Stein's method for normal approximation

Stein's method, initiated by C. Stein in his celebrated monograph [22] (after the seminal contribution [21]), is a general device to achieve approximation of probability measures by a fixed target measure, typically the normal distribution.

It is based on the so-called Stein lemma which expresses that a given real integrable random variable X, defined on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$, has law the standard Gaussian distribution $\mathcal{N}(0,1)$ if and only if, for any differentiable function $f : \mathbb{R} \to \mathbb{R}$ such that $Xf(X)$ and $f'(X)$ are integrable,

$$\mathbb{E}\big(Xf(X)\big) \,=\, \mathbb{E}\big(f'(X)\big). \tag{1}$$

Whenever X has distribution $\mathcal{N}(0,1)$, this is of course the basic integration by parts formula with respect to the Gaussian density (cf. [1]). For a quick proof, apply (1) to the real and imaginary parts of the functions $f(x) = e^{itx}$, $x \in \mathbb{R}$, depending on the parameter $t \in \mathbb{R}$, to get that $\mathbb{E}(Xe^{itX}) = it\,\mathbb{E}(e^{itX})$ for every $t \in \mathbb{R}$. By the integrability of X, the characteristic function $\varphi_X(t) = \mathbb{E}(e^{itX})$, $t \in \mathbb{R}$, of the law of X is differentiable and $\varphi'_X(t) = i\,\mathbb{E}(Xe^{itX})$. Thus it solves the differential equation $\varphi'_X(t) = -t\varphi_X(t)$, $t \in \mathbb{R}$, implying that $\varphi_X(t) = e^{-\frac{1}{2}t^2}$, $t \in \mathbb{R}$. Weaker assumption on X may be considered [18].

The idea underlying this observation is a path towards approximating Gaussian distributions without involving these Gaussian distributions themselves. Typically, if N has law $\mathcal{N}(0,1)$ and X is any integrable random variable, the Kolmogorov distance between the respective laws of N and X may be estimated, in terms of the Stein lemma, as

$$\sup_{t\in\mathbb{R}} \big|\mathbb{P}(X \le t) - \mathbb{P}(N \le t)\big| \,\le\, 2\Big(\sup_{f\in\mathcal{D}} \big|E\big(Xf(X)\big) - \mathbb{E}\big(f'(X)\big)\big|\Big)^{\frac{1}{2}} \tag{2}$$

Drafted by M. L. **v1** March 2024

where $\mathcal{D}$ is the family of twice continuously differentiable functions $f : \mathbb{R} \to \mathbb{R}$ such that $\|f\|_\infty \leq 1$, $\|f'\|_\infty \leq 1$, $\|f''\|_\infty \leq 1$.

The post is a brief survey of some ideas and results around Stein's method. Complete expositions include [22, 6, 11, 18, 9, 10] among others.

Table of contents

1 Solving Stein's equation

Let $d\gamma_1(x) = e^{-x^2/2}\frac{dx}{\sqrt{2\pi}}$ be the standard normal distribution on the real line $\mathbb{R}$. Given a (bounded, measurable) function $g : \mathbb{R} \to \mathbb{R}$, the classical Stein equation looks for a solution $f : \mathbb{R} \to \mathbb{R}$ of the equation

$$f'(x) - x\, f(x) \;=\; g(x) - \int_{\mathbb{R}} g\, d\gamma_1 \tag{3}$$

that is absolutely continuous and such that there exists a version of the derivative f' verifying (3) for every $x \in \mathbb{R}$.

Actually, a solution may be easily be produced as

$$f(x) \;=\; c\, e^{\frac{1}{2}x^2} + e^{\frac{1}{2}x^2} \int_{-\infty}^{x} \big[g(y) - \textstyle\int_{\mathbb{R}} g\, d\gamma_1\big] e^{-\frac{1}{2}y^2} dy, \quad x \in \mathbb{R}, \tag{4}$$

where $c \in \mathbb{R}$. When $c = 0$, f is the unique solution of (3) such that $\lim_{x\to\pm\infty} e^{-\frac{1}{2}x^2} f(x) = 0$. For a proof, simply observe that Stein's equation (3) may be rewritten as

$$e^{\frac{1}{2}x^2}\, \frac{d}{dx}\big[e^{-\frac{1}{2}x^2}\, f(x)\big] \;=\; g(x) - \int_{\mathbb{R}} g\, d\gamma_1.$$

The claim when $c = 0$ is a consequence of the dominated convergence theorem.

2 Stein's inequality for the total variation distance

When the input function g in Stein's equation (3) is bounded, for example taking values in $[0,1]$, it is not difficult to observe that the solution f is bounded as well as its derivative. More precisely, f may be chosen so that $\|f\|_\infty \leq \sqrt{2\pi}\,\|g\|_\infty$ and $\|f'\|_\infty \leq 4\,\|g\|_\infty$ (cf. e.g. [18]). This remark has significant importance towards bounds on distances between probability measures.

For example, recall the total variation distance between a probability measure μ on $\mathbb{R}$ and γ_1,

$$\|\mu - \gamma_1\|_{\mathrm{TV}} \,=\, \sup_{A \in \mathcal{B}(\mathbb{R})} \big[\mu(A) - \gamma_1(A)\big] \,=\, \frac{1}{2} \sup \bigg[\int_{\mathbb{R}} g\, d\mu - \int_{\mathbb{R}} g\, d\gamma_1 \bigg]$$

where the last supremum is taken over all bounded measurable $g : \mathbb{R} \to \mathbb{R}$ with $\|g\|_\infty \leq 1$. The previous comments thus yield the following basic approximation bound, sometimes called Stein's inequality.

Proposition 1 (Stein's inequality)**.** *In the preceding notation,*

$$\|\mu - \gamma_1\|_{\mathrm{TV}} \,\leq\, \sup \bigg| \int_{\mathbb{R}} f' d\mu - \int_{\mathbb{R}} x\, f\, d\mu \bigg| \tag{5}$$

where the supremum runs over all continuously differentiable functions $f : \mathbb{R} \to \mathbb{R}$ such that $\|f\|_\infty \leq \sqrt{\frac{\pi}{2}}$ and $\|f'\|_\infty \leq 2$.

As already mentioned, one of the main interests in (5) lies in the fact that only the measure μ is involved in the upper bound via explicit integrals. It has been used in a wide range of applications quantifying the convergence to a normal distribution (see the general references).

The bound (2) on the Kolmogorov distance displayed in the introduction is achieved in a similar way (see [9]).

3 Application to Berry-Esseen-type bounds

Stein's method may be illustrated in a relevant way on the standard central limit theorem. Consider the simplest instance of a sequence $(X_k)_{k\geq 1}$ of independent real random variables equally distributed as a random variable X with mean zero and variance one. The central limit theorem thus expresses that the sequence $\big(\frac{S_n}{\sqrt{n}}\big)_{n\geq 1}$, where $S_n = X_1 + \cdots + X_n$, converges weakly to a random variable N with the standard normal law $\mathcal{N}(0,1)$.

The central limit theorem may be quantified by the celebrated Berry-Esseen inequality, under a third moment assumption (cf. [2]). What follows does not reach the full strength of

this inequality, but illustrates in a simple manner the use of Stein's method in the form of the inequality (2).

Towards the use of (2), let $f : \mathbb{R} \to \mathbb{R}$ be a function such that $\|f\|_\infty \leq 1$, $\|f'\|_\infty \leq 1$, $\|f''\|_\infty \leq 1$, and consider the expression in $n \geq 2$,

$$\mathbb{E}\Big(\frac{S_n}{\sqrt{n}}\, f\Big(\frac{S_n}{\sqrt{n}}\Big)\Big) - \mathbb{E}\Big(f'\Big(\frac{S_n}{\sqrt{n}}\Big)\Big).$$

By identical distribution $\mathbb{E}\big(\frac{S_n}{\sqrt{n}} f\big(\frac{S_n}{\sqrt{n}}\big)\big) = \sqrt{n}\,\mathbb{E}\big(X_n f\big(\frac{S_n}{\sqrt{n}}\big)\big)$. By a Taylor expansion of f at the second order around $\frac{S_{n-1}}{\sqrt{n}}$,

$$f\Big(\frac{S_n}{\sqrt{n}}\Big) \,=\, f\Big(\frac{S_{n-1}}{\sqrt{n}} + \frac{X_n}{\sqrt{n}}\Big) \,=\, f\Big(\frac{S_{n-1}}{\sqrt{n}}\Big) + \frac{X_n}{\sqrt{n}}\, f'\Big(\frac{S_{n-1}}{\sqrt{n}}\Big) + \frac{X_n^2}{2n}\, f''(\Theta)$$

for some (random) Θ. Using the independence between X_n and S_{n-1}, the moment conditions $\mathbb{E}(X_n) = 0$, $\mathbb{E}(X_n^2) = 1$, and the hypothesis $\|f''\|_\infty \leq 1$, it follows that

$$\Big|\mathbb{E}\Big(X_n f\Big(\frac{S_n}{\sqrt{n}}\Big)\Big) - \frac{1}{\sqrt{n}}\,\mathbb{E}\Big(f'\Big(\frac{S_{n-1}}{\sqrt{n}}\Big)\Big| \,\leq\, \frac{1}{2n}\,\mathbb{E}\big(|X|^3\big).$$

In the same way,

$$\Big|\mathbb{E}\Big(f'\Big(\frac{S_n}{\sqrt{n}}\Big)\Big) - \mathbb{E}\Big(f'\Big(\frac{S_{n-1}}{\sqrt{n}}\Big)\Big| \,\leq\, \frac{1}{\sqrt{n}}\,.$$

As a conclusion,

$$\Big|\mathbb{E}\Big(\frac{S_n}{\sqrt{n}}\, f\Big(\frac{S_n}{\sqrt{n}}\Big)\Big) - \mathbb{E}\Big(f'\Big(\frac{S_n}{\sqrt{n}}\Big)\Big)\Big| \,\leq\, \frac{1}{2\sqrt{n}}\,\mathbb{E}\big(|X|^3\big) + \frac{1}{\sqrt{n}}\,.$$

Together with (2), for any $n \geq 1$,

$$\sup_{t\in\mathbb{R}} \Big|\mathbb{P}\Big(\frac{S_n}{\sqrt{n}} \leq t\Big) - \mathbb{P}(N \leq t)\Big| \,\leq\, \Big(\frac{2}{\sqrt{n}}\,\mathbb{E}\big(|X|^3\big)\Big)^{\frac{1}{2}}.$$

As announced, due to the square root, this is not the Berry-Esseen theorem. It may however be reached with some more efforts along the same lines (see e.g. [18]).

4 Stein's inequality for multivariate Gaussian variables

The question of analogues of the Stein inequality (5) in higher dimension has been raised in various studies and context. The delicate point is that Stein's equation in a multivariate setting is not always explicitly solvable. Nevertheless, some substitutes may be considered.

Let γ_n be the standard Gaussian measure on the Borel sets of $\mathbb{R}^n$ with density $\frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}|x|^2}$, $x \in \mathbb{R}^n$, with respect to the Lebesgue measure.

To address Stein's equation in $\mathbb{R}^n$, the notion of Stein kernel associated to an unknown distribution is of significant interest. Given a centered probability measure μ on $\mathbb{R}^n$, a Stein kernel of μ is a measurable matrix-valued map τ^μ on $\mathbb{R}^n$ such that for every smooth test function $g : \mathbb{R}^n \to \mathbb{R}$,

$$\int_{\mathbb{R}^n} x \cdot \nabla g \, d\mu \, = \, \int_{\mathbb{R}^n} \tau^\mu \cdot \nabla^2 g \, d\mu$$

where ∇g stands for the gradient of g, with the scalar product between vectors in $\mathbb{R}^n$, and $\nabla^2 g$ stands for the Hessian of g, with the Hilbert-Schmidt scalar product between (symmetric) $n \times n$ matrices. The choice for g of the coordinate maps $x \mapsto x_k$, $k = 1, \ldots, n$, justifies the centering hypothesis. With respect to the differential equation (3), the picture here lies at a second differential order.

Stein kernels appear implicitly in the literature about Stein's method (see the original monograph [22, Lecture VI] of C. Stein, as well as [8, 9, 12, 13]...), while second order operators in a multivariate setting were considered in [7, 14, 17].... They gained momentum in connection with probabilistic approximations involving random variables living on a Gaussian (Wiener) space in [18].

According to the standard Gaussian integration by parts formula

$$\int_{\mathbb{R}^n} x_k \, h \, d\gamma_n \, = \, \int_{\mathbb{R}^n} \partial_k h \, d\gamma_n$$

applied $h = \partial_k g$, $k = 1, \ldots, n$, the identity matrix Id in $\mathbb{R}^n$ is a Stein kernel for γ_n. The proximity of τ^μ with Id thus indicates that μ should be close to the Gaussian distribution γ_n. Therefore, whenever such a Stein kernel τ^μ exists, the quantity, called Stein discrepancy (of μ with respect to γ_n, and associated to the underlying kernel τ^μ),

$$\mathrm{S}_2\big(\mu \,|\, \gamma_n\big) \, = \, \bigg(\int_{\mathbb{R}^n} |\tau^\mu - \mathrm{Id}|^2 \, d\mu \bigg)^{\frac{1}{2}}$$

(with $|\cdot|$ the Hilbert-Schmidt norm) becomes relevant as a measure of the proximity of μ and γ_n.

In dimension one, Stein's inequality (5) (with $f = g'$) precisely indicates that

$$\|\mu - \gamma_1\|_{\mathrm{TV}} \, \leq \, 2 \int_{\mathbb{R}} |\tau^\mu - 1| d\mu,$$

and therefore, by Jensen's inequality,

$$\|\mu - \gamma_1\|_{\mathrm{TV}} \, \leq \, 2 \, \mathrm{S}_2\big(\mu \,|\, \gamma_1\big),$$

justifying the interest in the Stein discrepancy.

It is the purpose of the following proposition from [16] (after the prior investigation [17] for the W_1 Wasserstein distance) to emphasize the corresponding inequality in $\mathbb{R}^n$ with respect to the Kantorovich metric W_2. Given probability measures μ and ν on the Borel sets of $\mathbb{R}^n$ with a finite second moment, let

$$\mathrm{W}_2(\mu,\nu) \,=\, \inf_{\pi} \bigg(\int_{\mathbb{R}^n \times \mathbb{R}^n} |x-y|^2 d\pi(x,y) \bigg)^{\frac{1}{2}},$$

where the infimum is taken over all couplings π on $\mathbb{R}^n \times \mathbb{R}^n$ with respective marginals μ and ν, be the quadratic Kantorovich (Wasserstein) distance between μ and ν.

Proposition 2 (A multivariate version of Stein's inequality)**.** *In the preceding notation,*

$$\mathrm{W}_2\big(\mu, \gamma_n\big) \,\le\, \mathrm{S}_2\big(\mu \,|\, \gamma_n\big).$$

For such a result to be useful and of interest, it is necessary to determine and describe suitable kernels τ^μ of the probability μ to be approximated by the Gaussian distribution γ_n. In dimension $n = 1$, if μ has a density ρ with respect to the Lebesgue measure on $\mathbb{R}$, the Stein kernel τ^μ is uniquely determined (up to sets of zero Lebesgue measure), and under standard regularity assumptions on ρ, a version of τ^μ is given by

$$\tau^\mu(x) \,=\, \frac{1}{\rho(x)} \int_x^\infty y \rho(y) dy$$

for x inside the support of ρ. In higher dimension, Stein kernels are not always unique and even may not exist.

5 Second order Poincaré inequalities

In several illustrations and applications, the unknown probability measure μ is actually of more concrete nature, allowing for explicit descriptions of a kernel. A typical instance is the example of the law μ of $F(X)$ where

$$F \,=\, (F_1, \ldots, F_n) : \mathbb{R}^N \,\to\, \mathbb{R}^n$$

is measurable and X is a standard Gaussian random vector on $\mathbb{R}^N$ (on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$). In other words, μ, as a probability measure on the Borel sets of $\mathbb{R}^n$, is the law of F under γ_N, and the following studies its proximity to the standard Gaussian distribution γ_n on $\mathbb{R}^n$. It is possible to consider more general distributions for X, such as the Wiener

measure in infinite dimension [18]. In most applications, the function F is also assumed to be reasonably regular in order to perform a number of differentiation and integration by parts operations, smoothness that will always be implicit below (polynomials is a class of examples). Since μ should be centered, $\mathbb{E}(F(X)) = 0$.

In this setting, a Stein kernel τ^μ for the law μ of $F(X) = (F_1, \ldots, F_n)(X)$ on $\mathbb{R}^n$ may be represented by a regular version of the conditional (matrix-valued) expectation

$$\mathbb{E}\big(T(X) \,|\, F(X)\big)$$

of $T(X)$ with respect to the σ-field generated by $F(X)$, where $T = \nabla(-\mathrm{L})^{-1}F \cdot \nabla F$ with L the Ornstein-Uhlenbeck generator on $\mathbb{R}^N$ (cf. [3]). Proposition 2 therefore yields

$$\begin{aligned} \mathrm{W}_2\big(\mu, \gamma_n\big)^2 \,&\leq\, \int_{\mathbb{R}^d} |\tau^\mu - \mathrm{Id}|^2 \, d\mu \\ &=\, \mathbb{E}\Big(\big|\mathbb{E}(T(X) \,|\, F(X)) - \mathrm{Id}\big|^2\Big) \\ &\leq\, \mathbb{E}\big(|T(X) - \mathrm{Id}|^2\big) \,=\, \int_{\mathbb{R}^N} |T - \mathrm{Id}|^2 d\gamma_N \end{aligned} \tag{6}$$

after the use of Jensen's inequality in the conditional expectation.

The form of T is of particular interest for eigenfunctions of L, as developed in the works by I. Nourdin and G. Peccati in their investigation of asymptotics of multiple stochastic integrals and Wiener chaos [20, 18]. In general, the inverse operator $(-\mathrm{L})^{-1}$ embedded in the definition of T may be analyzed via the underlying Ornstein-Uhlenbeck semigroup with infinitesimal generator L, to provide handful expressions (cf. [3]). The following statement is one illustration of what may be achieved along these lines, in a form which has taken the name of second order Poincaré inequalities [8, 19] (as the Gaussian Poincaré inequality [4] is used at the level of the gradients along the Ornstein-Uhlenbeck semigroup).

Proposition 3. *In the preceding notation, provided that $F(X)$, with law μ on the Borel sets of $\mathbb{R}^n$, has covariance matrix the identity, and that $F : \mathbb{R}^N \to \mathbb{R}^n$ is smooth enough,*

$$\mathrm{W}_2\big(\mu, \gamma_n\big)^2 \,\leq\, 3 \left(\int_{\mathbb{R}^N} \left[\sum_{k=1}^n |\nabla F_k|^2 \right]^2 d\gamma_N \right)^{\frac{1}{2}} \left(\int_{\mathbb{R}^N} \left[\sum_{k=1}^n |\nabla^2 F_k|^2 \right]^2 d\gamma_N \right)^{\frac{1}{2}}. \tag{7}$$

Arbitrary covariances are considered in [19]. The preceding statement applies in dimension $n = 1$ also for the total variation distance. Inequalities such as (7) have been exploited in [19] to study central limit theorems on Wiener space (cf. [18]) and in [8] to control the distance of the law of traces of random matrices to the Gaussian distribution.

Variations on Proposition 3 have been illustrated in [15] in the study of rates of convergence of linear statistics along polynomials of the spectral measure of random matrices from the Gaussian Unitary Ensemble [5].

References

[1] Some basics on Gaussian measures and variables. *The Gaussian Blog*.

[2] The Central Limit Theorem. *The Gaussian Blog*.

[3] Mehler kernel, and the Ornstein-Uhlenbeck operator. *The Gaussian Blog*.

[4] The Gaussian Poincaré inequality. *The Gaussian Blog*.

[5] Gaussian random matrix ensembles. *The Gaussian Blog*.

[6] *An introduction to Stein's method.* A. Barbour, L. Chen, Editors. Lecture Notes Series, Institute for Mathematical Sciences, National University of Singapore, Vol. 4. Singapore University Press (2005).

[7] A. Barbour. Stein's method for diffusion approximations. *Probab. Theory Rel. Fields* 84, 297–322 (1990).

[8] S. Chatterjee. Fluctuations of eigenvalues and second order Poincaré inequalities. *Probab. Theory Related Fields* 143, 1–40 (2009).

[9] S. Chatterjee. A short survey of Stein's method. *Kyung Moon Sa*, 1–24. Seoul (2014).

[10] L. Chen. Stein's method of normal approximation: some recollections and reflections. *Ann. Statist.* 49, 1850–1863 (2021).

[11] L. Chen, L. Goldstein, Q.-M. Shao. *Normal approximation by Stein's method.* Probability and its Applications. Springer (2011).

[12] L. Goldstein, G. Reinert. Stein's method and the zero bias transformation with application to simple random sampling. *Ann. Appl. Probab.* 7, 935–952 (1997).

[13] L. Goldstein, G. Reinert. Zero biasing in one and higher dimensions, and applications. *Stein's method and applications*, Lect. Notes Ser. Inst. Math. Sci. Natl. Univ. Singap. 5, 1–18. Singapore Univ. Press (2005).

[14] F. Götze. On the rate of convergence in the multivariate CLT. *Ann. Probab.* 19, 724–739 (1991).

[15] G. Lambert, M. Ledoux, C. Webb. Quantitative normal approximation of linear statistics of beta-ensembles. *Ann. Probab.* 47, 2619–2685 (2019).

[16] M. Ledoux, I. Nourdin, G. Peccati. Stein's method, logarithmic Sobolev and transport inequalities. *Geom. and Funct. Anal.* 25, 256–306 (2015).

[17] E. Meckes. On Stein's method for multivariate normal approximation. *High dimensional probability V: the Luminy volume*, 153–178, Inst. Math. Stat. Collection 5 (2009).

[18] I. Nourdin, G. Peccati. *Normal approximations with Malliavin calculus: from Stein's method to universality*. Cambridge Tracts in Mathematics. Cambridge University Press (2012).

[19] I. Nourdin, G. Peccati, G. Reinert. Second order Poincaré inequalities and CLTs on Wiener space. *J. Funct. Anal.* 257, 593–609 (2009).

[20] D. Nualart, G. Peccati. Central limit theorems for sequences of multiple stochastic integrals. *Ann. Probab.* 33, 177–193 (2005).

[21] C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 2, 583–602. University of California Press (1972).

[22] C. Stein. *Approximate computation of expectations*. Institute of Mathematical Statistics Lecture Notes – Monograph Series 7. Institute of Mathematical Statistics (1986).