namely:

$$\sum_{i=1}^N \langle f, e_i \otimes e_i \rangle_{H \otimes 2}^2 \leq \sum_{i,j=1}^N \langle f, e_i \otimes e_j \rangle_{H \otimes 2}^2 = \|f\|_{H \otimes 2}^2,$$  \hspace{1cm} (4.18)

by Parseval’s identity. The trace is square summable, so $f$ is a Hilbert-Schmidt operator.

The reason why we consider the completion of the differential operator is that we are interested in solving a differential equation (the Stein equation). $S_p$ is not large enough to contain all the solutions of our interest. The latter seem to be lying in holes “in between” $S_p$, comparable to $\sqrt{2}$, the positive solution of $x^2 = 2$, lying in between the rational numbers.

4.2.2 Derivatives for general Hilbert space-valued random variables

Let $V$ be another real separable Hilbert space with norm $\|\cdot\|_V$ and inner product $\langle \cdot, \cdot \rangle_V$. It is also convenient to have the above constructions for random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ with values in $V$. $\mathcal{F}$ is again generated by an isonormal Gaussian process $W$. The procedure is analogous. The spaces $L^p(\Omega; V)$ denote the random variables $X : \Omega \to V$ such that $\|X\|_{L^p(V)} := \mathbb{E}[\|X\|_V^p]^{1/p} < \infty$. For $p = 2$, this again is a Hilbert space. We consider the space of smooth random variables $S_V$ consisting of:

$$F = \sum_{i=1}^n F_i v_i, \quad F_i \in S_p, \quad v_j \in V.$$  \hspace{1cm} (4.19)

Then the derivative operators are defined as $D^k F = \sum_{i=1}^n D^k F_i \otimes v_i$, which belong to all $L^p(\Omega; H^\otimes k \otimes V)$, $p \in [1, \infty)$. Similar properties hold:

1. $D^k : S_V \subset L^p(\Omega; V) \to L^p(\Omega; H^\otimes k \otimes V)$ are closable for $p \in [1, \infty)$.

2. Therefore, $D^k$ can again be extended to $\mathbb{D}^{k,p}(V)$, the closure with respect to:

$$\|F\|_{\mathbb{D}^{k,p}(V)} = \left(\mathbb{E}[\|F\|_V^p] + \mathbb{E}[\|F\|_{H^\otimes k \otimes V}^p] + \ldots + \mathbb{E}[\|F\|_{H^\otimes k \otimes V}^p]\right)^{1/p}.$$  \hspace{1cm} (4.20)

3. 1. and 2. of Remark 4.2.8 have direct analogs here.

4. If $V = H^\otimes k$ and $F \in \mathbb{D}^{k+1,p}$, then $D(D^k F) = D^{k+1} F$. For $V = \mathbb{R}$, $D$ coincides with the definition of last section. So the notation is consistent.
### 4.2.3 The divergence operator

The dual $\delta^k$ of $D^k : \mathbb{D}^{k,2} \to L^2(\Omega; H^{\otimes k})$ will now be investigated. Recall section 2.1.

**Definition 4.2.12.** The divergence operator of order $k$, $\delta^k : \text{dom}(\delta^k) \subset L^2(\Omega; H^{\otimes k}) \to L^2(\Omega)$, where:

$$\text{dom}(\delta^k) := \{ u \in L^2(\Omega; H^{\otimes k}) \mid (\exists C > 0)(\forall F \in \mathcal{S}_p) \left( |E[\langle D^k F, u \rangle_{H^{\otimes k}}]| \leq C \|F\|_{L^2}\right) \} ,$$

is uniquely defined by:

$$E[F\delta(u)] = E[\langle D^k F, u \rangle_{H^{\otimes k}}], \quad F \in \mathcal{S}_p.$$  \hfill (4.21)

'F $\in \mathcal{S}_p$' may be replaced by 'F $\in D^k,2$' by extending the definitions continuously. For notational convenience, $\delta^0$ denotes the identity operator and $\delta := \delta^1$.

**Remark 4.2.13 (Basic properties of $\delta^k$).**

1. This operator is well-defined, by Riesz’ representation theorem, Theorem B.3.4. Namely, if $u \in \text{dom}(\delta^k)$, then $L^2(\Omega) \to \mathbb{R}; F \mapsto E[\langle D^k F, u \rangle_{H^{\otimes k}}]$.

   defines a continuous linear operator. Thus we obtain a representation as an inner product in $L^2(\Omega)$. Also note that dom($\delta^k$) is the largest domain possible: if a dual element $\delta^k(u) \in L^2(\Omega)$ exists, then it defines a continuous linear functional on $L^2(\Omega)$ by taking the inner product. As was already noted in section 2.1, $\delta^k$ is closed as the dual of a linear operator.

2. $E[\delta^k(u)] = 0$ for $k \geq 1$, by taking a constant $F$ in (4.21).

3. By (4.2.6), every $h \in H$ is included in dom($\delta$), with $\delta(h) = W(h)$.

4. Similarly for $h_1, \ldots, h_k \in H$, (4.12) yields that $f := h_1 \otimes \ldots \otimes h_k \in \text{dom}(\delta^k)$ with $\delta^k(f)$ a polynomial expression of order $k$ in $W(h_1), \ldots, W(h_k)$. Note that $\delta^k$ may also be defined for non-symmetric tensors. By continuity, Theorem 4.2.19, we will see that $H^{\otimes k} \subset \text{dom}(\delta^k)$.

We will now relate the domain of $\delta^k$ to the domain $\mathbb{D}^{k,p}$ of $D$ and outline some explicit computation rules and continuity properties for $\delta^k$. Many elements of $\mathbb{D}^{1,2}(H)$ can be computed conveniently:

**Proposition 4.2.14.** Let $F \in \mathbb{D}^{1,2}$, $u \in \text{dom}(\delta)$ and $F u \in L^2(\Omega; H^{\otimes k})$. Then $F u \in \text{dom}(\delta)$ and

$$\delta(Fu) = F\delta(u) - \langle DF, u \rangle_H,$$  \hfill (4.22)

given that the right hand side of (4.22) is square integrable.
Proof. Computing $\delta$ often comes down to extending (4.11) continuously. From Proposition 4.2.11, we have $D(FG) = GDF + FDG \in L^2(\Omega; H)$ for any $G \in S_b$.

$$\mathbb{E}[\langle DG, Fu \rangle_H] = \mathbb{E}[\langle FDG, u \rangle_H] = \mathbb{E}[\langle -GDF + D(FG), u \rangle_H]$$

$$= \mathbb{E}[G(F\delta(u) - \langle DF, u \rangle_H)],$$

where duality can be applied because $FG \in D^{1,2}$. Since the right hand side of formula (4.22) is square integrable, it defines a dual element for the $L^2$-norm. 

It can be seen that $S_H$ is contained in $\text{dom}(\delta)$ and we may compute for $F_i \in S_p$ and $h_i \in H$:

$$\delta \left( \sum_{i=1}^n F_i h_i \right) = \sum_{i=1}^n (F_i W(h_i) - \langle DF_i, h_i \rangle_H).$$

(4.23)

This can be used to infer a useful commutation relation between $D$ and $\delta$, termed the Heisenberg commutation relation in quantum mechanics:

Lemma 4.2.15 (Heisenberg commutation relation). For $u \in S_H$, $h \in H$:

$$\langle D\delta(u), h \rangle_H - \delta(D^h u) = \langle u, h \rangle_H.$$  

(4.24)

$D^h$ denotes the inner product of $h$ with the differentiated component, see (4.26).

Remark 4.2.16. This is often just denoted by $D\delta(u) - \delta(Du) = u$. To make it precise, a definition for $\delta$ in $L^2(\Omega; H^{0,2})$ is needed. Also, $Du$ may be asymmetric, so it is important to state which component to evaluate $h$ in. It was already remarked that $D^h$ represents a directional derivative.

Proof. By (4.23), and $u = \sum_i F_i h_i$ as before, $\delta(u) \in D^{1,2}$. By Proposition 4.2.11:

$$D\delta(u) = \sum_{i=1}^n (W(h_i)DF_i + F_i h_i - \langle D^2F_i, h_i \rangle_H),$$

(4.25)

with the convention of (4.17). On the other hand, again using (4.22):

$$D^h u = \sum_{i=1}^n \langle DF_i, h \rangle_H h_i,$$

(4.26)

$$\delta(D^h u) = \sum_{i=1}^n \delta(\langle DF_i, h \rangle_H h_i) = \sum_{i=1}^n (W(h_i)\langle DF_i, h \rangle_H - \langle D^2F_i, h \otimes h \rangle_{H^{0,2}}),$$

(4.27)

from which the result can be read off. 

This computation has many nice implications. For instance:
Theorem 4.2.17. \( \mathbb{D}^{1,2}(H) \subset \text{dom}(\delta) \) and the restricted operator,
\[
\delta|_{\mathbb{D}^{1,2}} : \mathbb{D}^{1,2}(H) \to L^2(\Omega); u \mapsto \delta(u),
\]  
(4.28)
is continuous. In particular for \( u, v \in \mathbb{D}^{1,2}(H) \):
\[
\mathbb{E}[\delta(u)\delta(v)] = \mathbb{E}[\langle u, v \rangle_H] + \mathbb{E}[\text{tr}(Du \circ Dv)].
\]  
(4.29)

Remark 4.2.18. Actually, it yields an isomorphism into \( L^2(\Omega) \), but a method to infer this uses some spectral properties that will not be explained further. Recall that the trace of an operator \( A : H \to H \), is defined as \( \text{tr}(A) := \sum_i \langle Ae_i, e_i \rangle_H \) if it exists for an orthonormal base \( (e_i)_{i \in \mathbb{N}}. \) \( Du \circ Dv \) denotes the composition of operators \( H \to H \), using (4.17).

Proof. Consider \( u, v \in \mathcal{S}_H \) and fix an orthonormal basis \( (e_i)_{i \in \mathbb{N}} \) for \( H \) (\( \mathbb{N} \) countable). In the previous proof, it was noted that \( \delta(v) \in \mathbb{D}^{1,2}. \)
\[
\mathbb{E}[\delta(u)\delta(v)] = \mathbb{E}[\langle u, D\delta(v) \rangle_H] = \sum_{i \in \mathbb{N}} \mathbb{E}[\langle u, e_i \rangle_H \langle v, e_i \rangle_H + \langle u, e_i \rangle_H \delta(D^e_i v)]
\]
\[
= \mathbb{E}[\langle u, v \rangle_H] + \mathbb{E}[\sum_{i \in \mathbb{N}} \langle D\langle u, e_i \rangle_H, D^e_i v \rangle_H].
\]

We have used the above commutation relation for the second equality. The last term is the expectation of the trace of the composition of operators. Taking \( u = v \) and using (4.18), we infer:
\[
\mathbb{E}[\delta(u)^2] \leq \mathbb{E}[\|u\|_{H}^2] + \mathbb{E}[\text{tr}(Du \circ Du)] \leq \|u\|_{\mathbb{D}^{1,2}(H)}^2.
\]  
(4.30)

We thus obtain the required continuity on \( \mathcal{S}_H \). By density, (4.29) can now be extended to \( \mathbb{D}^{1,2}(H) \): for any approximating sequence \( u_n \in \mathcal{S}_H \) for \( u \) in \( \mathbb{D}^{1,2}(H) \), it yields that \( (\delta(u_n))_{n \in \mathbb{N}} \) is Cauchy in \( L^2(\Omega) \). So \( u \in \text{dom}(\delta) \) and \( \delta(u_n) \to \delta(u) \).

More general;

Theorem 4.2.19 (Meyer inequalities). For any integers \( k \geq l \geq 1 \), \( \mathbb{D}^{k,2}(H^{\otimes l}) \subset \text{dom}(\delta^l) \) and we obtain continuity: there exist \( c_{k,l} \in \mathbb{R} \) such that
\[
\|\delta^k(u)\|_{\mathbb{D}^{k-l,2}} \leq c_{k,l}\|u\|_{\mathbb{D}^{k,2}(H^{\otimes l})}.
\]  
(4.31)

Remark 4.2.20. The proof may be found in [20], Proposition 1.5.7. It can be extended to continuity on \( \mathbb{D}^{k,p} \) spaces for any \( p \in [1, \infty) \).
4.2.4 Divergence of Hilbert space-valued variables

Again fix a real separable Hilbert space $V$. It is again possible to define divergences for Hilbert space-valued random variables as dual operators for the corresponding derivative operators. A general definition will however not be needed. Following [18], we only define them directly for deterministic vectors.

1. Namely, for $k \geq 1$:

$$\delta^k : V \otimes H^\otimes k \to L^2(\Omega; V) : \sum_{i=1}^{n} v_i \otimes h_i^{[k]} \mapsto \sum_{i=1}^{n} v_i \delta^k(h_i^{[k]}), \quad (4.32)$$

where $v_i \in V$, $h_i^{[k]} \in H^\otimes k$. We have only defined $\delta^k$ on finite linear combinations. It extends by continuity, Theorem 4.2.19.

2. For $f \in H^\otimes(k+l)$, $k, l$ integers, $\delta^k(f)$ may be defined as $\delta^k$ on $H^\otimes k \otimes V$, where $V = H^\otimes l$. Here, $\delta^k$ acts on the first $k$ components. For symmetric $f$, it is unambiguous. It is consistent for the general definition of Hilbert-valued divergences, meaning $\delta^{k+l}(f) = \delta^l(\delta^k(f))$. Actually, we will always symmetrize before evaluating. Otherwise, the results are not very meaningful.

Given these definitions, the commutation relation may be iterated:

**Proposition 4.2.21.** Let $k \in \mathbb{N}$, $u \in H^\otimes k$. Then $\delta^k(u) \in D^{1,2}$ and:

$$D\delta^k(u) = k\delta^{k-1}(u). \quad (4.33)$$

**Proof.** First, we have by (4.12) that tensors $f$ in an orthogonal base of $H^\otimes k$ have $\delta^k(f) \in \mathcal{S}_p$. By continuity, Theorem 4.2.19, it then follows that $\delta^k(u) \in D^{1,2}$. The commutation relation for $k = 1$ is proven in Lemma 4.2.15. Assume for $k \geq 2$ the validity of the formula for $k - 1$. Taking $h \in H$:

$$D^h\delta^k(u) = \delta(D^h\delta^{k-1}(u)) + \langle \delta^{k-1}(u), h \rangle_H$$

$$= (k-1)\delta(\langle \delta^{k-2}(u), h \rangle_H) + \langle \delta^{k-1}(u), h \rangle_H = k\langle \delta^{k-1}(u), h \rangle_H,$$

where we have used symmetry in the last step. \qed

**Remark 4.2.22.** Most $H^\otimes k$-valued random variables that we consider will be symmetric. For orthonormal $e_1$ and $e_2$ in $H$, we have with the above convention for $\delta(e_1 \otimes e_2)$ that it equals $W(e_1)e_2$. Since this is not symmetric, it can not be a multiple of $D\delta^2(e_1 \otimes e_2)$. 

73
4.3 The generator and Ohrnstein-Uhlenbeck process

To encode properties in all directions of $W$, derivatives were best described multidimensionally. This however does not allow to solve Stein equations for real-valued functions. Therefore, we consider the generator:

$$L : \text{dom}(L) \subset L^2(\Omega) \rightarrow L^2(\Omega); LF := -\delta(DF),$$

$$(4.34)$$

$\text{dom}(L) = \{F \in D^{1,2} | DF \in \text{dom}(\delta)\}.$

It has the advantage of being symmetric on $\text{dom}(L)$, i.e. for $F, G \in \text{dom}(L)$, Green’s formula holds: $E[L(F)G] = E[FL(G)]$. Also, compositions of the form $-\delta_Z \circ D$ are the natural form for many infinitesimal descriptions of evolution equations such as Schrödinger’s equation. In this way, they are often the generator of an associated Markov process. Now, what does it look like here?

**Proposition 4.3.1.** Let $F \in S_p$ have representation (4.4). Then $F \in \text{dom}(L)$ and:

$$LF = \sum_{i,j}^{n} \partial_{ij} f(W(h_1), \ldots, W(h_n))(h_i, h_j)_H - \sum_{i=1}^{n} \partial_i f(W(h_1), \ldots, W(h_n)W(h_i)$$

$$(4.35)$$

This is a direct consequence of the computation rule of Proposition 4.2.14. So $L$ is actually an infinite-dimensional partial differential operator. The Stein equation will then be $LF = v(w) - \mathbb{E}[v(W)]$, where $v : \mathbb{R}^H \rightarrow \mathbb{R}$ is measurable for the product $\sigma$-algebra of $\mathbb{R}^H$. For our purposes, we will restrict attention to functions of finitely many independent Gaussians. We have already noted in the one-dimensional case that $L_N = -\delta_N \circ D$ describes a Sturm-Liouville problem (SL), where $N \sim N(0, 1)$. It is singular due to an unbounded domain of integration $\mathbb{R}$. Namely,

$$Lf = \frac{(qf')'}{q}, \quad f \in \text{dom}(L_N),$$

$$$(4.36)$$

where $q(x) = e^{-x^2/2}/2\pi$. Note that zero-boundary conditions are fulfilled for solutions in $\text{dom}(L_N)$ as extension of $S_p$. A lot of structure of (one-dimensional) Sturm-Liouville problems can be inferred from the eigenfunctions, $Lf = \lambda f$, $\lambda \in \mathbb{R}$, even for many singular SL problems. Here, $f \neq 0$ is called an eigenfunction associated with the eigenvalue $\lambda$. Often, it exhibits the following properties:

1. $L$ has countably many simple real eigenvalues $\lambda_1 > \lambda_2 > \ldots \rightarrow -\infty$. Simple means that the eigenfunctions associated with $\lambda_n$ generate a one-dimensional vector space.

2. The normalized eigenfunctions form an orthonormal base for $L^2(q)$. It is understood, that one eigenfunction $f_n$ is chosen per eigenvalue $\lambda_n$. 

74
3. The eigenfunctions are oscillating and $f_n$ has $n - 1$ zero points.

Various behavior, classified by conditions on the endpoint of the domain of integration, can occur. See chapter 10 of [31] for more information. In our current case, it may be noted that $L$ describes an infinite-dimensional Sturm-Liouville problem that is isotropic in every direction (due to the condition of isonormality). Thus we might suspect similar properties as for the one-dimensional problem. We will not go deep into explicit Sturm-Liouville theory. Instead, we work the necessary theory out “by hand”. The reason why these properties are mentioned, is to raise attention to what may be suspected for general Stein functionals. As it turns out, the eigenfunctions for the normal distribution are the Hermite polynomials that we mentioned shortly before. Finally, together with a connection to semi-group theory, these properties prove to be very useful to infer regularity. More information on semi-groups may be found in [21] and [10].

4.3.1 Spectral decomposition of $L^2(\Omega)$

The structural Sturm-Liouville spectral decomposition is key for the strength of Malliavin calculus in Stein’s method. The eigenfunctions may be found by using the ladder operator $\delta$ (the name is due to Remark 4.3.3).

**Definition 4.3.2.** For $n \in \mathbb{Z}^+$, the $n$-th Hermite polynomial is defined as:

$$H_n : \mathbb{R} \to \mathbb{R}; H_n(x) = (-1)^n e^{x^2/2} \frac{d^n}{dy^n} \left( y \mapsto e^{-y^2/2} \right) \bigg|_{y=x}. \quad (4.37)$$

Also define $W_{-1} := 0$.

**Remark 4.3.3.** For $N \sim N(0,1)$, this corresponds to $H_n = \delta_N^n(1)$, as may be verified from (2.5) (the Stein kernel is 1, so $\delta_N = \delta_N$). In our current situation, it may also be verified from (4.12) that $H_n(W(h)) = \delta^n(h \otimes \ldots \otimes h)$, where we consider the $n$-fold tensor product of $h \in H$, $\|h\|_H = 1$.

In the following $\delta_{n,m} := \mathbb{1}_{n=m}$ denotes the Kronecker delta symbol and $0^0 := 1$. The properties of the one-dimensional Hermite polynomials can be read off by putting $H = \mathbb{R}$ and noting that the distribution of $W(e_1)$ and Lebesgue measure are mutually absolutely continuous (replace $W(e_1)$ by $x \in \mathbb{R}$).

**Proposition 4.3.4** (Properties of Hermite polynomials).

1. **Eigenfunctions:** For $n \geq 0$, $H_n' = nH_{n-1}$ and $LH_n(W(h)) = -nH_n(W(h))$.

2. **Moment generating function:** $e^{tx-t^2/2} = \sum_{k=0}^{\infty} H_k(x) \frac{t^k}{k!}$, $\forall t \in \mathbb{R}$.

3. If $X, Y \sim N(0,1)$ are jointly normal, then $\mathbb{E}[H_n(X)H_m(Y)] = n!\delta_{n,m}(\mathbb{E}[XY])^n$. 

75
4. The linear space generated by
\[ \left\{ H_n(W(h)) \mid n \in \mathbb{Z}^+, h \in H, \|h\|_H = 1 \right\} \] (4.38)
is dense in $L^p(\Omega)$ for any $p \in [1, \infty)$.

**Proof.** 1. By the commutation relation (4.33) and Remark 4.3.3, $DH_n(W(h)) = nH_{n-1}(W(h))$. Thus we have $H'_n = nH_{n-1}$ almost everywhere, and by continuity everywhere. From this, we also have $LH_n(W(h)) = -nH_n(W(h))$.

2. Note that:
\[
\begin{align*}
\frac{d^k}{dt^k} e^{tx-t^2/2} \bigg|_{t=0} &= \frac{d^k}{dt^k} e^{x^2/2-\frac{(t-x)^2}{2}} \bigg|_{t=x} = (-1)^k e^{x^2/2} \frac{d^k}{dt^k} e^{-t^2/2} \bigg|_{t=x} = H_k(x).
\end{align*}
\]
So, for every fixed $x$, we get uniform convergence over compacta for $t \in \mathbb{R}$ by known properties of Taylor series.

3. Since we have convergence over compacta for $t \in \mathbb{R}$ and $E[e^{tX}] < \infty$ for any $t \in \mathbb{R}$, we can differentiate under the integral sign to infer:
\[
E[H_n(X)H_m(Y)] = \frac{d^{n+m}}{dt^nds^m} E \left[ e^{(t-s)X+sY-s^2/2} \right] \bigg|_{t=0,s=0} = \frac{d^{n+m}}{dt^nds^m} E[e^{st}E[XY]] \bigg|_{t=0,s=0} = n! \delta_{m,n} (E[XY])^n.
\]

4. The linear combinations of $H_n(W(h))$, $n \in \mathbb{Z}^+$ contain all $W(h)^k$, $k \in \mathbb{Z}^+$. To infer density, we can again approximate Fourier transforms as in the proof of Lemma 4.2.3.

\[ \square \]

**Remark 4.3.5.** The Hermite polynomials also exhibit the following properties, for $n \geq 0$:

1. **Recurrence relation:** $H_{n+1}(x) - xH_n(x) = -nh_{n-1}(x)$.

2. With $[n/2]$ the largest integer smaller than $n/2$ (floor function):
\[ H_n(x) = n! \sum_{j=0}^{[n/2]} (-1)^j x^{n-2j} / j!(n-2j)!2^j. \] (4.39)

3. $H_n(-x) = (-1)^n H_n(x)$.

4. $H_{2n}(0) = (-1)^n (2n)!/2^n n!$

**Definition 4.3.6.** For $n \geq 0$, the $n$-th Wiener chaos is defined as:
\[ \mathcal{H}_n := \left\{ H_n(W(h)) \mid h \in H, \|h\|_H = 1 \right\} \subset L^2(\Omega). \] (4.40)
This denotes the closed linear subspace in $L^2(\Omega)$ generated by the given set. By point 3. and 4. of Proposition 4.3.4, it corresponds to the eigenspace of the eigenvalue $-n$ of $L$. 

76
It implies an orthogonal decomposition of $L^2(\Omega)$:

$$L^2(\Omega) = \mathcal{H}_0 \oplus \mathcal{H}_1 \oplus \mathcal{H}_2 \oplus \ldots$$  \hspace{1cm} (4.41)

This means that for any $F \in L^2(\Omega)$ there exist unique $F_n \in \mathcal{H}_n$ such that:

$$F = \mathbb{E}[F] + \sum_{n=1}^{\infty} F_n.$$  \hspace{1cm} (4.42)

$F_0 = \mathbb{E}[F]$ because the $F_n$ have mean zero for $n \geq 1$. As is known, the $F_n$ are the orthogonal projections of $F$ onto $\mathcal{H}_n$, or equivalently the best approximation $F_n = \text{argmin}_{G \in \mathcal{H}_n} \|F - G\|$. For the acquainted reader, we mention that Itô’s representation theorem relates to (4.42). We can refine this representation formula some more by taking into account the dimensions of $H$.

**Definition 4.3.7.** For $k \geq 1$, $f \in H^\otimes k$, the $k$-th multiple integral of $f$ is defined as $I_k(f) := \delta^k(f)$.

For $l \leq k$ and $f \in H^\otimes k$, $I_l(f)$ can be defined as for the Hilbert space-valued divergences (the divergence acts on the first $l$ components). The multiple integrals make up the Wiener chaos:

**Lemma 4.3.8** (Properties of multiple integrals). For integers $k, l \geq 1$, $f \in H^\otimes k$, $g \in H^\otimes l$:

1. **Differentiability:** $I_k(f) \in \mathbb{D}^{\infty,2}$ and:

   $$D^l I_k(f) = \mathbb{1}_{l \leq k} \frac{k!}{(k-l)!} I_{k-l}(f).$$  \hspace{1cm} (4.43)

2. **Isometry property:**

   $$\mathbb{E}[I_k(f)I_l(g)] = \delta_{k,l}k!(f,g)_{H^\otimes k}.$$  \hspace{1cm} (4.44)

3. **Relation to Wiener chaos:** We get an isometric isomorphism:

   $$I_k : H^\otimes k \rightarrow \mathcal{H}_k; f \mapsto I_k(f).$$  \hspace{1cm} (4.45)

We stress again that it is important that the tensors are symmetric.

**Proof.**

1. By the iterated commutation relation, (4.33), we have that $I_k(f) \in \mathbb{D}^{1,2}$ and the above formula holds for $l = 1$. We may then iterate the argument, yielding $I_k(f) \in \mathbb{D}^{\infty,2}$ and the formula for all $l \geq 1$.

2. Suppose $k \geq l$. Due to point 1., we infer:

   $$\mathbb{E}[I_k(f)I_l(g)] = \mathbb{E}[\delta^k(f)\delta^l(g)] = \mathbb{E}[(f, D^k I_l(g))_{H^\otimes k}] = \delta_{k,l}k!(f,g)_{H^\otimes k}.$$  \hspace{1cm} (4.44)
3. In Remark 4.3.3 it was noted that $H_n(W(h)) = I_n(h \otimes \ldots \otimes h)$ if $\|h\|_H = 1$. Because $I_n$ is an isometry into $L^2(\Omega)$ by 2. and linearity, the $n$-th Wiener chaos is contained in the image of $I_n$. Consider now an orthonormal base $(e_i)_{i \in N}$ for $H$ and a multi-index $\alpha \in (\mathbb{Z}^+)^k$ for some $k \geq 1$. Then it can also be computed with (4.12) that

$$I_k \left( (|\alpha|!)^{-1} \sum_{\sigma} e_\sigma(\alpha) \right) = H_{\alpha_1}(W(e_1)) \ldots H_{\alpha_k}(W(e_k)).$$

Because the $n$-th Wiener chaos is closed, it contains the image of $I_k$. \hfill \Box

**Remark 4.3.9.**
1. In Corollary 4.3.19 we will see that the $L^p$-norms are equivalent in every fixed Wiener chaos. So the Lemma also implies that $I_k(f) \in \mathbb{D}^{\infty,p}$ for any $p \in [1, \infty)$.

2. In the case of Gaussian measures, Example 4.1.3, the $I_k(f)$ coincide with multiple Wiener-Itô integrals. See sections 1.1–1.3 in [20] or 2.7.1 in [18] for more information. Hence the name multiple integral.

This implies the following expansion:

**Corollary 4.3.10 (Stroock formula).**
1. For any $F \in L^2(\Omega)$, there exist unique $f_k \in H^{\otimes k}, k \geq 1$ such that:

$$F = \mathbb{E}[F] + \sum_{k=1}^{\infty} I_k(f_k) \quad \text{in } L^2(\Omega). \tag{4.46}$$

2. Let $l \geq 1$. $F \in \mathbb{D}^{l,2}$ if and only if $\sum_{k=l}^{\infty} k^l k! \|f_k\|_{H^{\otimes k}}^2 < \infty$. Then:

$$D^l F = \sum_{k=l}^{\infty} \frac{k!}{(k-l)!} I_{k-l}(f_k). \tag{4.47}$$

3. If $F \in \mathbb{D}^{k,2}$, then $f_k = \frac{1}{k!} \mathbb{E}[D^k F]$.

3. gives a computation rule for the expansion.

**Proof.** 1. is a combination of the Wiener chaos decomposition of $L^2(\Omega)$ and part 3. of Lemma 4.3.8.

For 2., if $\sum_{k=l}^{\infty} k^l k! \|f_k\|_{H^{\otimes k}}^2 < \infty$, we can approximate $F$ in $L^2(\Omega)$ by $\mathbb{E}[F] + \sum_{k=1}^{n} I_k(f_k)$ for $n \to \infty$, while the $l$-th derivatives form a Cauchy sequence in $L^2(\Omega; H^{\otimes l})$. Indeed, let $n \geq m \geq l$:

$$\mathbb{E} \left[ \left\| D^l \left( \sum_{k=m}^{n} I_k(f_k) \right) \right\|_{H^{\otimes l}}^2 \right] = \mathbb{E} \left[ \left\| \sum_{k=m}^{n} \frac{k!}{(k-l)!} I_{k-l}(f_k) \right\|_{H^{\otimes l}}^2 \right] \leq \sum_{k=m}^{n} \left( \frac{k!}{(k-l)!} \right)^2 (k-l)! \|f_k\|_{H^{\otimes k}}^2 \tag{4.48}$$

78
We have used the isometry property (4.44).

On the other hand, suppose \( F \in \mathbb{D}_l^2 \). \( (\mathcal{H}_n(\mathcal{H}^{\otimes l}))_{n \geq 0} \) form an orthogonal decomposition of \( L^2(\Omega; H^{\otimes l}) \), where \( \mathcal{H}_n(\mathcal{H}^{\otimes l}) \) is the closure in \( L^2(\Omega; H^{\otimes l}) \) of variables of the form \( \sum_{i=1}^m F_i g_i, F_i \in \mathcal{H}_n, g_i \in H^{\otimes l} \). Indeed, \( S(\mathcal{H}^{\otimes l}) \) is dense and we may approximate this space by the \( \mathcal{H}_n(\mathcal{H}^{\otimes l}) \). Thus we find unique \( g_k \in \mathcal{H}_k(\mathcal{H}^{\otimes l}) \) such that:

\[
D_l^k F = E[D_l^k F] + \sum_{k=1}^{\infty} g_k, \quad \text{in} \quad L^2(\Omega; H^{\otimes l}).
\]

Because \( D_l \) maps \( \mathcal{H}_{n+l} \) on \( \mathcal{H}_n(H^{\otimes l}) \) and \( \delta^l \) maps it back, \( g_k = D_l I_{k+l}(f_{k+l}) \). The isometry property yields the converse.

3. can be read off from the formula in 2.: \( E[D_l^k F] = l! f_l \).

Before moving on, we will need an expansion for products of multiple integrals:

**Lemma 4.3.11 (Product formula).** \( k, l \geq 1, f \in H^{\otimes k}, g \in H^{\otimes l} \). Then:

\[
I_k(f)I_l(g) = \sum_{r=0}^{k\wedge l} r! \binom{k}{r} \binom{l}{r} I_{k+l-2r}(f^{\otimes r}g).
\]

(4.49)

Recall the definition of the symmetrized contraction \( \tilde{\otimes}_r \) from the appendix. The proof combines Stroock’s formula (4.46) with the Leibniz rule, Proposition 4.2.11.

### 4.3.2 The Ornstein-Uhlenbeck semigroup

We have constructed generators with stationary distribution \( \mathbb{P} \circ W^{-1} \), i.e. \( E[Lf(W)] = 0 \) for all measurable \( f : \mathbb{R}^H \to \mathbb{R} \) such that \( f(W) \) is in \( \text{dom}(L) \). A spectral decomposition was investigated to infer more structure. As an illustration:

**Proposition 4.3.12 (Spectral decomposition of the generator).** Let \( F \in L^2(\Omega) \) have the form (4.46). Then \( F \in \text{dom}(L) \) if and only if:

\[
\sum_{k=1}^{\infty} p^2 E[I_k(f_k)^2] < \infty.
\]

(4.50)

In this case, \( LF = -\sum_{k=1}^{\infty} k I_k(f_k) \).

On Wiener chaos of order \( n \), this representations is valid. Then, it follows by Corollary 4.3.10. There is yet another useful connection to infer regularity of Stein solutions. Semi-group theory allows representations in terms of a semi-group of operators. On top of that, it can often be given a stochastic representation by a Markov process. The famous Feynman-Kac formula and Feynman integrals arise in this way. In Gaussian analysis the connection with semi-groups can be given without too much technicalities. The Markov process corresponds to an infinite-dimensional isonormal
Ornstein-Uhlenbeck process. However, we will only use the marginals of the distribution. For a construction of the whole Markov Process, see Proposition 1.4.1 in [20] or Remark 2.8.5 in [18].

For the isonormal Gaussian process \( W \), fix an independent copy \( W' = \{W'(h) \mid h \in H\} \) on the product probability space \((\Omega \times \Omega', \mathcal{F} \otimes \mathcal{F}', \mathbb{P} \times \mathbb{P}')\). The dashed symbols indicate copies of the non-dashed objects. For ease of consideration, we assume that \( W \) is measurable with respect to the first component of this measure space and \( W' \) with respect to the second. In this way we get for any measurable \( f : \mathbb{R}^H \to \mathbb{R} \) that:

\[
\mathbb{E}[f(W)] = \mathbb{E}'[f(W')], \quad \mathbb{E}[f(W')] = f(W') \quad \text{a.s.}
\]

It is well-known that if \( F \) is a measurable function with respect to \( \sigma(W(h) \mid h \in H)\), then a measurable \( f : \mathbb{R}^H \to \mathbb{R} \) may be found such that \( F = f(W) \) a.s. It is \( \mathbb{P} \circ W^{-1} \)-a.s. unique.

**Definition 4.3.13.** The Ornstein-Uhlenbeck semi-group \((P_t)_{t \in [0, \infty)}\) is defined on \( L^1(\Omega) \) by Mehler’s formula:

\[
P_t : L^1(\Omega) \to L^1(\Omega); F = f(W) \mapsto \mathbb{E}' \left[ f \left( e^{-t}W + \sqrt{1 - e^{-2t}}W' \right) \right]. \quad (4.51)
\]

**Remark 4.3.14.** A comment is in order on whether it is well-defined. First, for \( \mathbb{P} \)-a.s. every \( \omega \in \Omega, \omega' \mapsto f \left( e^{-t}W(\omega) + \sqrt{1 - e^{-2t}}W'(\omega') \right) \) is \( \mathcal{F} \)-measurable. Next, for any Borel \( A \subset \mathbb{R}^H \):

\[
(\mathbb{P} \times \mathbb{P}') \left[ e^{-t}W + \sqrt{1 - e^{-2t}}W' \in A \right] = \mathbb{P} \left[ W \in A \right]. \quad (4.52)
\]

In other words, they have the same distribution. This may be seen as follows. The distribution is determined by \((W(e_i))_{i \in N}\) \((N \subset \mathbb{N} \) countable), for an orthonormal base \((e_i)_{i \in N}\) of \( H \). Denote the projection of \( A \) on the components \((e_1, \ldots, e_n)\) as \( A_n \). Then, it is sufficient to show that for all \( n \in \mathbb{N} \):

\[
(\mathbb{P} \times \mathbb{P}') \left[ \left( e^{-t}W(h) + \sqrt{1 - e^{-2t}}W'(h) \right)_{h \in \{e_1, \ldots, e_n\}} \in A_n \right] = \mathbb{P} \left[ (W(h))_{h \in \{e_1, \ldots, e_n\}} \in A_n \right],
\]

by taking limits (decreasingness of probability measures). Then, because the components are independent and for standard normal random variables \( N \) and \( N' \), \( e^{-t}N + \sqrt{1 - e^{-2t}}N' \) again has a standard normal distribution, these equalities are valid. This implies:

\[
\mathbb{E} \left[ \mathbb{E}' \left[ f \left( e^{-t}W + \sqrt{1 - e^{-2t}}W' \right) \right] \right] = \mathbb{E}[f(W)] < \infty, \quad (4.53)
\]

by Fubini. Thus, the inner expectation is finite \( \omega \)-a.s. We conclude that \( P_t(f(W)) \in L^1(\Omega) \) by the property \( |\mathbb{E}[F]| \leq \mathbb{E}[|F|] \) for \( \mathcal{F} \)-measurable \( F \).

Finally, \((P_t)_{t \in [0, \infty)}\) indeed forms a semi-group. \( P_0 \) is the identity operator and \( P_s \circ P_t = P_{s+t} \) for \( s, t \geq 0 \). The latter may be inferred by fixing another independent copy \( W'' \) of \( W \), and again using an argument of equality in distribution as (4.52). For any \( v : \mathbb{R}^H \to \mathbb{R} \) we then have that \( v(W) - \mathbb{E}[v(W)] = P_0(v(W)) - P_\infty(v(W)) \), where \( P_\infty(F) = \lim_{t \to \infty} P_t(F) \).
We may also consider $P_t : L^p(\Omega) \to L^p(\Omega)$, $p \in [1, \infty)$, which is continuous:

**Proposition 4.3.15** (Linear contraction). For $t \geq 0$, $p \in [1, \infty)$, $\|P_t\|_{L^p(\Omega) \to L^p(\Omega)} \leq 1$.

**Proof.** We just have to use Hölder’s inequalities and (4.52) for $f(W) \in L^p(\Omega)$:

\[
E[|P_t f(W)|^p] = E\left[ E'\left( f\left( e^{-t}W + \sqrt{1-e^{-2t}}W'\right) \right) \right]^p 
\leq E\left[ E'\left( f\left( e^{-t}W + \sqrt{1-e^{-2t}}W'\right) \right) \right] = E[|f(W)|^p].
\]

\[\square\]

**Remark 4.3.16.** We may again define $P_t$ on $L^p(\Omega; H^\otimes k)$ by first considering $S^\otimes k_H$ (vector-valued integrals) and then extending continuously. Most properties carry over.

**Theorem 4.3.17** (Spectral decomposition of $P_t$). For $F \in L^2(\Omega)$ with Wiener chaos decomposition as in (4.46):

\[
P_t(F) = \sum_{k=0}^{\infty} e^{-kt}I_k(f_k).
\] (4.54)

From this formula, we see that the norm decays if $F$ is not a constant. This formula may be proven by showing that the identity holds for $F$ of the form $e^{W(h)-\|h\|^2/2}$, $h \in H$, noting point 2. of 4.3.4. That is because the span of these variables is dense in $L^2(\Omega)$ (again use the Hahn-Banach theorem in symbiosis with moment generating functions, Lemma 4.2.3). Due to the analogy with previous proofs, we do not outline it here.

By this representation on Wiener chaos, we can directly see that $L$ is the infinitesimal generator of the Ornstein-Uhlenbeck process. That is,

\[
\lim_{t \to 0^+} \frac{P_tF - F}{t} = LF \quad \text{in } L^2(\Omega), \quad F \in \text{dom}(L).
\]

The decay of the norms can be quantified in a particularly strong **hypercontractivity** property:

**Theorem 4.3.18** (Nelson). For $F \in L^p(\Omega)$, $p \in (1, \infty)$ and any $t \geq 0$:

\[
E\left[ |P_tF|^{1+e^{2t(p-1)}}\right]^{1/(1+e^{2t(p-1)})} \leq E[|F|^{p}]^{1/p}.
\] (4.55)

See [18] for a proof. Also, continuity can not be better, by inspecting $F = \exp(\lambda W(h))$, $\|h\|_H = 1$. $\|P_tF\|_{L^p(\Omega)} = \exp(\lambda^2(r e^{-2t} + 1 - e^{-2t})/2)$, $\|F\|_{L^p(\Omega)} = \exp(\lambda^2 p/2)$. For $r > 1 + e^{2t(p-1)}$, take $\lambda \to \infty$ to see that $\|P_tF\|_{L^p(\Omega)} / \|P_tF\|_{L^p(\Omega)} \to \infty$.

As promised, this implies that the $L^p$-norms are equivalent in every fixed Wiener chaos:
Corollary 4.3.19. Let $F \in \mathcal{H}_n$, $n \geq 1$ and $r > p > 1$. Then

$$
\mathbb{E}[|F|^p]^{1/p} \leq \mathbb{E}[|F|^r]^{1/r} \leq \left(\frac{r-1}{p-1}\right)^{n/2} \mathbb{E}[|F|^p]^{1/p}.
$$

(4.56)

The second inequality is just a consequence of Nelson's theorem and the explicit representation of $P_t$ on Wiener chaos: $\|P_t I_n(f)\|_{L^r(\Omega)} = e^{-nt} \|F\|_{L^r(\Omega)}$.

Now that we have developed the basis of the language of Malliavin calculus, we can start to exploit it in full strength.

4.4 Inversion of $L$ and integration by parts

Consider again the Stein equation $LF = v(W) - \mathbb{E}[v(W)]$, where $v \in \mathcal{H}$, a suitable function class as outlined at the very beginning of this text. Since $\ker(L)$ only consists of constants (note the spectral decomposition), the range has codimension one [27]. Furthermore, we are interested in solutions with zero-boundary conditions. A spectral formula is direct:

**Definition 4.4.1** (Spectral decomposition of $L^{-1}$). The pseudo-inverse of $L$ is defined as:

$$
L^{-1} : L^2(\Omega) \rightarrow L^2(\Omega); \sum_{k=0}^{\infty} I_k(f_k) \mapsto -\sum_{k=1}^{\infty} \frac{1}{k} I_k(f_k),
$$

(4.57)

where we consider Stroock’s formula (4.46).

The definition was extended to the whole space by “ignoring constant terms”. From the spectral representation, we infer for $F \in L^2(\Omega)$ that $L^{-1}F \in \text{dom}(L)$ and $LL^{-1}F = F - \mathbb{E}[F]$.

We first state an integration by parts formula in the new terminology. It will be fundamental to applications in Stein’s method, because it allows to compare derivatives directly.

**Proposition 4.4.2** (Integration by parts). Let $F, G \in \mathbb{D}^{1,2}$ and $g \in C^1(\mathbb{R})$ with bounded derivative. Then:

$$
\mathbb{E}[F g(G)] = \mathbb{E}[F] \mathbb{E}[g(G)] + \mathbb{E}[g'(G) \langle DG, -DL^{-1}F \rangle_H].
$$

(4.58)

**Proof.** By the chain rule, Theorem 4.2.9, it is seen that $g(G) \in \mathbb{D}^{1,2}$ and $Dg(G) = g'(G) DG$. We just have to use duality with $L = -\delta_D$:

$$
\mathbb{E}[(F - \mathbb{E}[F]) g(G)] = \mathbb{E}[(LL^{-1}F) g(G)] = \mathbb{E}[(Dg(G), -DL^{-1}F)_H] = \mathbb{E}[g'(G) \langle DG, -DL^{-1}F \rangle_H].
$$

\[\square\]
It is nothing really new, since this coincides with the previous integration by parts formula. If for example $F = W(h)$, $\|h\|_H = 1$, the spectral decomposition yields $L^{-1}W(h) = -W(h)$, $DW(h) = h$ and so:

$$\mathbb{E}[W(h)g(W(h))] = \mathbb{E}[g'(W(h))],$$

gives integration by parts for a standard normal distribution.

The spectral formula for $L^{-1}$ is not convenient to infer regularity properties from. Here, semi-group theory comes to the rescue. It is often possible to interpolate:

$$F_v := \int_0^\infty (P_\infty[v(W)] - P_t[v(W)])dt. \quad (4.59)$$

First note that the quantity inside the integral is a random variable. Since the integration domain is unbounded, more properties need to be used to show that the integral converges. First, $P_\infty[v(W)] = \mathbb{E}[v(W)]$, $P_0[v(W)] = v(W)$. Then, if everything acts nicely, we would expect:

$$LF_v = \int_0^\infty \left[ -L P_t[v(W)] dt = P_0[v(W)] - P_\infty[v(W)] = v(W) - \mathbb{E}[v(W)]. \right] \quad (4.60)$$

In regular cases, geometric ergodicity holds however, meaning that $d_{TV}$ or $d_W$ decay exponentially (see [10] for more information). In the next chapter we will check convergence by hand for $v : \mathbb{R}^d \to \mathbb{R}$ Lipschitz with bounded derivative. Also, the representation allows to let certain operations act on the “most regular part”, as in the following Lemma (which is not a heuristic):

**Lemma 4.4.3.** For $F \in \mathbb{D}^{1,2}$ with $\mathbb{E}[F] = 0$, we obtain:

$$-DL^{-1}F = \int_0^\infty e^{-t}P_t[DF]dt = -(L - I)^{-1}DF. \quad (4.61)$$

So we are able to put the derivative on $F$. Also note that by the spectral decomposition, it may be seen that $L - I$, where $I$ is the identity operator, is injective.

**Proof.** Technicalities with convergence in (4.59) can be evaded by spectral decomposition. If $F = I_k(f_k)$ for $f_k \in H^{\odot k}$ and $k \geq 1$, then the formula is direct: $-DL^{-1}F = D(1/k)F = I_{k-1}(f_k)$. Also:

$$\int_0^\infty e^{-t}P_t[DF]dt = \int_0^\infty e^{-t}e^{-(k-1)t}kI_{k-1}(f_k)dt = I_{k-1}(f_k).$$

And finally: $(L - I)^{-1}DF = \frac{k}{-k(k-1)}I_{k-1}(f_k)$. Then the formulas extend by linearity to linear combinations and by $\mathbb{D}^{1,2} \to L^2(\Omega, H)$ continuity of the three formulas to every $F \in \mathbb{D}^{1,2}$. For example:

$$\mathbb{E} \left[ \left\| \int_0^\infty e^{-t}P_t[DF]dt \right\|_H^2 \right] \leq \int_0^\infty e^{-t}\mathbb{E}[\|P_tDF\|_H^2]dt$$

$$\leq \int_0^\infty e^{-t}\mathbb{E}[\|DF\|_H^2]dt = \mathbb{E}[\|DF\|_H^2],$$

83
for which Jensen’s inequality and the contraction property of $P_t$ were used (keeping in
mind Remark 4.3.16).

Before concluding this chapter, we note an important property that states that $\mathbb{E}[\|DF\|^2_H]$ is actually sufficient to describe the zero-boundary Sobolev functions $\mathbb{D}^{1,2}$ with zero mean (instead of $\|\cdot\|_{\mathbb{D}^{1,2}}$):

**Theorem 4.4.4** (Poincaré inequality). Let $F \in \mathbb{D}^{1,2}$. Then $\text{var}[F] \leq \mathbb{E}[\|DF\|^2_H]$.

*Proof.* We may pick an interpolation or spectral argument as a favorite. For the latter: if $F = I_k(f_k), f_k \in H^\otimes k, k \geq 1$, then $\text{var}[F] = k!\|f_k\|^2_H = \mathbb{E}[\|DF\|^2_H]$ by the isometry property (4.44). By the orthogonal decomposition, both sides may be extended linearly. Finally, they extend continuously with respect to $\|\cdot\|_{\mathbb{D}^{1,2}}$.

We finally mention that variance expansions can also be given in terms of derivatives, by using the spectral decomposition and an interpolation argument. [18] for example states for $F \in \mathbb{D}^{\infty,2}$:

$$\text{var}(F) = \sum_{k=1}^{\infty} \frac{1}{k!} \mathbb{E}[\|D^k F\|^2_{H^\otimes k}],$$

(4.62)

which is a direct consequence of developing $F$ in the Sturm-Liouville eigenbasis (the Wiener chaoses here, but similar formulas can be obtained for other distributions). If $\mathbb{E}[\|D^k F\|^2_{H^\otimes k}]/k! \to 0$, this can be strengthened to:

$$\text{var}(F) = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k!} \mathbb{E}[\|D^k F\|^2_{H^\otimes k}].$$

(4.63)
Chapter 5
Gaussian approximation

In this chapter, we sketch some of the interesting results for which the Mallaivin-Stein symbiosis may be used. Before, methods based on moments and cumulants used to be an important tool for Gaussian approximation. These are based on showing for a sequence $(f_n(W))_{n \in \mathbb{N}}$ that all moments or all cumulants converge to the respective values of a normal distribution. The new tools show that for fixed Wiener chaoses an equivalence actually exists based on the first four moments. Using the same tools, it is then possible to infer approximation results for general functionals. In the end, they allow to study dependence that is carried through a whole sequence of random variables for central limit theorems. This may be contrasted to results as Theorem 1.1.12. It is reported in [18] that these new tools mostly improve on the method of moments and cumulants. The reader is also referred to [18] for complete arguments on the results. First, we illustrate interpolation to infer Stein approximations.

5.1 Multivariate normal approximation

We emphasized that the interpolation formula (4.59) enables to obtain regularity much more efficiently. The only technicality that remains is the verification of integrability assumptions. We are now in an excellent position to illustrate this for $d$-dimensional normal distributions, $d \in \mathbb{N}$. Let $A = (A_{ij}), B = (B_{ij}) \in \mathbb{R}^{d \times d}$. Recall from the previous chapter the Hilbert-Schmidt norm, which actually extends to an inner product:

$$\langle A, B \rangle_{HS} := tr(AB^T) = \sum_{i,j=1}^{d} A_{ij} B_{ij}. \quad (5.1)$$

Furthermore, let $\Sigma \in \mathbb{R}^{d \times d}$ be a positive semi-definite matrix. Then it may first of all be verified that if $(N_1, \ldots, N_d) \sim N(0, \Sigma)$ is multivariate normal distributed (with Definition 4.1.1, even if $\Sigma$ is not strictly positive definite):

$$\mathbb{E}[N_i f(N_1, \ldots, N_k)] = \sum_{j=1}^{d} \Sigma_{ij} \mathbb{E}[\partial_j f(N_1, \ldots, N_k)], \quad (5.2)$$
for \( i \in \{1, \ldots, d\} \) and \( f \in C^1(\mathbb{R}^d) \) with bounded derivatives. Indeed, if \( \Sigma \) is strictly positive definite, it follows from one-dimensional integration by parts as before (chapter 1). If it is singular, than we may first infer the formula with \( \Sigma^{(2)} := \Sigma + \epsilon I, I \in \mathbb{R}^{d \times d} \) the identity matrix and \( \epsilon > 0 \). Then let \( \epsilon \to 0 \) and use distributional convergence. It is possible to put these results together nicely by going one order higher, to the generator \( L \). For all \( f \in C^2(\mathbb{R}^d) \) with bounded first and second derivative, we have

\[
\mathbb{E}[(N, \nabla f(N))_{\mathbb{R}^d}] = \mathbb{E}[(\Sigma, \text{Hess} f(N))_{HS}],
\]

where we denote the Hessian matrix \( \text{Hess} f(N) = (\partial_{ij} f)_{i,j=1,\ldots,d} \). Indeed, just replace \( f \) by \( \partial_i f \) in (5.2). This is the basis for a Stein characterization:

**Theorem 5.1.1** (Multivariate normal Stein characterization). Let \( \Sigma = (\Sigma_{ij})_{i,j=1,\ldots,d} \) be positive semi-definite and consider an \( \mathbb{R}^d \)-valued random variable \( X = (X_1, \ldots, X_d) \) with finite first moment. Then \( X \sim N(0, \Sigma) \) if and only if

\[
\mathbb{E}[(X, \nabla f(X))_{\mathbb{R}^d}] = \mathbb{E}[(\Sigma, \text{Hess} f(X))_{HS}],
\]

for any \( f \in C^2(\mathbb{R}^d) \) with bounded first and second derivative.

**Proof.** The ‘if’ direction was already shown. One way to proceed for the other direction is reverting the problem to one-dimensional Stein characterizations (Remark 1.1.8): take \( g \in C^1_b(\mathbb{R}) \) and \( k = (k_1, \ldots, k_d) \in \mathbb{R}^d \). Define \( G(z) = \int_0^z g(y)dy \) and \( f(x_1, \ldots, x_d) = G(k_1 x_1 + \ldots + k_d x_d) \). (5.3) reads:

\[
\mathbb{E}[(k_1 X_1 + \ldots + k_d X_d) g(k_1 X_1 + \ldots + k_d X_d)] = k^T \Sigma k \mathbb{E}[g'(k^T X)].
\]

Note that these \( g \) encompass all solutions of the corresponding Stein equation for indicator functions if \( k^T \Sigma k \neq 0 \) (by rescaling to Lemma 1.1.1), and \( X \) must be 0 if \( k^T \Sigma k = 0 \). So we obtain that \( k^T C \sim N(0, k^T \Sigma k) \). Definition 4.1.1 then gives \( X \sim N(0, \Sigma) \). \( QED \)

Note that (5.3) coincides with \( \mathbb{E}[L f(N_1, \ldots, N_d)] = 0 \), Proposition 4.3.1, if encoded with \( h \in \mathbb{R}^d \) (see the proof of Proposition 5.1.2). An approach that is more apt to generalization may be based on interpolation techniques or techniques from Markov process theory to find unique stationary measures of \( L \).

### 5.1.1 Standard normal distributions

For a \( d \)-dimensional standard normal distributions \( N \sim N(0, I) \), the previous Stein characterization gives rise to a Stein equation:

\[
\triangle g(x) - \langle x, \nabla g(x) \rangle_{\mathbb{R}^d} = h(x) - \mathbb{E}[h(N)], \quad x \in \mathbb{R}^d.
\]

\( \triangle \) denotes the Laplace operator \( \partial_{11} + \ldots + \partial_{dd} \) and \( h : \mathbb{R}^d \to \mathbb{R} \in L^1(N) \). We would like to solve it with \( f \in C^2 \) if \( h \) is continuous or a weakly differentiable \( f \) if \( h \) is not. Also, we are interested in the solution with zero-boundary conditions (\( \in \mathcal{D}^{1,2} \)). We show the promised interpolation formula:
Proposition 5.1.2. For \( h \in \text{Lip}(\mathbb{R}^d) \), a solution to (5.4) is given by the \( C^2(\mathbb{R}^d) \)-function

\[
g_h(x) = \int_0^\infty \mathbb{E}[h(N) - h(e^{-t}x + \sqrt{1-e^{-2t}}N)]dt,
\]

and we have

\[
\sup_{x \in \mathbb{R}^d} \|\text{Hess}g_h(x)\|_{HS} \leq \sqrt{d}\|h\|_{Lip}.
\]

Proof. As stated, the proof just comes down to checking a string of integrability properties if there are no general properties available yet for (4.60). First note that the term in the integral is integrable:

\[
\left| h(N) - h(e^{-t}x + \sqrt{1-e^{-2t}}N) \right| \leq \|h\|_{Lip} \left( e^{-t}\|x\|_{\mathbb{R}^d} + (1 - \sqrt{1-e^{-2t}})\|N\|_{\mathbb{R}^d} \right),
\]

because \( (1 - \sqrt{1-e^{-2t}}) = e^{-2t}/(\sqrt{1-e^{-2t}} + 1) \leq e^{-2t} \) for \( t \geq 0 \). Next, it is possible to differentiate under the integral sign and expectation, since the terms are integrable (noting again Theorem B.3.3), \( i \in \{1, \ldots, d\} \):

\[
\partial_i g_h(x) = - \int_0^\infty e^{-t}\mathbb{E}[(\partial_i h)(e^{-t}x + \sqrt{1-e^{-2t}}N)]dt
\]

\[
= - \int_0^\infty \frac{e^{-t}}{\sqrt{1-e^{-2t}}}\mathbb{E}[h(e^{-t}x + \sqrt{1-e^{-2t}}N)N_i]dt,
\]

where the last equality follows from integration by parts. This reasoning is very important, it allows to exploit the regularity where there is room. Because now it is possible to differentiate again (\( \int_0^\infty e^{-2t}/\sqrt{1-e^{-2t}}dt < \infty \)):

\[
\partial_{ij} g_h(x) = - \int_0^\infty \frac{e^{-2t}}{\sqrt{1-e^{-2t}}}\mathbb{E}[(\partial_{ij} h)(e^{-t}x + \sqrt{1-e^{-2t}}N)N_i]dt
\]

\[
= - \int_0^\infty \frac{e^{-2t}}{1-e^{-2t}}\mathbb{E}[h(e^{-t}x + \sqrt{1-e^{-2t}}N)H_{ij}(N)]dt,
\]

\( j \in \{1, \ldots, n\}, \ H_{ij}(N) = N_iN_j \) for \( i \neq j \) and \( H_{ii}(N) = N_i^2 - 1 \) the corresponding multiple integral or multi-dimensional Hermite-polynomial. The last step is again an application of integration by parts. It allows to infer that \( g_h \in C^2(\mathbb{R}^d) \). For the Stein approximation, take \( B \in \mathbb{R}^{d \times d} \) and use the first representation:

\[
\|(\text{Hess}g_h(x), B)\|_{HS} = \int_0^\infty \frac{e^{-2t}}{\sqrt{1-e^{-2t}}}\mathbb{E}\left[\langle B^T N, \nabla h(e^{-t}x + \sqrt{1-e^{-2t}}N)\rangle_{\mathbb{R}^d}\right] dt
\]

\[
\leq \left( \sup_{x \in \mathbb{R}^d} \|\nabla h(x)\|_{\mathbb{R}^d} \right) \mathbb{E}[\|B^T N\|_{\mathbb{R}^d}] \int_0^\infty \frac{e^{-2t}}{\sqrt{1-e^{-2t}}dt}.
\]

The Cauchy-Schwarz inequality was used in the last step. Now \( \int_0^\infty e^{-2t}/\sqrt{1-e^{-2t}}dt = (1/2)\int_0^1 (1 - x)^{-1/2}dx = 1, \ \|\nabla h(x)\|_{\mathbb{R}^d} \leq \sqrt{d}\|h\|_{Lip} \) and:

\[
\mathbb{E}[\|B^T N\|_{\mathbb{R}^d}]^2 \leq \mathbb{E}[\|B^T N\|_{\mathbb{R}^d}^2] = \mathbb{E} \left[ \sum_{i=1}^d \left( \sum_{j=1}^d B_{ij}N_i \right)^2 \right] = \|B\|_{HS}^2.
\]

87
Putting everything together yields \(|\langle \text{Hess} g_h(x), B \rangle_{HS} \| \leq \sqrt{d} \| h \|_{Lip} \| B \|_{HS}|. Taking B = \text{Hess} g_h(x) then gives (5.6). Now it need only be checked that it is a solution. Encoding \((N_1, \ldots, N_d) = (W(e_1), \ldots, W(e_n))\) for \(e_i\) the standard basis vector of \(\mathbb{R}^d\) and \(W\) an isonormal Gaussian encoded by \(H = \mathbb{R}^d\), we have that \(f_h(N) \in \text{dom}(L)\): the Hessian is bounded, thus \(f_h\) and \(\partial_i f\) are at most of polynomial growth, implying \(f \in \mathcal{D}_{1,2}, DF \in \mathcal{D}_{1,2}(H)\). By extending Proposition 4.3.1 continuously:

\[
\Delta g_h(N) - \langle N, \nabla g_h(N) \rangle_{\mathbb{R}^d} = L g_h(N) = -\int_0^\infty L P_t h(N) dt = h(N) - \mathbb{E}[h(N)].
\]

We have used (4.60), after it has been noted that \(L\) and the integral can be interchanged by integrability of (5.7) and (5.8). Because \(N\) is mutually absolutely continuous with the Lebesgue measure and \(g_h\) is continuous, the equality holds for every \(x \in \mathbb{R}^d\) instead of \(N\). 

### 5.1.2 Non-singular normal distributions

By rescaling, the regularity of Stein solutions of general non-singular normal distributions can be inferred. We just state the result. For a worked out version of the rescaling argument, we refer to [18].

Let \(N \sim N(0, \Sigma)\) and \(\Sigma \in \mathbb{R}^{d \times d}\) strictly positive definite. The Stein equation now reads for \(h : \mathbb{R}^d \to \mathbb{R}\) with \(\mathbb{E}[h(N)] < \infty\):

\[
\langle \Sigma, \text{Hess} g(x) \rangle_{HS} - \langle x, \nabla g(x) \rangle_{\mathbb{R}^d} = h(x) - \mathbb{E}[h(N)], \quad x \in \mathbb{R}^d.
\]

And the Stein approximation becomes:

**Proposition 5.1.3.** For \(h \in \text{Lip}(\mathbb{R}^d)\), a solution to (5.9) is given by the \(C^2(\mathbb{R}^d)\) function

\[
g_h(x) = \int_0^\infty \mathbb{E}[h(N) - h(e^{-t} x + \sqrt{1 - e^{-2t}} N)] dt,
\]

and we have:

\[
\sup_{x \in \mathbb{R}^d} \| \text{Hess} g_h(x) \|_{HS} \leq \sqrt{d} \| h \|_{Lip} \| \Sigma^{-1} \|_{\mathbb{R}^d \to \mathbb{R}^d} \| \Sigma \|_{\mathbb{R}^d \to \mathbb{R}^d}^{1/2}.
\]

\(\| \cdot \|_{\mathbb{R}^d \to \mathbb{R}^d}\) denotes the strong operator norm (Definition B.1.6).

### 5.2 Normal approximation in a new language

For the rest of this chapter, assume that all random variables are considered on a probability space \((\Omega, \mathcal{F}, \mathbb{P})\) where \(\mathcal{F}\) is generated by an isonormal Gaussian process \(W\) indexed by the Hilbert space \(H\).

We now come to the fundamental part where Stein bounds are reformulated in the language of Malliavin calculus for one- and multi-dimensional normal approximation. We get the following approximation property:
Proposition 5.2.1. Let $F \in \mathbb{D}^{1,2}$ with $\mathbb{E}[F] = 0$, $\mathbb{E}[F^2] = 1$. Furthermore consider $f \in \text{Lip}(\mathbb{R})$ and assume that either $F$ has a density with respect to Lebesgue measure or that $f \in C^1(\mathbb{R})$. Then $\mathbb{E}[f'(F)]$ is well defined and:

$$|\mathbb{E}[f'(F)] - \mathbb{E}[Ff(F)]| \leq \|f\|_{\text{Lip}} \mathbb{E}|1 - \langle DF, -DL^{-1}F \rangle_H|.$$  \hspace{1cm} (5.12)

If moreover $F \in \mathbb{D}^{1,4}$, then $\langle DF, -DL^{-1}F \rangle_H$ is square integrable and we can estimate

$$\mathbb{E}|1 - \langle DF, -DL^{-1}F \rangle_H| \leq \sqrt{\text{var}[(DF, -DL^{-1}F)_H]}.$$  \hspace{1cm} (5.13)

If $f \in C^1$, it is just a consequence of the integration by parts formula, Proposition 4.4.2. By convoluting with a Dirac approximation, we can then approximate any Lipschitz function by $C^\infty$ functions with bounded derivative. The second part is a consequence of Lemma 4.4.3:

$$\mathbb{E}[(DF, -DL^{-1}F)_H^2] \leq \sqrt{\mathbb{E}[\|DF\|_H^4] \sqrt{\mathbb{E}[\|DL^{-1}F\|_H^4]}},$$  \hspace{1cm} (5.14)

by using Cauchy-Schwarz inequality twice. By the contraction property of $P_t$, we then get as in the proof of Lemma 4.4.3:

$$\mathbb{E}[\|DL^{-1}F\|_H^4] \leq \mathbb{E} \left[ \left\| \int_0^\infty e^{-t}P_tDFdt \right\|_H^4 \right] \leq \mathbb{E}[\|DF\|_H^4].$$

Because $\mathbb{E}[\langle DF, -DL^{-1}F \rangle_H] = 1$ by again using integration by parts, the inequality of the Proposition is just another application of Cauchy-Schwarz.

In this way, we retrieve:

Proposition 5.2.2. Let $F \in \mathbb{D}^{1,2}$ with $\mathbb{E}[F] = 0$ and $\mathbb{E}[F^2] = \sigma^2 > 0$ and $N \sim N(0, \sigma^2)$. Then:

$$d_W(F, N) \leq \frac{\sqrt{2} \sigma}{\sigma^2} \mathbb{E}|\sigma^2 - \langle DF, -DL^{-1}F \rangle_H|.$$  \hspace{1cm} (5.15)

Remark 5.2.3. Similar formulas can also be proven for the total variation and Kolmogorov distance if $F$ has a density with respect to Lebesgue measure. To practically bound the right hand side, (5.13) may be used.

Proof. This is just a consequence of the previous proposition together with Lemma 1.1.3 and Lemma 1.3.10.

A multi-dimensional statement based on the previous section reads:

Proposition 5.2.4. Let $d \geq 2$ and $F = (F_1, \ldots, F_d)$, $F_i \in \mathbb{D}^{1,4}$ with $\mathbb{E}[F_i] = 0$ for $i = 1, \ldots, d$. If $\Sigma \in \mathbb{R}^{d \times d}$ is positive definite and $N \sim N(0, \Sigma)$:

$$d_W(F, N) \leq \sqrt{d} \|\Sigma^{-1}\|_{\mathbb{R}^d} \|\Sigma\|_{\mathbb{R}^d}^{1/2} \left( \sum_{i,j=1}^d \mathbb{E} \left[ (\Sigma_{ij} - \langle DF_j, -DL^{-1}F_i \rangle_H)^2 \right] \right)^{1/2}.$$  \hspace{1cm} (5.16)
When there is interest in the particular bounds on the Zolotarev-type distances and \( \text{var}(F) \) is difficult to estimate, it is possible to obtain bounds which only show \( \Sigma \) within \( (\Sigma_{ij} - \langle DF_j, -DL^{-1}F_i \rangle_H)^2 \). This can again be bounded by a fourth moment if all \( F_i \in \mathbb{D}^{1.4} \). One may find these results in chapters 5 and 6 of [18] for comparisons on twice absolutely continuous \( h \) with bounded second derivative (Theorem 6.1.2 for example). On the other hand, it is possible to consider singular \( \text{var}(F) \) then.

### 5.3 Normal approximation for Gaussian functionals

With the language of Malliavin calculus and the Stein bounds from the previous chapter, we are left to calculate the latter in practical instances. It is by no means trivial to put everything together however. For Wiener chaoses \( F \in \mathcal{H}_k, L^{-1}F = -k^{-1}F \) and the results of the previous section translate to:

**Proposition 5.3.1.** For \( k \in \mathbb{N}, F \in \mathcal{H}_k, \mathbb{E}[F^2] = \sigma^2 > 0 \) and \( N \sim N(0, \sigma^2) \):

\[
d_W(F, N) \leq \frac{1}{\sigma} \sqrt{\text{var} \left( \frac{2}{k \pi} \| DF \|_H^2 \right)} \leq \frac{1}{\sigma} \sqrt{\frac{2k - 2}{3k \pi} \left( \mathbb{E}[F^4] - 3\sigma^4 \right)}. \tag{5.16}
\]

For analogous results in the total variation or Kolmogorov distance, one first needs to show that the multiple integrals have a density with respect to Lebesgue measure. The second inequality is based on the following important calculations (due to the product and isometry rules):

**Lemma 5.3.2.** For \( k \geq 2, F = I_k(f), f \in H^\otimes k \):

\[
\text{var} \left( \frac{1}{k} \| DF \|_H^2 \right) = \frac{1}{k^2} \sum_{r=1}^{k-1} r^2 \frac{k}{r} \left( 2k - 2r \right)! \| f \otimes_r f \|_{H^\otimes (2k - 2r)}^2, \tag{5.17}
\]

\[
\mathbb{E}[F^4] - 3\mathbb{E}[F^2]^2 = \frac{3}{k} \sum_{r=1}^{k-1} r^2 \frac{k}{r} \left( 2k - 2r \right)! \| f \otimes_r f \|_{H^\otimes (2k - 2r)}^2 \tag{5.18}
\]

\[
= \sum_{r=1}^{k-1} k^4 \left( \frac{k}{r} \right)^2 \left( \| f \otimes_r f \|_{H^\otimes (2k - 2r)}^2 + \left( \frac{k - 2r}{k} \right) \| f \otimes_r f \|_{H^\otimes (2k - 2r)}^2 \right).
\]

\[
\mathbb{E}[\| D^2F \otimes_1 D^2F \|_{H^\otimes 2}^2] \leq k^4 (k - 1)^4 \sum_{r=1}^{k-1} (r - 1)^2 \left( \frac{k - 2}{r - 1} \right)^4 (2k - 2r)! \| f \otimes_r f \|_{H^\otimes (2k - 2r)}^2. \tag{5.20}
\]

Namely, it gives the important estimates:

\[
\text{var} \left( \frac{1}{k} \| DF \|_H^2 \right) \leq \frac{k - 1}{3k} \left( \mathbb{E}[F^4] - \mathbb{E}[F^2]^2 \right) \leq (k - 1) \text{var} \left( \frac{1}{k} \| DF \|_H \right). \tag{5.21}
\]
For every random variable, it may be shown that $\mathbb{E}[F^4] - 3\mathbb{E}[F^2]^2 \geq 0$ and that this is 0 for a normal distribution. Together with the other calculations of Lemma 5.3.2 and taking into account Corollary 4.3.19, it is now possible to obtain the fundamental theorem:

**Theorem 5.3.3** (Fourth moment theorem). For a sequence of random variables $F_n = I_k(f_n)$, $f_n \in \mathcal{H}^k$, $k \geq 2$ such that $\mathbb{E}[F_n^2] \to \sigma^2 > 0$ for $n \to \infty$, the following are equivalent for $n \to \infty$:

1. $F_n \to N(0, \sigma^2)$ in distribution.
2. $\mathbb{E}[F_n^4] \to 3\sigma^4$.
3. $\text{var} \left( \|DF_n\|_{H}^2 \right) \to 0$.
4. $\|f \otimes_r f\|_{H^2(2k-2r)}^2 \to 0$ for all $r = 1, \ldots, k - 1$.
5. $\|f \otimes_r f\|_{H^2(2k-2r)}^2 \to 0$ for all $r = 1, \ldots, k - 1$.

4. and 5. give a practical way to test for distributional convergence from the Hilbert space encoding scheme. Because the bounds in (5.16) are uniform in $k$, point 2. above also implies for a sequence $(F_n)_{n \in \mathbb{N}}$ ranging over different Wiener chaoses that the distribution converges to a normal one. This notion is delicate however: it is not possible for variables in a fixed sum of Wiener chaoses to converge in probability to a normally distributed random variable [18].

In a similar spirit, it is possible to derive a so-called second-order Poincaré inequality. It is based on an iteration of the normal Poincaré inequality, Theorem 4.4.4. First, similar calculations lead to the following bound:

**Lemma 5.3.4.** For $F, G \in \mathbb{D}^{2,4}$ with $\mathbb{E}[F] = \mathbb{E}[G] = 0$:

$$
\mathbb{E}\left| \mathbb{E}[FG] - \langle DF, -DL^{-1}G \rangle_{H} \right| \leq \mathbb{E}[\|D^2F\|^4_{H \to H}]^{1/4} \mathbb{E}[\|DG\|^4_{H}]^{1/4} + \frac{1}{2} \mathbb{E}[\|D^2G\|^4_{H \to H}]^{1/4} \mathbb{E}[\|DF\|^4_{H}]^{1/4}.
$$

Here, one has to recall the convention (4.17). Putting this together with Proposition 5.2.1:

**Theorem 5.3.5** (Second order Poincaré inequality). For $F \in \mathbb{D}^{2,4}$ such that $\mathbb{E}[F] = 0$, $\mathbb{E}[F^2] = \sigma^2 > 0$ and $N \sim N(0, \sigma^2)$:

$$
d_W(F, N) \leq \frac{3}{\sqrt{2\pi \sigma^2}} \mathbb{E}[\|D^2F\|^4_{H \to H}]^{1/4} \mathbb{E}[\|DF\|^4_{H}]^{1/4}. \quad (5.22)
$$

And a random contraction inequality holds:

$$
\mathbb{E}[\|D^2F\|^4_{H \to H}]^{1/4} \leq \mathbb{E}[\|D^2F \otimes_1 D^2F\|^2_{H \otimes 1}]. \quad (5.23)
$$
Armed with this inequality, it is possible to investigate distributional convergence to a normal distribution for a centered sequence \((F_n)_{n \in \mathbb{N}}\) in \(D^2\), such that \(\text{var}(F_n) \to \sigma^2 > 0\). \(DF\) should be tested for boundedness. It should not converge to 0 because then \(F\) is expected to approach a constant. Then it should be tested that \(D^2F_n\) converges to 0. This means that the fluctuations of \(F_n\) around a normal distribution should not be too large. Indeed, if \(F_n = f(N)\), \(N\) a normal distribution, then \(D^2F = f''(N)\). If this tends to zero, then \(F\) is expected to tend to \(aN + b\) for \(a, b \in \mathbb{R}\). See [19] for an application to a CLT for linear functionals of Gaussian subordinated fields.

As another illustration of how the above results may be used, a nice application can be found in [13]. There it is proven that the parameter \(\theta\) of a Langevin dynamics
\[
dX_t = -\theta X_t dt + \sigma dB_t, \quad (B_t)_{t \in \mathbb{R}^+}\text{ a (fractional) Brownian motion}
\]
can be estimated by a so-called least squares estimator \(\hat{\theta}_T\), namely that it converges a.s. to \(\theta\) and a CLT holds for \(T \to \infty\):
\[
\hat{\theta}_T = \frac{\int_0^T X_t dX_t}{\int_0^T X^2_t dt}.
\]

### 5.4 Breuer-Major theorem

We want to mention a last application of the new language, because it is an excellent illustration of how dependence can be dealt with. By carrying out an analogous program as the last section for multi-dimensional Wiener chaoses, conditions of a CLT can be formulated in terms of the Stroock expansion, Corollary 4.3.10. Consider a sequence \((F_n)_{n \in \mathbb{N}}\) in \(L^2(\Omega)\), with \(\mathbb{E}[F_n] = 0\, f_{n,k} \in H^\otimes k, k \geq 1:\)
\[
F_n = \sum_{k=1}^{\infty} I_k(f_{n,k}). \tag{5.24}
\]

**Theorem 5.4.1** (CLT via chaos decomposition). If there exist \(\sigma_k^2 \geq 0\) such that for \(n \to \infty\):

1. \(k!\|f_{n,k}\|^2_{H^\otimes k} \to \sigma_k^2\), for all \(k \geq 1\).
2. \(\sigma^2 := \sum_{k=1}^{\infty} \sigma_k^2 < \infty\).
3. \(\|f_{n,k} \otimes_r f_{n,k}\|_{H^\otimes(2k-2r)} \to 0\) for \(k \geq 2, r = 1, \ldots, k - 1\).
4. \(\lim_{M \to \infty} \sup_{n \geq 1} \sum_{k=M+1}^{\infty} k!\|f_{n,k}\|^2_{H^\otimes k} = 0\).

Then \(F_n \to N(0,\sigma^2)\) in distribution for \(n \to \infty\).

3. may be verified with the help of Theorem 5.3.3. 4. indicates that the \(F_n\) may be approximated uniformly by Wiener chaoses on \(L^2(\Omega)\). This allows to investigate partial sums of a stationary sequence:
\[
S_n = \frac{1}{\sqrt{n}} \sum_{l=1}^{n} f(X_l), \tag{5.25}
\]
1. \( f \in L^2(N), N \sim N(0,1) \) has an expansion \( f(x) = \sum_{k=1}^{\infty} a_k H_k(x) \) with Hermite rank

\[
R := \inf \{ k \geq 1 | a_1 = \ldots = a_{k-1} = 0, a_k \neq 0 \}.
\]

Such an expansion exists by Proposition 4.3.4 if \( f \) is non-zero and \( \mathbb{E}[f(N)] = 0 \).

2. \((X_k)_{k \in \mathbb{Z}}\) is a centered stationary sequence with unit variance. Stationary means that \((X_{k+l})_{k \in \mathbb{Z}} \sim (X_k)_{k \in \mathbb{Z}}\) for all \( l \in \mathbb{Z} \), i.e., the distribution is shift-invariant. The correlation function is defined as \( \rho(l) = \mathbb{E}[X_0 X_l] \). In particular, \( X_k \sim N(0,1) \).

With the chaos decomposition, it is possible to obtain:

**Theorem 5.4.2 (Breuer-Major).** If \( \sum_{l \in \mathbb{Z}} |\rho(l)|^R < \infty \), then \( S_n \to N(0,\sigma^2) \) in distribution for \( n \to \infty \), where:

\[
\sigma^2 := \sum_{k=R}^{\infty} k! a_k^2 \sum_{l \in \mathbb{Z}} \rho(l)^k < \infty.
\]

In this way, weak conditions are obtained on the correlation function for a CLT to hold. It may be compared to CLTs for stationary ergodic Markov chains, see [10]. Note that the two sided condition only relies on one because \( \rho(l) = \rho(-l) \) by stationarity. It may be applied with \( X_i \) the increments of a fractional Brownian motion [18].
Chapter 6
Analysis and conclusion

Given random variables $X$ and $Z$, it has been illustrated that Stein’s method is often able to successfully estimate

$$\mathbb{E}[h(X)] - \mathbb{E}[h(Z)], \quad h \in C,$$

(6.1)
in two stages. It uses one variable to construct a Sturm-Liouville transform that translates the properties of $C$, or some regular subset $\mathcal{H}$, well. Then it remains to estimate a functional based on the distribution of $X$. In this way, the distributions of $X$ and $Z$ do not need to be directly compared over $C$.

It was already observed in the one-dimensional case that Stein’s reasoning for the normal distribution could be replicated well for other distributions. The calculations with density functions did not make clear exactly how. Malliavin calculus offers a complete framework that sheds much light on the necessary constructions. It combines the following. First, an encoding scheme that efficiently describes families of normal variables. Then, the construction of an operator that has a specific distribution as its only invariant distribution (Stein characterization). Finally an analysis of its inverse in terms of spectral decomposition and interpolation. Some properties of Gaussians were exploited, but it would however seem much can be replicated for other distributions.

Efficient encoding scheme

Gaussian processes are not the only interesting processes that arise in stochastic models. They represent the central distributions for diffusion processes, having continuous paths. For models with jumps, such as population models, Poisson jump processes are central. Here, the Poisson process resembles an analog of Brownian motion. The definition of a Gaussian process based on characteristic functions, Definition 4.1.1, does not readily translate to other families of distributions. However, Malliavin calculus is based on the idea Wiener-Itô integrals, related to Example 4.1.3. In this text they have been used abstractly as multiple integrals $I_k(f_k)$. Analogs exist for jump processes based on intensity measures. See chapter 1 of [11] for an introduction and more references. ‘Isotropic schemes’ are useful to extrapolate the Sturm-Liouville theory to multiple dimensions.
Sturm-Liouville transformation

With the ideas from section 2.1, a linear operator $D$ that describes the local behavior of a distribution can be dualized with respect to $Z$ to obtain $\delta$. This can be readily replicated for a large class of probability structures. Because the construction depends on the distribution itself, some properties may need to be investigated again. Also, dualization depends on the encoding scheme so that different spaces may be mapped to each other (such as $L^2(\Omega) \rightarrow L^2(\Omega; H)$ before). This can be resolved by considering $L = -\delta \circ D$. In all instances, it had the distribution of $Z$ as a unique stationary measure, meaning that a Stein characterization holds:

$$\mathbb{E}[Lf(X)] = 0, \forall f \in \text{dom}(L) \iff X \sim Z,$$

by comparing to all distributions $X$ which use similar information as $Z$, such as having the same state space with nested support of the distribution, or being measurable with respect to a specific $\sigma$-algebra.

What makes Malliavin calculus most efficient is the exploitation of deeper theory of PDE. Inverting $L$ on $\mathcal{H}_\perp$ gave rise to a Sturm-Liouville problem:

$$Lf = h - \mathbb{E}[h(Z)], \text{ under zero-boundary conditions.}$$

These zero-boundary conditions could be formulated as finding the solution $f$ in a completed space with respect to $C^\infty_0(\mathbb{R}^n)$, such as $\mathbb{D}^{1,2}$. They consist of the most regular functions and give rise to Poincaré inequalities. In section 4.3 it was commented that in a plethora of situations, it is “easier” to first understand the eigenfunctions $Lf = \lambda f$. They give an orthogonal decomposition of $L^2(\Omega)$ and because $L$, $D$ and $\delta$ do not “explode” on them (they are continuous with continuous inverse), much more structure may be inferred.

Furthermore, using a representation for solutions in terms of a semi-group allows to translate regularity more efficiently. This is because here only small local changes need to be investigated and regularity can be stored where there is room, as in the proof of Proposition 5.1.2. Also, there was an interpretation in terms of a Markov process possible for the Gaussian families. [21] and [10] could be investigated for generalization.

The above two enable much more strength to infer regularity of Stein solutions. The decomposed subspaces may first be investigated (giving rise to the fourth moment theorem), and then extensions to the full space can be sought, such as in the CLTs based on Stroock’s formula. This successful combination tackled many Gaussian approximation problems (see the webpage of Ivan Nourdin for an overview).

Yet, the above two methods are not restricted to Gaussian alone. Sturm-Liouville theory is more widely available, as discussed at the beginning of section 4.3. Schoutens outlines decompositions for Pearson family of distributions in [24]. In this case the eigenfunctions happen to be orthogonal polynomials. Moreover, he also mentions certain duality results, such as Farvard’s theorem, to replicate the situation for discrete
distributions of Ord’s family. There, both situations also allow a semi-group representation and interpretation in terms of Markov processes. However, a specific tilt of the Stein equation was used. When investigating the forms of the eigenfunctions of

\[ L(g) = \frac{(qTg')'}{q} = T(x)g''(x) + (\mu - x)g'(x) = \lambda g(x), \]

(6.2)

with \( T \) the Stein kernel of a distribution with Lebesgue-density \( q \), the eigenfunctions seem to coincide with \( g_n = q^{-1}\delta^n(qT^n) \).

In trying to generalize, one easily runs into some technicalities such as the tilts. However, Poincaré inequalities, variance expansions and such can be expected. In this way, a similar program could perhaps be partially replicated for Poisson jump process approximation (which may already be studied in the literature).

Theory is drawn from various related mathematical subjects, sometimes stated in different language and notation. The following relations were summarized:

Stein’s method starts from the view of a measure \( \mu \) and constructs a differential \( L \) having \( \mu \) as a stationary measure. Theory of PDE mostly starts from \( L \). The above interrelations are however also central to Markov process theory and stochastic analysis. This was illustrated with some references throughout the text.

It was raised that Stein’s method has the advantage of offering quantitative estimates in stochastic approximation. Multiple related distances could be considered, section 1.3. There were some illustrations of how it led to results where dependence could be taken care of, such as Theorem 1.1.12 and Theorem 5.4.2. Finally, it allows much insight into classes of random variables. Namely, the decomposition of Gaussian functionals into Wiener chaos expansions enables to infer general approximation results.
Appendix A

Nederlandstalige samenvatting

Een belangrijke overweging bij het wiskundig modelleren van wereldfenomenen is de robuustheid van het model. In stochastische modellen leidt dit tot vraagstukken van stochastische approximatie. Hierbij wordt de overeenkomstigheid tussen het gedrag van twee toevallige veranderlijken $X$ en $Z$ op een kansruimte $(\Omega, \mathcal{F}, \mathbb{P})$ vaak gemeten aan de hand van de volgende statistieken:

$$E[h(X)] - E[h(Z)], \quad h \in \mathcal{C}. \quad (A.1)$$

$\mathcal{C}$ is hierin een klasse meetbare functies $S \to \mathbb{R}$ die het gedrag van de kansmaten bepaalt, waarbij $X$ en $Z$ hun waarden nemen in een ruimte $S$. De indicator functies $\mathcal{C} = \{1_{(\infty, x]} | x \in \mathbb{R}\}$ geven voor $S = \mathbb{R}$ aanleiding tot het vergelijken van de distributiefuncties. Als voor $\mathcal{C}$ de continuë functies op $S$ beschouwd worden, geeft dit aanleiding tot zogenaamde zwakke convergentie. Wanneer de statistieken (A.1) dicht bij 0 liggen wordt verwacht dat $X$ en $Z$ een gelijkaardig gedrag of een gelijkvaardige verdeling hebben. Populaire methoden die hierin inzicht winnen zijn gebaseerd op genererende en karakteristieke functies (Fouriertransformatie). Lévy’s continuïteitsstelling geeft zo aanleiding tot de centrale limietstelling die fundamenteel is voor het gebruik van normaalverdelingen in vele wetenschappen. Centrale limietstellingen vormen een belangrijk thema doorheen deze tekst. Voor gestandardizeerde steekproeven,

$$\sum_{i=1}^{n} \frac{X_i - E[X_i]}{\sqrt{n\sigma_n^2}}, \quad (A.2)$$

formuleren ze voorwaarden waaronder de verdeling een standaard normale distributie benadert voor grote waarden van $n$. Hierin is $\sigma_n^2 = \text{var}(\sum_{k=1}^{n} X_i)$.

In de jaren ’70-’80 ontwikkelde Charles Stein een alternatieve methode om de statistieken (A.1) te begrenzen in het geval $Z$ een normaalverdeling heeft. Hiervoor wordt de structuur van een partiële differentiaalvergelijking benut. Ze laat eerst en vooral toe andere waarschijnlijkheidsstructuren te beschouwen waarin afhankelijkheden kunnen optreden. Bovendien geeft ze in tegenstelling tot Fouriertransformatie een kwantitatief beeld over (A.1). Verder zijn verschillende metrische structuren op de toevallige veranderlijken gelijkaardig te behandelen, zoals beide keuzes voor $\mathcal{C}$ hierboven.
Daar komt nog bij dat de methode succesvol repliceerbaar bleek voor andere distributies dan een normaalverdeling. Hiermee verkrijgt men inzicht in (centrale) limietstellingen, populatiemodellen, de invloed van zogenaamde 'prior'-distributies in Bayesi-anse statistiek maar ook bijvoorbeeld het gedrag van schatters voor diffusieprocessen. Het is echter niet dadelijk duidelijk hoe en waarom Stein’s methode er precies in slaagt (A.1) in alle gevallen goed te begrenzen. Daarom is het hoofddoel van dit werk dit te onderzoeken. Namelijk, hoe werkt Stein’s methode? Wanneer is het voordelig ze te beschouwen? En hoe kan ze vertaald worden naar andere situaties?

Eerst worden één-dimensionale discrete en absoluut continu verdelingen bestudeerd. Een aantal van de voorvermelde gebruikssituaties wordt verkend en unificerende taal gepromoot. We bestuderen kort hoe de statistieken (A.1) alternatief kunnen berekend worden. Op de besproken voorbeelden worden soms lichte uitbreidingen verkend. Voor Gaussiaanse analyse, de studie van functionalen van Gaussische processen, is er een succesvol raamwerk beschikbaar dat geworteld zit in de taal van ‘Malliavin calculus’. Er wordt een grondige analyse gemaakt van haar argumenten om erna de cruciale mechanismen uit te kristalliseren. Zo pogen we een algemeen beeld over Stein’s methode te formuleren. We schetsen eveneens een aantal van de interessante resultaten waarin de symbiose met ‘Malliavin calculus’ resulteert.

Zoals vermeld maakt de methode gebruik van een gepaste differentiaal- of differ- entievergelijking. Hierbij is een centraal probleem de regulariteit te bepalen van een klasse oplossingen. Eén-dimensionaal kan men gebruik maken van berekeningen met densiteiten, maar er blijkt een grote winst mogelijk door spectraaleigenschappen en semi-groeprepresentaties te benutten. Dit laat toe stochastische variabelen onder te verdelen in reguliere deelruimten waaruit sterke approximatieresultaten kunnen afgeleid worden. Zo wordt bijvoorbeeld een Breuer-Major resultaat bekomen. Deze formuleert een centrale limietstelling voor identiek verdeelde, maar niet onafhankelijke functies van normaal verdeelde veranderlijken. Bovendien stellen we vast dat bovenstaande verbanden zich niet hoeven te beperken tot Gaussiaanse analyse.

De beste manier om Stein’s methode te leren kennen, is door ze in actie te zien. De lezer wordt daarom hartelijk uitgenodigd een kijkje te nemen in dit werk.
Appendix B

A summary of important notions and results

B.1 Some conventions

We state that $\mathbb{N} = \{1, 2, \ldots\}$ is the set of natural numbers. When it is preferable to include 0, this will be indicated by $\mathbb{Z}^+ = \{0, 1, 2, \ldots\}$, the set of non-negative integers.

Definition B.1.1. Given a probability space $(\Omega, \mathcal{F}, P)$, we denote the completion of $\mathcal{F}$ as

$$\mathcal{F}^P := \sigma \left( \mathcal{F}, \left\{ A \subset \Omega \mid (\exists \tilde{A} \in \mathcal{F}) (A \subset \tilde{A} \text{ and } P[\tilde{A}] = 0) \right\} \right).$$

Throughout this text, it may always be assumed that the measure space is complete to avoid ambiguities with measurability. To distinguish probabilities from general expectations, following the usual convention, the symbol $\mathbb{E}$ is reserved for the expectation operator associated with $P$. For a measurable $f : \Omega \to \mathbb{R}$ that is positive or integrable:

$$\mathbb{E}[f] = \int_{\Omega} fd\mathbb{P}.$$ 

For general $\sigma$-finite measure spaces $(\Omega, \mathcal{F}, \mu)$, this distinction will not be made:

$$\mu(f) := \int_{\Omega} fd\mu.$$ 

For $p \in [0, \infty]$, $L^p(\Omega, \mathcal{F}, \mu)$ will shortly be denoted by $L^p(\mu)$ or if $X \sim \mu$ is a random variable with the distribution of $\mu$, $L^p(X)$. If the measure is understood, $L^p(\Omega)$ will mainly be used. This is the Banach space of $p$-integrable functions $g : \Omega \to \mathbb{R}$ with norm:

$$\|g\|_p = \|g\|_{L^p} = \begin{cases} \left( \int_{\Omega} |g|^p d\mu \right)^{1/p} & \text{if } p < \infty \\ \sup_{\Omega} |g| & \text{if } p = \infty \end{cases}.$$ 

Now let $\Omega$ be a topological space and $\mathcal{F}$ its Borel-$\sigma$-algebra (or the completion; see the following definition). Then $L^p_{\text{loc}}(\Omega, \mathcal{F}, \mu)$ denotes the space of locally $p$-integrable functions, the functions that are $p$-integrable on all compacta. Similar shorthand notations are in use.
Finally, let $a, b \in \mathbb{R} \cup \{\pm \infty\}$, $a < b$. $dx$ or $dz$ denotes Lebesgue measure if it is stated on its own. $L^p(a, b)$ is another shorthand for $p$-integrable functions with respect to the Lebesgue measure on $(a, b)$.

**Definition B.1.2.** A Polish space $S$ is a topological space such that a metric $d : S \times S \to [0, +\infty)$ exists that turns $S$ into a complete metric space. We reserve $\mathcal{B}$ or $\mathcal{B}(S)$ for the $\sigma$-algebra of Borel sets:

$$\mathcal{B} := \sigma(B \subset S \text{ open}).$$

When considering a probability measure $\mu$ on $S$, it may also denote its completion for $\mu$. There may be multiple metrics that turn $S$ into a complete space. These may exhibit distinct properties (a totally bounded metric may for instance be found, see [9]). Polish spaces are the spaces on which the usual probabilistic properties hold, such as the existence of conditional expectations. Due to separability, $d$ is measurable.

For a random variable $X$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, $\mathbb{P}_X := \mathbb{P} \circ X^{-1}$ denotes its distribution.

On $(S, \mathcal{B})$, we denote the probability measures by $\mathcal{P}(S)$ and the probability measures with **finite first moment** as:

$$\mathcal{P}^1(S) := \{ \mu \in \mathcal{P}(S) \mid \int_S d(x, x_0)\mu(dx) \},$$

where $x_0 \in S$ is arbitrary. The definition does not depend on $x_0$.

**Definition B.1.3.** Let $a, b \in \mathbb{R} \cup \{\pm \infty\}$. The following shorthand notation is in use throughout this text:

$$a \wedge b = \min\{a, b\}.$$  

$$a \vee b = \max\{a, b\}.$$  

**Definition B.1.4.** Let $n \in \mathbb{N}$, $k \in \mathbb{N} \cup \{\infty\}$. Consider a Polish space $S$ and an open subset $U$ of $\mathbb{R}^n$. We denote:

$$C(S) \supset C_b(S) \supset C_0(S),$$

the set of continuous functions $g : S \to \mathbb{R}$ without additional restrictions, that are bounded and that are compactly supported, respectively. Furthermore,

$$\text{Lip}(S) := \{ f \in C(S) \mid (\exists L > 0)(\forall x, y \in S)(d(f(x), f(y)) \leq Ld(x, y)) \} \supset \text{Lip}_b(S),$$

denote the Lipschitz and bounded Lipschitz functions respectively. Also write

$$\|h\|_{Lip} := \inf\{ L \geq 0 \mid (\forall x, y \in S)(|h(y) - h(x)| \leq Ld(x, y))\}.$$  

Finally,

$$C^k(U) \supset C^k_b(U) \supset C^k_0(U)$$
denote the sets of all \( k \) times continuously differentiable functions \( g : U \to \mathbb{R} \) with respectively no supplementary restrictions, that are bounded with bounded derivatives and the ones that have compact support in \( U \).

Let \( l \in \{1,\ldots,k\} \) and \( i_1,\ldots,i_l \in \{1,\ldots,n\} \). The partial derivatives of \( g \in C^k(U) \) are equivalently denoted by:

\[
\partial_{i_1,\ldots,i_l} g \equiv \frac{\partial^l}{\partial x_{i_1} \ldots \partial x_{i_l}} g \equiv \partial^\alpha g,
\]

where the variables of \( g \) are denoted by \( (x_1,\ldots,x_n) \). \( \alpha = (\alpha_1,\ldots,\alpha_n) \in (\mathbb{Z}^+)^n \) indicates multi-index notation. \( \alpha_j \) is the amount of derivatives on the \( j \)-th variable. Here, \( |\alpha| := \alpha_1 + \ldots + \alpha_n \leq k \). Furthermore, \( \{\beta \leq \alpha\} := \{ (\beta_1,\ldots,\beta_n) \in (\mathbb{Z}^+)^n \mid \beta_i \leq \alpha_i \ \text{for all} \ i \} \). In this way, the Leibniz rule reads for \( f, g \in C^k(U) \):

\[
\partial^\alpha (fg) = \sum_{\beta \leq \alpha} \binom{\alpha}{\beta} \partial^\beta f \partial^{\alpha-\beta} g,
\]

(B.6)

where \( \binom{\alpha}{\beta} \) denotes \( \alpha!/[\beta!(\alpha-\beta)!] \) and \( \alpha! = \alpha_1! \ldots \alpha_n! \).

**Definition B.1.5.** A locally integrable function \( g : U \subset_{\text{open}} \mathbb{R}^n \to \mathbb{R} \) is weakly differentiable iff there exist locally integrable \( f_i : U \to \mathbb{R} \) for \( i \in \{0,\ldots,n\} \) such that for all \( \varphi \in C_0^\infty(U) \):

\[
\int_U f_i(x) \varphi(x) dx = - \int_U g(x) \partial_i \varphi(x) dx.
\]

(B.7)

The \( f_i \) are then denoted as \( \delta_i g \). Let \( k \in \mathbb{N} \), \( p \in [1,\infty) \). The Sobolev space \( W^{k,p}(U) \) is the Banach space of functions \( g : \mathbb{R} \to \mathbb{R} \) that are \( k \) times weakly differentiable with \( p \)-integrable derivatives. Its norm is given by:

\[
\|g\|_{W^{k,p}} = \begin{cases} 
\left( \sum_{i=0}^k \|D^i g\|_{L^p(U)}^p \right)^{1/p} & \text{if } p < \infty, \\
\max_{i=0,\ldots,k} \|D^i g\|_\infty & \text{if } p = \infty.
\end{cases}
\]

Here, \( D^i g \) is a vector consisting of all partial derivatives of order \( i \) (or alternatively a tensor product, see chapter 4). The Euclidian norm is considered for \( \|D^i g\| \). An important property is that \( C^\infty(U) \cap W^{k,p}(U) \) is dense in \( W^{k,p}(U) \) for \( p < \infty \). If \( n = 1 \), and \( U = (a,b) \), \( a < b \), a fundamental theorem of calculus holds:

\[
g(y) - g(x) = \int_x^y g'(z) dz.
\]

In particular, \( g \) is continuous. The Sobolev space with zero-boundary conditions \( W_0^{k,p}(U) \) is defined as the closure of \( C_0^\infty(U) \) for the norm \( \|\cdot\|_{W^{k,p}(U)} \). These spaces exhibit many nice embedding and interpolation theorems. For more information, see [6] or [27]. The most important notions will however be stated or worked out in this text.

With \( G^{k,p}(U) \) we denote the space of \( k \)-times weakly differentiable functions such that the highest order derivatives are bounded in \( L^p(U) \).
Similarly, when another σ-finite measure $\mu$ is considered on $U \subset \mathbb{R}^n$, we can replace the Lebesgue $p$-norms by the alternative $p$-norms. If $Z \sim \mu$, then we denote these spaces by $W^{k,p}(Z)$, $G^{k,p}(Z)$ and so on. In this case, we use this notation mainly as a shorthand to state that the derivatives are bounded in the sense of the $L^p(Z)$ norms.

**Definition B.1.6.** For a linear operator between normed spaces $A : X \to Y$ and $G \subset X$, we denote the semi-norms $|A|_G := \sup_{x \in G} \|Ax\|$. The strong operator norm is in this way given by $\|A\| = |A|_{B_X(0,1)}$, where $B_X(0,1)$ denotes the unit ball in $X$. For a Banach space $X$, we denote its dual by $X'$. This is the vector space of linear continuous functionals $X \to \mathbb{R}$. When no topology is mentioned, we implicitly assume the one given by the strong operator norm. The kernel of $A$ is defined as:

$$\ker(A) = \{ x \in X \mid Ax = 0 \}.$$  

Consider a linear operator $A : \text{dom}(A) \subset X \to Y$, where now $X$ is a Hilbert space with inner product $\langle \cdot, \cdot \rangle_X$ and $\text{dom}(A)$ is a linear subspace of $X$. The **dual operator** $A^T : \text{dom}(A^T) \subset Y \to X$ is defined as the linear operator on:

$$\text{dom}(A^T) := \left\{ y \in Y \mid (\exists C > 0) (\forall x \in \text{dom}(A)) \left( |\langle Ax, y \rangle_X| \leq C \sqrt{|\langle Ax, y \rangle_X|} \right) \right\},$$  

such that for all $x \in \text{dom}(A)$:

$$\langle Ax, y \rangle_X = \langle x, A^Ty \rangle_X.$$  

This is well defined by Riesz’ representation theorem, Theorem B.3.4. More information can be found in section 2.1 and chapter 4.

**B.2 Tensor products of Hilbert spaces**

We recall here the notion of Hilbert spaces of tensor products. Let $H$ be a real separable Hilbert space with inner product $\langle \cdot, \cdot \rangle_H$. Consider the space of **pure k-fold tensor products** for $k \in \mathbb{N}$:

$$\{ h_1 \otimes \ldots \otimes h_k \mid h_1, \ldots, h_k \in H \}.$$  

They generate a linear space where $h_1 \otimes \ldots \otimes (ah_i + bg_i) \otimes \ldots \otimes h_k = a(h_1 \otimes \ldots \otimes h_i \otimes \ldots \otimes h_k) + b(h_1 \otimes \ldots \otimes g_i \otimes \ldots \otimes h_k)$, $a, b \in \mathbb{R}$. Recall that they may be defined as formal objects or as multilinear maps on $H^k$. So if a function on tensor products is considered, it should be checked that it is well-defined\(^2\). The canonical inner product defined on pure tensors,

$$\langle g_1 \otimes \ldots \otimes g_k, h_1 \otimes \ldots \otimes h_k \rangle_{H^\otimes k} := \langle g_1, h_1 \rangle_H \cdots \langle g_k, h_k \rangle_H,$$  

\(^2\)Consider the definition on the free vector space $F(H^k)$ and verify that it is multilinear.

\(^1\)A rigorous description of the linear space they generate, can be given by considering formal sums of elements in $H^k$, i.e. the free vector space $F(H^k)$ and imposing the equivalence classes under the relation of linear combinations of components. For instance $(ah_1, h_2, \ldots, h_n) \sim (ah_1, h_2, \ldots, h_n)$. Refer to [27] for more information.
extends by multilinearity to the linear space they generate. Then, the Hilbert space $H^\otimes k$ is obtained by taking the completion. It can be described explicitly in the following way. Suppose $H$ is infinite dimensional and fix an orthonormal basis $(e_i)_{i \in \mathbb{N}}$ ($N = \{1, \ldots, n\}$ or $N = \mathbb{N}$). Then $H^\otimes k$ consists of vectors

$$f := \sum_{i_1, \ldots, i_k = 1}^{\infty} a_{i_1, \ldots, i_k} e_{i_1} \otimes \ldots \otimes e_{i_k},$$

where $a_{i_1, \ldots, i_k} \in \mathbb{R}$ and

$$\|f\|_{H^\otimes k}^2 := \langle f, f \rangle_{H^\otimes k} = \sum_{i_1, \ldots, i_k = 1}^{\infty} a_{i_1, \ldots, i_k}^2 < \infty.$$ 

The limit in (B.10) is taken with respect to $\|\cdot\|_{H^\otimes k}$. Again, the reader should be careful with the fact of identifications based on a multilinear combination in a fixed component. Note that the tensor products $\{e_{i_1} \otimes \ldots \otimes e_{i_k}\}$ form a countable orthogonal basis so that the $a_{i_1, \ldots, i_k}$ are unique and there is an isometry from the square summable sequences $(a_{i_1, \ldots, i_k})$ to $H^\otimes k$.

$H^\otimes 1$ is isomorphic to $H$ and will just be denoted by $H$.

A tensor $f \in H^\otimes k$ is called symmetric iff for all permutations $\sigma$ of $\{1, \ldots, k\}$ and any $h_1, \ldots, h_k \in H$:

$$\langle f, h_{\sigma(1)} \otimes \ldots \otimes h_{\sigma(k)} \rangle_{H^\otimes k} = \langle f, h_1 \otimes \ldots \otimes h_k \rangle_{H^\otimes k}.$$ 

(B.11)

The closed subspace of $H^\otimes k$ consisting of symmetric tensor products gets denoted by $H^\otimes k$. $e_1 \otimes e_2 + e_2 \otimes e_1$ is a non-pure symmetric tensor product in $H^\otimes 2$ for example. Given a tensor $f \in H^\otimes k$, we can turn it into a symmetric tensor $\tilde{f} \in H^\otimes k$ by (canonical) symmetrization:

$$\langle \tilde{f}, h_1 \otimes \ldots \otimes h_k \rangle_{H^\otimes k} := \frac{1}{k!} \sum_\sigma \langle f, h_{\sigma(1)} \otimes \ldots \otimes h_{\sigma(k)} \rangle_{H^\otimes k}.$$ 

(B.12)

The sum is taken over all permutations $\sigma$ of $\{1, \ldots, k\}$. If $f$ already was symmetric, then $\tilde{f} = f$.

The $r$-th contraction of two tensor products $f \in H^\otimes k$ and $g \in H^\otimes l$, $f \otimes_r g \in H^\otimes k+l-2r$, where $0 \leq r \leq k \wedge l$, is defined by extending

$$(h_1 \otimes \ldots \otimes h_k) \otimes_r (v_1 \otimes \ldots \otimes v_k) := \left( \prod_{i=1}^{r} \langle h_i, v_i \rangle_H \right) h_{r+1} \otimes \ldots \otimes h_k \otimes v_{r+1} \otimes \ldots \otimes v_k,$$

$h_i, v_i \in H$. If $r = 0$, the product is replaced by 1. So if we take representations with respect to $\{e_{i_1} \otimes \ldots \otimes e_{i_k}\}$, (B.10), this becomes:

$$f = \sum_{i_1, \ldots, i_k = 1}^{\infty} a_{i_1, \ldots, i_k} e_{i_1} \otimes \ldots \otimes e_{i_k},$$

103
\[ g = \sum_{j_1, \ldots, j_l=1}^{\infty} b_{j_1, \ldots, j_l} e_{j_1} \otimes \cdots \otimes e_{j_l}, \]

\[ f \otimes_r g := \sum_{i_1, \ldots, i_k=1}^{\infty} \sum_{j_1, \ldots, j_k=1}^{\infty} a_{i_1, \ldots, i_k} b_{j_1, \ldots, j_k} (e_{i_1} \otimes \cdots \otimes e_{i_k}) \otimes_r (e_{j_1} \otimes \cdots \otimes e_{j_k}) \]

(B.13)

\[ = \sum_{z_1, \ldots, z_{k+l-2r}=1}^{\infty} (a \ast_r b)_{z_1, \ldots, z_{k+l-2r}} e_{z_1} \otimes \cdots \otimes e_{z_{k+l-2r}}, \]

(B.14)

Note that it is well-defined by the Cauchy-Schwartz inequality (the numerical sequence \(a \ast_r b\) is square summable) and that \(\otimes_r\) is bilinear and continuous. If \(r = 0\), we just take the usual tensor product. If \(r = k = l\), then the inner product \(\langle \cdot, \cdot \rangle_{H^\otimes k}\) is obtained.

When \(f \in H^\otimes k\) and \(g \in H^\otimes l\) are symmetric, then the following expression is also well-defined:

\[ f \otimes_r g = \sum_{y_1, \ldots, y_r=1}^{\infty} \langle f, e_{y_1} \otimes \cdots \otimes e_{y_r} \rangle_{H^\otimes r} \otimes \langle g, e_{y_1} \otimes \cdots \otimes e_{y_r} \rangle_{H^\otimes r}. \]

(B.15)

This is because we do not need to specify with respect to which components the inner product is taken. However, it is not necessarily symmetric. Therefore we reserve \(\hat{\otimes}_r\) for the symmetrization:

\[ f \hat{\otimes}_r g := \frac{1}{(k + l - 2r)!} \sum_{\sigma} \sum_{z_1, \ldots, z_{k+l-2r}=1}^{\infty} (a \ast_r b)_{z_1, \ldots, z_{k+l-2r}} e_{z_{\sigma(1)}} \otimes \cdots \otimes e_{z_{\sigma(k+l-2r)}}, \]

(B.16)

where \(\sigma\) runs through the permutations of \(\{1, \ldots, k + l - 2r\}\).

### B.3 Important theorems and lemmas

Let \(I\) be an arbitrary index set and consider Borel spaces \((\Omega_i, \mathcal{F}_i)_{i \in I}\) and probability measures \((P_j)_{J \subseteq I \text{ finite}}\), \(P_j\) being defined on \((\times_{j \in J} \Omega_j, \otimes_{j \in J} \mathcal{F}_j)\). We want to glue together these probability spaces into a probability space \((\Omega, \mathcal{F}, \mathbb{P})\). Here \(\Omega := \times_{i \in I} \Omega_i\), \(\mathcal{F} := \otimes_{i \in I} \mathcal{F}_i\), \(\mathbb{P} \circ \pi_J = P_J\) for \(J \subseteq I \text{ finite}\). Here, \(\pi_J\) denotes the projection on the coordinates corresponding to \(J\). Thus, we are interested in finding the projective limit of the \((P_J)\).

The \((P_J)\) are called **consistent** if for every two finite \(J_1 \subset J_2 \subset I\), \(\mathbb{P}_{J_2} \circ \pi_{J_1}^{-1} = \mathbb{P}_{J_1}\).


**Theorem B.3.1** (Extension theorem of Daniell Kolmogorov). If the \((P_J)_{J \subseteq I \text{ finite}}\) are consistent, then there exists a unique probability measure \(\mathbb{P}\) on \((\Omega, \mathcal{F})\) such that for all finite \(J \subseteq I\), \(\mathbb{P} \circ \pi_J^{-1} = \mathbb{P}_J\).

Let \(n \in \mathbb{N}\) and \(U \subseteq \mathbb{R}^n\) be open. Consider \(a < b\) in \(\mathbb{R} \cup \{\pm \infty\}\).
Lemma B.3.2 (Sobolev product rule). Let \( p, p' \in [1, \infty) \), \( p + p' - 1 = 1 \), \( f \in W^{1,p}(U) \) and \( g \in W^{1,p'}(U) \). Then \( fg \in W^{1,1}(U) \) and:

\[
(fg)' = f'g + fg'.
\]

By a Fubini argument, it remains true if \( f, g \in W^{1,1}(a,b) \) and \( f'g \) and \( fg' \) are integrable.

See [6] for more information on Sobolev spaces.

Theorem B.3.3 (Rademacher’s theorem). Consider \( U \subset \mathbb{R}^n \) and a Lipschitz function \( f : U \to \mathbb{R} \). Then \( f \) is Lebesgue-almost everywhere differentiable\(^3\) with \( \|f'\|_{\infty} < \infty \). If \( n = 1 \) and \( U \) is an interval, it holds for all \( a < b \) in \( U \) that:

\[
f(b) - f(a) = \int_a^b f'(x)dx. \tag{B.17}
\]

Actually, a similar theorem holds for the second derivative of convex functions (Alexandrov’s theorem). Also, there need not be a uniform Lipschitz condition (Stepanov’s theorem).

Theorem B.3.4 (Riesz’ representation theorem, real version). Given a real Hilbert space \( H \) with inner product \( \langle \cdot, \cdot \rangle_H \) and dual space \( H' \), the following is an isometric isomorphism:

\[
J : H \to H'; \ x \mapsto \langle \cdot, x \rangle_H. \tag{B.18}
\]

Recall Definition B.1.6. This implies that for every linear continuous functional \( f \in H' \), there exists an \( x \in H \) with the same norm such that \( f(y) = \langle y, x \rangle_H \), for all \( y \in H \).

Theorem B.3.5 (Dual of \( L^p(\Omega) \)). Let \( (\Omega, \mathcal{F}, \mu) \) be a \( \sigma \)-finite measure space, \( p \in [1, \infty) \) and \( p^{-1} + p'^{-1} = 1 \). Then we get an isometric isomorphism:

\[
J : L^{p'}(\Omega) \to (L^p(\Omega))'; \ g \mapsto \left( J(g) : L^p(\Omega) \to \mathbb{R}; \ f \mapsto \int_{\Omega} fg d\mu \right). \tag{B.19}
\]


A class \( C \) of functions \( X \to \mathbb{R} \) is said to separate points iff for all \( x, y \in X \), there exists an \( f \in C \) such that \( f(x) \neq f(y) \).

Theorem B.3.6 (Stone-Weierstrass, real version). Let \( X \) be a compact Hausdorff space and \( C \subset C_b(X) \) an algebra that separates points. Then \( C \) is dense in \( (C_b(X), \|\cdot\|_{\infty}) \).

It can be found in chapter 15 of [15].

Lemma B.3.7 (Ulam’s tightness theorem). If \( S \) is a separable complete metric space, then every probability measure \( \mu \) on the Borel sets is tight. This means that \( \forall \epsilon > 0 \), there exists a compact \( K \subset S \) such that \( \mu(K) \geq 1 - \epsilon \).

\(^3\)Meaning that the set of points such that \( f \) is not differentiable form a null set.
Let \((\Omega, \mathcal{F})\) be a measure space. \(\mu : \mathcal{F} \to \mathbb{R}\) is called a signed measure on \((\Omega, \mathcal{F})\) if it is \(\sigma\)-additive: for any sequence of pairwise disjoint sets \(A_1, A_2, \ldots\):

\[
\mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i).
\]

**Theorem B.3.8** (Jordan decomposition). If \(\mu\) is a signed measure on \((\Omega, \mathcal{F})\), then there exists a disjoint decomposition \(\Omega = \Omega^+ \cup \Omega^-\) and uniquely determined finite positive measures \(\mu^+\) and \(\mu^-\) such that \(\mu = \mu^+ - \mu^-\) and:

\[
\mu^+(A) = 0, \quad \forall A \subset S^- \text{ measurable}, \quad \mu^-(B) = 0, \quad \forall B \subset S^+ \text{ measurable}.
\]

It may be found in chapter 15 of [15].

**Theorem B.3.9** (Hahn-Banach approximation theorem). Let \(X\) be a normed space and consider a linear subspace \(X_0 \subset X\). Then \(X_0\) is dense in \(X\) iff for every \(\varphi \in X'\) such that \(\varphi|_{X_0} = 0\), it holds that \(\varphi = 0\).

It is also true in locally convex spaces. See chapter 18 of [27].

A sequence of measures \((\mu_n)_{n \in \mathbb{N}}\) on a metric space \((S, d)\) is tight if and only if for all \(\epsilon > 0\), there exists a compact \(K \subset S\) such that \(\inf_{n \in \mathbb{N}} \mu_n(K) \geq 1 - \epsilon\). It is called weakly relatively compact iff for every subsequence \((\mu_{n(k)})_{k \in \mathbb{N}}\) there exists another subsequence that converges weakly in \(\mathcal{P}(S)\). See chapter 13 of [15].

**Theorem B.3.10** (Prohorov). Let \((S, d)\) be a metric space and consider measures \((\mu_n)_{n \in \mathbb{N}}\) on \((S, \mathcal{B}(S))\). Then:

1. If \((\mu_n)_{n \in \mathbb{N}}\) is tight, then it is weakly relatively compact.

2. If \(S\) is Polish and \((\mu_n)_{n \in \mathbb{N}}\) is weakly relatively compact, then it is tight.

A form of the Arzelà-Ascoli theorem is given by:

**Theorem B.3.11** (Arzelà-Ascoli). Let \((T, d)\) be a compact metric space and \(C \subset C(T)\). Then \(C\) is totally bounded for \(\|\cdot\|_{\infty}\) if and only if it is uniformly bounded and equicontinuous, thus uniformly equicontinuous.

Uniformly bounded means that \(C\) is contained in some ball \(\{h \in C(T) \mid \|h - h_0\| \leq K\}\) in \(C(T)\) \((h_0 \in C(T), \ K > 0)\). Totally bounded means that for all \(\epsilon > 0\), there exists finitely many \(f_1, \ldots, f_n \in C\) such that for all \(f \in C\), there exists a \(j \in \{1, \ldots, n\}\) such that \(\|h - f_j\| \leq \epsilon\). It thus means that \(C\) can be covered by finitely many balls of any size. Finally, equicontinuous means:

\[
(\forall \epsilon > 0)(\exists \delta > 0)(\forall f \in C)(\forall x, y \in T)(d(x, y) < \delta \Rightarrow |f(x) - f(y)| < \epsilon).
\]

See for example chapter 2 of [9].
Appendix C

Additional arguments

C.1 Zolotarev-type distances

This section contains supplementary arguments for the results in section 1.3.

**Lemma C.1.1.** Suppose $\mu \in P_1(\mathbb{R})$ is absolutely continuous with respect to Lebesgue measure $dx$ with density $f$ that is bounded. Then we have for any $\nu \in P_1(\mathbb{R})$ that:

$$d_K(\mu, \nu) \leq \sqrt{\|f\|_{\infty}}d_W(\mu, \nu).$$

**Proof.** Let $x \in \mathbb{R}$. We want to smoothen $\mathbb{1}_{(-\infty, x]}$. First consider the “upper smoothed” $h_\epsilon(y) := (1 \wedge \frac{1}{\epsilon}(x + \epsilon - y)) \vee 0$. Then, $\|h_\epsilon\|_{\text{Lip}} = 1/\epsilon$ and

$$\int_{-\infty}^{x} d\nu - \int_{-\infty}^{x} d\mu \leq \int h_\epsilon d\nu - \int h_\epsilon d\mu + \int_{x}^{x+\epsilon} f(x)dx \quad (C.1)$$

$$\leq \frac{1}{\epsilon}d_W(\mu, \nu) + \epsilon\|f\|_{\infty}.$$

The $\nu$-integral just enlarges when replacing the indicator function. In the $\mu$-integral of (C.1), we need to compensate by an error that can be bounded by 1 on $(x, x + \epsilon)$. Considering the “lower smoothed” $g_\epsilon(y) := (1 \wedge \frac{1}{\epsilon}(x - y)) \vee 0$ and reasoning similarly yields the total bound (after taking the supremum):

$$d_K(\mu, \nu) \leq \frac{d_W(\mu, \nu)}{\epsilon} + \|f\|_{\infty}\epsilon.$$  

Optimizing with $\epsilon = \sqrt{d_W(\mu, \nu)/\|f\|_{\infty}}$ yields the claim. □

By a similar argument we have for $\mu, \nu \in P(\mathbb{R})$, where $\mu$ is absolutely continuous with respect to Lebesgue measure with density $f$:

$$d_K(\mu, \nu) \leq d_{BW}(\mu, \nu) + \sqrt{\|f\|_{\infty}}d_W(\mu, \nu) \leq (\sqrt{2} + \sqrt{\|f\|_{\infty}})\sqrt{d_{BW}(\mu, \nu)}, \quad (C.2)$$

because $d_{BW}$ is bounded by 2. Here, $\|h_\epsilon\|_{\infty} + \|h_\epsilon'\|_{\infty} = 1 + \epsilon^{-1}$.

Before being able to proof the other claims, we need to show that the bounded Wasserstein distance metrizes weak convergence.
C.1.1 Weak convergence of probability measures

If $T$ is a Polish space with Borel-sigma-algebra $\mathcal{T}$, then laws $\mu_n \in \mathcal{P}(T)$ converge weakly to $\mu \in \mathcal{P}(T)$ iff:

$$\int h d\mu_n \to \int h d\mu, \quad \forall h \in C_b(T). \quad (C.3)$$

We fix a metric $d$ that metrizes $T$. Recall from section B.1 that we denote the bounded Lipschitz functions on $T$ by $\text{Lip}_b(T)$. Some properties of the portemanteau$^1$ are:

**Lemma C.1.2** (Portemanteau). Let $\mu_n, \mu \in \mathcal{P}(T)$, $n \in \mathbb{N}$. The following are equivalent:

1. $\mu_n \to \mu$ weakly in $\mathcal{P}(T)$ as $n \to \infty$.

2. For all $f \in \text{Lip}_b(T)$, as $n \to \infty$:

$$\int f d\mu_n \to \int f d\mu. \quad (C.4)$$

3. For all upper semicontinuous $f : S \to \mathbb{R}$ that are bounded from above:

$$\limsup_{n \to \infty} \int f d\mu_n \leq \int f d\mu. \quad (C.5)$$

4. $\limsup_{n \to \infty} \mu_n(C) \leq \mu(C)$ for all $C \subset \text{closed} \ T$.

$Lip_b(T)$ can be normed by:

$$\|h\|_{BL} := \|h\|_{\infty} + \|h\|_{Lip}. \quad (C.6)$$

If $T$ is compact, the Stone-Weierstrass theorem, Theorem B.3.6 implies that $\text{Lip}_b(T)$ is dense in $C(T)$ for $\|\cdot\|_{\infty}$. Recall the Bounded Wasserstein distance:

$$d_{bW}(X, Z) = \sup \{ \|E[h(X)] - E[h(Z)]\| \mid \|h\|_{BL} \leq 1 \}. \quad \text{for general Polish } T,\text{ it will be possible to restrict to convenient compacts in } T. \text{ By the portemanteau theorem, weak convergence is equivalent to convergence over all bounded Lipschitz functions.}

**Proposition C.1.3.** For a separable metric space $(T, d)$, probability laws $\mu_n$ converge weakly to a law $\mu$ on $T$ if and only if $d_{bW}(\mu_n, \mu) \to 0$.  

---

$^1$Eng. coathanger.
Proof. \(\cdot\Leftrightarrow\) Thi implication is direct from the portemanteau theorem.

\(\cdot\Rightarrow\): Since we can consider the completion \(\overline{T}\) of \(T\) and extend any \(h \in BL(T, d)\) to \(\overline{h} \in BL(\overline{T}, d)\) with \(\|h\|_{BL} = \|\overline{h}\|_{BL}\), we may assume that \(T\) already is complete. On Polish spaces we may find for any \(\epsilon > 0\) a compact \(K \subset T\) such that \(\mu(K) > 1 - \epsilon\) (Lemma B.3.7). \(H := \{h \in BL(T, d) \mid \|h\|_{BL} \leq 1\}\) restricted to \(\mathcal{H}_K := \{h|_K \mid h \in \mathcal{H}\}\) forms a compact subset of \((C(K), \|\cdot\|_\infty)\) by the Arzelà-Ascoli theorem, Theorem B.3.11. Because compactness implies totally boundedness in metric spaces, we find \(h_1, \ldots, h_k \in H\) such that for any \(h \in H\), there is a \(j\) with \(\sup_{y \in K} |h(y) - h_j(y)| < \epsilon\).

If we denote \(K_\epsilon := \{y \in T \mid d(y, K) < \epsilon\}\), then we can strengthen this with the Lipschitz condition to:

\[
|\int h d\mu_n - \int h d\mu| \leq \int_{T \setminus K^*} |h - h_j| d\mu_n + \int h_j d\mu_n - \mu \leq 2(\mu_n + \mu)(T \setminus K^*) + (3\epsilon) \cdot 2 + \left|\int h_j d(\mu_n - \mu)\right|.
\]

To bound the first term, we used \(|h - h_j| \leq 2\) and (C.7). We obtain from \(\mu_n(K^*) > 1 - 2\epsilon\) for \(n\) sufficiently large. We infer for \(h \in \mathcal{H}\) and corresponding \(h_j\):

\[
d_{bW}(\mu_n, \mu) \leq 12\epsilon + \max_{j=1,\ldots,k} \left|\int h_j d(\mu_n - \mu)\right|.
\]

This can be made smaller than \(13\epsilon\) by weak convergence when taking \(n\) sufficiently large. Since \(\epsilon > 0\) was arbitrary, we have shown convergence in the bounded Wasserstein metric.

Metrizability yields nice properties on the space \(\mathcal{P}(T)\) equipped with weak convergence. It can for example be used to prove Prohorov’s theorem (see chapter 11 of [9]).

For the purpose of Stein’s method, \(d_{bW}\) may be less convenient since \(\delta^{-1}\) may not translate \(\|\cdot\|_{BL}\) transparantly. Also, the stronger Wasserstein metric can often be considered, which is defined for laws \(\mu\) on \(T\) with first moments. Recall from Definition B.1.1 the notation \(\mathcal{P}^1(T)\) for probability measures with first moments.

**Proposition C.1.4.** Let \((T, d)\) be a separable metric space. On \(\mathcal{P}^1(T)\), \(d_W\) metrizes convergence of Lipschitz functions.
This means that $d_W(\mu_n, \mu) \to 0$ if and only if for all Lipschitz $h \in C(T)$ we have $\int h d\mu_n \to \int h d\mu$. This relates to the Rubinstein-Kantorovich theorem, Theorem 1.3.6, or [29].

For most spaces $T$, the Wasserstein metric is strictly stronger than the bounded Wasserstein metric. Consider $\mathbb{R}$ with the usual norm and $\mu := \delta_0$, $\mu_n = (1 - 1/n)\delta_0 + (1/n)1_{(n, n+1)}$. Then $\mu_n$ converges to $\mu$ for $d_{bW}$ but not for $d_W$ (consider $h(x) = x$). They are however equivalent on sets of probability laws $Q$ that have uniformly integrable first moments, i.e. for an arbitrary $x_0 \in T$:

$$(\forall \epsilon > 0)(\exists K_\epsilon \subseteq_{\text{compact}} T)(\sup_{\mu \in Q} \int_{K_\epsilon} d(x, x_0) d\mu < \epsilon). \quad \text{(U.I.)}$$

**Lemma C.1.5.** For $Q \subset P^1(T)$ with uniformly integrable first moments, convergence in $d_{bW}$ implies convergence in $d_W$.

**Proof.** With this assumption, we may restrict integration to bounded functions. For $\epsilon > 0$, take $K_\epsilon$ as in (U.I.). If $h$ is a Lipschitz function, then it is bounded by a constant $C$ on $K_\epsilon$. Suppose we have $d_{bW}(\mu_n, \mu) \to 0$ for a laws $Q$:

$$\left| \int h d(\mu_n - \mu) \right| \leq \left| \int_{K_\epsilon} [(h \wedge C) \vee (-C)] d(\mu_n + \mu) \right|$$

$$+ \int_{K_\epsilon} (d(x, x_0) + |h(x_0)|) d(\mu_n + \mu)$$

$$\leq \left| \int_T [(h \wedge C) \vee (-C)] d(\mu_n + \mu) \right| + \tilde{C}\epsilon,$$

for some $\tilde{C} > 0$. The first term can be made smaller than $\epsilon$ for $n$ sufficiently large since $[(h \wedge C) \vee (-C)]$ is a bounded Lipschitz function. \qed

The relation with $d_K$ is summed up in the following lemma. Its result is well-known if ‘convergence for $d_{bW}$’ is replaced by distributional convergence (see for instance point (i) of Lemma 2.2 in [28] and use compactness of the image of the distribution function, [0, 1], for the converse).

**Lemma C.1.6.** Let $T = \mathbb{R}$ and consider random variables $(X_n)_{n \in \mathbb{N}}, Z$. If $d_K(X_n, Z) \to 0$ for $n \to \infty$, then $d_{bW}(X_n, Z) \to 0$. If $Z$ has a continuous distribution function, the converse holds.

This implies parts 2. and 3. of Proposition 1.3.1.

Again for most $T$, $d_{TV}$ and $d_K$ are strictly stronger than $d_{bW}$. If we consider $X_n = (1/n, \ldots, 1/n)$ in $\mathbb{R}^d$, then $X_n$ converges weakly to $Z = (0, \ldots, 0)$ but not in $d_K$ or $d_{TV}$.
Lemma C.1.7 (Total variation distance). Let $X$ and $Z$ be $S$-valued random variables and $a < b$ real numbers. Then:

$$d_{TV}(X, Z) = \frac{1}{b-a} \sup_{f:S \to [a,b]} |\mathbb{E}[f(X)] - \mathbb{E}[f(Z)]|.$$  

(C.8)

If $S$ is discrete, then:

$$d_{TV}(X, Z) = \frac{1}{2} \sum_{s \in S} |\mathbb{P}[X = s] - \mathbb{P}[Z = s]|.$$  

(C.9)

Proof. Direct computation and approximation may be tried. It can also be shown efficiently by the Jordan decomposition $S^+ \cup S^-$ of $\mu := \mathbb{P}_X - \mathbb{P}_Y$, Theorem B.3.8. So suppose $\mu(A) \geq 0$ for all Borel $A \subset S$ and $\mu(B) \leq 0$ for all Borel $B \subset S$. Then it is directly seen that:

$$d_{TV}(X, Z) = \mathbb{P}[X \in S^+] - \mathbb{P}[Y \in S^+] = -(\mathbb{P}[X \in S^-] - \mathbb{P}[Y \in S^-]).$$

By enlarging any measurable $f : S \to [a, b]$ on $S^+$ and shrinking it on $S^-$, the largest possible value for the right hand side of (C.8) is attained in $f = b1_{S^+} + a1_{S^-}$, which yields the left hand side.

Finally, if $S$ is discrete, it is possible to choose

$$A = \{x \in S \mid \mathbb{P}[X = x] > \mathbb{P}[Y = x]\},$$

in the definition of the total variation metric. The argument can be concluded from:

$$d_{TV}(X, Z) = \mathbb{P}[X \in A] - \mathbb{P}[Y \in A] = \sum_{s \in A} |\mathbb{P}[X = s] - \mathbb{P}[Y = s]|$$

$$= \sum_{s \in A^c} |\mathbb{P}[X = s] - \mathbb{P}[Y = s]|.$$

\qed
Bibliography


