

# Wasserstein bounds for non-linear Gaussian filters

Toni Karvonen and Simo Särkkä

**Abstract**—Most Kalman filters for non-linear systems, such as the unscented Kalman filter, are based on Gaussian approximations. We use Poincaré inequalities to bound the Wasserstein distance between the true joint distribution of the prediction and measurement and its Gaussian approximation. The bounds can be used to assess the performance of non-linear Gaussian filters and determine those filtering approximations that are most likely to induce error.

**Index Terms**—Kalman filtering, Gaussian filtering, non-linear filtering, Wasserstein distance, Poincaré inequalities

## I. INTRODUCTION

THE Kalman filter and its non-linear extensions have been ubiquitous tools for estimating the latent state of partially observed state-space Markov models since the 1960s. Their applications include, to name but a few relatively recent ones, various types of tracking such as vision-aided [1] and pedestrian tracking [2], elimination of noise from brain imaging data [3], and simultaneous localization and mapping [4].

The Kalman filter is the optimal linear estimator in the mean-square sense when the system is linear Gaussian [5]. Moreover, under natural observability and controllability assumptions the linear Kalman filter exhibits strong exponential stability properties; see [6], [7], and [8, App. C]. However, useful theoretical results for popular extensions of the Kalman filter for non-linear Gaussian systems, such as the extended and unscented Kalman filters, have proved far more elusive. While Lipschitzianity assumptions suffice for understanding the propagation of the filtering distribution through the system dynamics, the update step in which the distribution is conditioned on new measurements is much more difficult to analyse—even over one time-step. Accordingly, most error analysis for non-linear Kalman filters requires essentially unverifiable assumptions on error covariances and other filtering quantities (e.g., [9]–[11]) or assumes that the measurement model is linear [12].

Many non-linear Kalman filters are based on the idea of Gaussian moment-matching. These *Gaussian filters* replace all distributions that occur in filtering with moment-matched Gaussians that are far easier to operate with; we defer a detailed description to Section II. To moment-match one has to compute integrals that are rarely available in closed form. This has given rise to a class of non-linear filters that combine Gaussian moment-matching with numerical integration (see [13] and [14, Ch. 8]). The unscented Kalman filter is undoubtedly the most popular member of this class of filters.

T. Karvonen was supported by the Research Council of Finland grants 359183 and 368086. He also acknowledges the research environment provided by ELLIS Institute Finland.

T. Karvonen is with the School of Engineering Sciences, Lappeenranta–Lahti University of Technology LUT, Lappeenranta, Finland (email: toni.karvonen@lut.fi).

S. Särkkä is with the Department of Electrical Engineering and Automation, Aalto University, Espoo, Finland (email: simo.sarkka@aalto.fi).

## A. Prior work and contributions

In this article we study the accuracy of Gaussian filters over one time-step. The only prior work on assessing the effect of moment-matching we are aware of is by Morelande and García-Fernández [15]. By assuming a prediction distribution  $\mathbf{X} \sim \mathcal{N}(\mathbf{m}, \mathbf{P})$  (i.e.,  $\mathbf{X}$  is a Gaussian random vector with mean  $\mathbf{m}$  and covariance  $\mathbf{P}$ ) and a non-linear measurement model  $\mathbf{Y} = \mathbf{h}(\mathbf{X}) + \mathbf{V}$ , where  $\mathbf{V} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\mathbf{V}})$ , they compute that the Kullback–Leibler divergence between the moment-matched Gaussian approximation  $\tilde{\mathbf{Z}} = (\mathbf{X}, \tilde{\mathbf{Y}})$  and the true joint distribution  $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$  is<sup>1</sup>

$$D_{\text{KL}}(\mathbf{Z} \parallel \tilde{\mathbf{Z}}) = \frac{1}{2} \log \left[ \det \left( \mathbf{I} + \Sigma_{\mathbf{V}}^{-1} [\text{Cov}[\mathbf{Y}] - \Sigma_{\mathbf{V}} - \text{Cov}(\mathbf{X}, \mathbf{Y})^{\top} \mathbf{P}^{-1} \text{Cov}(\mathbf{X}, \mathbf{Y})] \right) \right]. \quad (1)$$

However, since the Kullback–Leibler divergence is not a metric, it is difficult to incorporate other errors, such as those propagated from previous time-steps or due to numerical integration in an unscented Kalman filter.

The aim of this article is to obtain upper bounds like (1) but in terms of the Wasserstein distance (see Section III) rather than the Kullback–Leibler divergence. To bound the Wasserstein distance between a non-linearly transformed Gaussian distribution and its moment-matched approximation we use recent second-order Poincaré inequalities, to be reviewed in Section IV. In Section V, we derive an upper bound, given in Theorem 8, for the Wasserstein distance between the true joint distribution of the prediction and measurement and that produced by the exact Gaussian Kalman filter. Unlike in [15], our bound is *over both the prediction and update step*. The bound is given in terms of expectations of the first and second derivatives of the dynamic and measurement model functions as well as the difference of real and approximated joint means and covariances. The bound vanishes when the model is affine, in which case the prediction and measurement distributions are Gaussian. Section VI contains a numerical example. While the bound is very conservative, our results will help in gauging the worst-case performance of Gaussian filters in safety-critical applications and in determining which approximations in non-linear filtering are most likely to induce large errors. We also hope that the results will bring about a new type of stability analysis for Gaussian filters.

## B. Notational conventions

Gradient of a function  $g: \mathbb{R}^d \rightarrow \mathbb{R}$  is a column-vector function  $\nabla g$  with elements  $[\nabla g]_i = \partial g / \partial x_i$ . For clarity, the gradient of a function  $g$  evaluated at  $\mathbf{f}(\mathbf{x})$  is occasionally written  $(\nabla g)[\mathbf{f}(\mathbf{x})]$ . The Hessian has elements  $[\mathbf{H}_g]_{ij} = \partial^2 g / \partial x_i \partial x_j$ .

<sup>1</sup>This is Eq. (23) of [15], which erroneously omits the coefficient 1/2.

For a function  $\mathbf{g}: \mathbb{R}^d \rightarrow \mathbb{R}^d$  with components  $g_i$ , the Jacobian has elements  $[\mathbf{J}_{\mathbf{g}}]_{ij} = \partial g_i / \partial x_j$ .

We use the Löwner partial ordering of positive-semidefinite matrices. For matrices  $\mathbf{A}$  and  $\mathbf{B}$ ,  $\mathbf{A} > \mathbf{B}$  ( $\mathbf{A} \geq \mathbf{B}$ ) means that  $\mathbf{A} - \mathbf{B}$  is positive-definite (positive-semidefinite). See [16, Ch. 8] for a review of properties of the Löwner ordering.

Norm is denoted  $\|\cdot\|$ . Depending on the object, this stands for the Euclidean norm  $\|\mathbf{x}\|$  of a vector  $\mathbf{x} \in \mathbb{R}^d$  or the spectral norm of a matrix  $\mathbf{A} \in \mathbb{R}^{d \times d}$ , which is defined as

$$\|\mathbf{A}\| = \sup\{\|\mathbf{Ax}\| / \|\mathbf{x}\| : \mathbf{0} \neq \mathbf{x} \in \mathbb{R}^d\} = \sqrt{\lambda_{\max}(\mathbf{A}^T \mathbf{A})},$$

where  $\lambda_{\max}$  denotes the largest eigenvalue. The  $L^p$ -norm  $\|\mathbf{X}\|_p$  of a random variable  $\mathbf{X}$  is  $\|\mathbf{X}\|_p = (\mathbb{E}[\|\mathbf{X}\|^p])^{1/p}$ . If  $\mathbf{A}$  is positive-definite,  $\sqrt{\mathbf{A}}$  stands for the unique positive-definite matrix such that  $\sqrt{\mathbf{A}}\sqrt{\mathbf{A}} = \mathbf{A}$ .

## II. NON-LINEAR GAUSSIAN FILTERING

In this article we consider a non-linear dynamic system

$$\mathbf{X} = \mathbf{f}(\mathbf{X}_p) + \mathbf{U}, \quad (2)$$

$$\mathbf{Y} = \mathbf{h}(\mathbf{X}) + \mathbf{V}, \quad (3)$$

where the *state*  $\mathbf{X} \in \mathbb{R}^n$  is obtained by propagating the  $n$ -dimensional Gaussian prior  $\mathbf{X}_p \sim \mathcal{N}(\mathbf{m}_p, \mathbf{P}_p)$  through the *dynamic model function*  $\mathbf{f}: \mathbb{R}^n \rightarrow \mathbb{R}^n$ . The *measurement*  $\mathbf{y} \in \mathbb{R}^m$  is a realization of the random variable  $\mathbf{Y}$  that is obtained via the *measurement model function*  $\mathbf{h}: \mathbb{R}^n \rightarrow \mathbb{R}^m$ . The state and measurement are corrupted by additive Gaussian noises  $\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\mathbf{U}})$  and  $\mathbf{V} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\mathbf{V}})$ . The covariance matrices  $\mathbf{P}_p$ ,  $\Sigma_{\mathbf{U}}$ , and  $\Sigma_{\mathbf{V}}$  are assumed positive-definite and the random variables  $\mathbf{X}_p$ ,  $\mathbf{U}$  and  $\mathbf{V}$  independent. The distribution of  $\mathbf{X}$  is called the *prediction distribution*. Unfortunately, the *filtering distribution*  $\mathbf{X} | \mathbf{Y} = \mathbf{y}$  is Gaussian only if the dynamic and measurement model functions are affine.

We use  $\text{pr}$  to denote Gaussian projection that transforms a given random variable into a Gaussian one while preserving the mean and covariance. That is, for an arbitrary random variable  $\mathbf{Z}$  with finite mean  $\mathbb{E}(\mathbf{Z})$  and covariance  $\text{Cov}(\mathbf{Z})$  we have

$$\text{pr } \mathbf{Z} \sim \mathcal{N}(\mathbb{E}(\mathbf{Z}), \text{Cov}(\mathbf{Z})), \quad (4)$$

which is to say that the moments of the Gaussian projection match those of the original random variable. In Gaussian filtering [14, Ch. 8], a Gaussian approximation to the filtering distribution is obtained via two projections:

1. *Prediction*. Approximate  $\mathbf{X}$  in (2) with the moment-matched Gaussian  $\tilde{\mathbf{X}} = \text{pr } \mathbf{X}$ . That is,

$$p(\mathbf{X}) \approx p(\tilde{\mathbf{X}}) = \mathcal{N}(\mathbf{m}_{\mathbf{X}}, \mathbf{P}_{\mathbf{X}}),$$

where  $\mathbf{m}_{\mathbf{X}} = \mathbb{E}(\mathbf{X}) = \mathbb{E}[\mathbf{f}(\mathbf{X}_p)]$  and

$$\mathbf{P}_{\mathbf{X}} = \text{Cov}(\mathbf{X}) = \text{Cov}[\mathbf{f}(\mathbf{X}_p)] + \Sigma_{\mathbf{U}}.$$

2. Use  $\tilde{\mathbf{X}}$  to construct a Gaussian approximation

$$\tilde{\mathbf{Z}} = (\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}) = \text{pr}(\tilde{\mathbf{X}}, \mathbf{h}(\tilde{\mathbf{X}}) + \mathbf{V})$$

of the joint distribution  $\mathbf{Z}$  of  $\mathbf{X}$  and  $\mathbf{Y}$ . The Gaussian approximation to  $\mathbf{Y}$  is then  $\tilde{\mathbf{Y}} \sim \mathcal{N}(\mathbf{m}_{\tilde{\mathbf{Y}}}, \mathbf{P}_{\tilde{\mathbf{Y}}})$ , where  $\mathbf{m}_{\tilde{\mathbf{Y}}} = \mathbb{E}(\tilde{\mathbf{Y}}) = \mathbb{E}[\mathbf{h}(\tilde{\mathbf{X}})]$  and

$$\mathbf{P}_{\tilde{\mathbf{Y}}} = \text{Cov}(\tilde{\mathbf{Y}}) = \text{Cov}[\mathbf{h}(\tilde{\mathbf{X}})] + \Sigma_{\mathbf{V}}.$$

Note that  $\tilde{\mathbf{Y}}$  does *not* equal  $\text{pr } \mathbf{Y}$ :

$$\tilde{\mathbf{Y}} = \text{pr}(\mathbf{h}(\text{pr } \mathbf{X}) + \mathbf{V}) \neq \text{pr}(\mathbf{h}(\mathbf{X}) + \mathbf{V}) = \text{pr } \mathbf{Y}.$$

The cross-covariance is

$$\mathbf{C}_{\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}} = \text{Cov}(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}) = \mathbb{E}[(\tilde{\mathbf{X}} - \mathbf{m}_{\tilde{\mathbf{X}}})(\mathbf{h}(\tilde{\mathbf{X}}) - \mathbf{m}_{\tilde{\mathbf{Y}}})^T].$$

3. *Update*. Use the standard Gaussian conditioning formulas to compute a Gaussian approximation to  $\mathbf{X} | \mathbf{Y} = \mathbf{y}$ :

$$p(\mathbf{X} | \mathbf{Y} = \mathbf{y}) \approx p(\tilde{\mathbf{X}} | \tilde{\mathbf{Y}} = \mathbf{y}) = \mathcal{N}(\mathbf{m}_{\tilde{\mathbf{X}}|\tilde{\mathbf{Y}}}, \mathbf{P}_{\tilde{\mathbf{X}}|\tilde{\mathbf{Y}}}),$$

where

$$\mathbf{m}_{\tilde{\mathbf{X}}|\tilde{\mathbf{Y}}} = \mathbf{m}_{\tilde{\mathbf{X}}} + \mathbf{C}_{\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}} \mathbf{P}_{\tilde{\mathbf{Y}}}^{-1} (\mathbf{y} - \mathbf{m}_{\tilde{\mathbf{Y}}}),$$

$$\mathbf{P}_{\tilde{\mathbf{X}}|\tilde{\mathbf{Y}}} = \mathbf{P}_{\tilde{\mathbf{X}}} - \mathbf{C}_{\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}} \mathbf{P}_{\tilde{\mathbf{Y}}}^{-1} \mathbf{C}_{\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}}^T.$$

Figure 1 shows the one-step structure of a Gaussian filter. The Gaussian approximation to the conditional distribution is then treated as  $\mathbf{X}_p$  and the procedure repeated for as many time-steps as called for. Within this scheme the Gaussian integrals

$$\int_{\mathbb{R}^n} \mathbf{g}_1(\mathbf{x}) \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \int_{\mathbb{R}^n} \mathbf{g}_1(\mathbf{x}) \mathbf{g}_2(\mathbf{x})^T \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

for  $\mathbf{g}_1, \mathbf{g}_2 \in \{\mathbf{f}, \mathbf{h}\}$  need to be computed, a task that is rarely possible analytically. In engineering literature, the most prevalent numerical integration methods for approximating these integrals are the fully symmetric cubature integration formulas of McNamee and Stenger [17], used in the unscented transform of Julier *et al.* [18]. See [13] and [14, Ch. 8] for other popular sigma-point methods. As our aim is not to compare accuracy or effect of different sigma-point schemes, the problem of integral computations is not taken into account in what follows (one interested in this aspect should see, e.g., [15] and [19]). That is, we consider an *exact Gaussian filter* which computes the moments exactly. Theorem 5, our main tool, would not even lend itself easily to inclusion of numerical integration errors. The triangle inequality could be used, but this would offer little additional insight.

## III. WASSERSTEIN DISTANCE

This section introduces the main metric of this article, the Wasserstein distance. Aside from the metrics discussed here, a great number of other metrics for measuring the distance between random variables exist [20]. We focus on the Wasserstein distance because Theorem 5, our main workhorse, is not available in sufficient generality for other distances. The  $p$ th *Wasserstein distance*  $W_p$ , often known as earth mover's distance, between random variables  $\mathbf{X}$  and  $\mathbf{Z}$  in  $\mathbb{R}^d$  is

$$W_p(\mathbf{X}, \mathbf{Z})^p = \inf \mathbb{E}(\|\mathbf{Y} - \mathbf{U}\|^p),$$

where the infimum is over all joint distributions of random variables  $\mathbf{Y}$  and  $\mathbf{U}$  whose marginal distributions coincide with those of  $\mathbf{X}$  and  $\mathbf{Z}$ . Wasserstein distances are important in the theory of optimal transport [21]. We use only the first Wasserstein distance  $W_1$  that will be from now simply called the Wasserstein distance and denoted  $d_W$ .

Many useful probability metrics are encompassed by the family of metrics defined as

$$d(\mathbf{X}, \mathbf{Z}) = \sup_{h \in \mathcal{H}} |\mathbb{E}[h(\mathbf{X})] - \mathbb{E}[h(\mathbf{Z})]|, \quad (5)$$

where  $\mathcal{H}$  is some class of functions from  $\mathbb{R}^d$  to  $\mathbb{R}$ . The two most important members of this class are the Wasserstein distance

$$d_W(\mathbf{X}, \mathbf{Z}) = W_1(\mathbf{X}, \mathbf{Z}) = \sup_{\text{Lip}(h) \leq 1} |\mathbb{E}[h(\mathbf{X})] - \mathbb{E}[h(\mathbf{Z})]|,$$

where  $\text{Lip}(h) = \sup_{\mathbf{x} \neq \mathbf{y}} |h(\mathbf{x}) - h(\mathbf{y})| / \|\mathbf{x} - \mathbf{y}\|$  is the *maximal Lipschitz constant*, and the *total variation distance*

$$d_{\text{TV}}(\mathbf{X}, \mathbf{Z}) = \sup_{|h| \leq 1/2} |\mathbb{E}[h(\mathbf{X})] - \mathbb{E}[h(\mathbf{Z})]|.$$

Note that the total variation distance is at most one and is thus more suitable for convergence analysis than for our purposes.

The following propositions can be found in [22]. When combined, they provide an upper bound on the Wasserstein distance between two Gaussian random variables.

**Proposition 1.** *If  $1 \leq q \leq p$ , then  $W_q(\mathbf{X}, \mathbf{Z}) \leq W_p(\mathbf{X}, \mathbf{Z})$  for any random variables  $\mathbf{X}$  and  $\mathbf{Z}$ .*

**Proposition 2.** *Let  $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X)$  and  $\mathbf{Z} \sim \mathcal{N}(\boldsymbol{\mu}_Z, \boldsymbol{\Sigma}_Z)$  with positive-definite  $\boldsymbol{\Sigma}_X$  and  $\boldsymbol{\Sigma}_Z$ . Set  $\mathbf{L}_X = \boldsymbol{\Sigma}_X^{1/2}$ . Then*

$$W_2(\mathbf{X}, \mathbf{Z})^2 = \|\boldsymbol{\mu}_X - \boldsymbol{\mu}_Z\|^2 + \text{tr} \boldsymbol{\Sigma}_X + \text{tr} \boldsymbol{\Sigma}_Z - 2 \text{tr} (\sqrt{\mathbf{L}_X \boldsymbol{\Sigma}_Z \mathbf{L}_X}).$$

In the notation of Proposition 2, these propositions yield

$$d_W(\mathbf{X}, \mathbf{Z})^2 \leq \|\boldsymbol{\mu}_X - \boldsymbol{\mu}_Z\|^2 + \text{tr} \boldsymbol{\Sigma}_X + \text{tr} \boldsymbol{\Sigma}_Z - 2 \text{tr} (\sqrt{\mathbf{L}_X \boldsymbol{\Sigma}_Z \mathbf{L}_X}). \quad (6)$$

Ley *et al.* [23] have derived an upper bound for the Wasserstein distance between two univariate Gaussians directly, without using Proposition 1. However, in the multivariate case bounds derived in this way seem to exist only for centered Gaussians [24, Exr. 4.5.3].

#### IV. STEIN'S METHOD AND POINCARÉ INEQUALITIES

This section introduces the two inequalities essential for the developments of this article. These are the *Poincaré inequalities* of Theorems 3 and 5 that provide means of controlling a transformed Gaussian random variable in terms of expectations of first and second derivatives of the transformation. These inequalities, particularly the second one, are closely linked to *Stein's method*, the basics of which we sketch next.

Stein's method has its basis on the familiar identity

$$\mathbb{E}[Xg(X)] = \mathbb{E}[g'(X)], \quad (7)$$

often called *Stein's identity* or *Stein's lemma*, that holds for every differentiable function  $g: \mathbb{R} \rightarrow \mathbb{R}$  if and only if  $X \sim \mathcal{N}(0, 1)$ . This identity allows one to assess the Gaussianity of a transformed random variable via the differential equation

$$g'(z) - zg(z) = h(z) - \mathbb{E}[h(X)],$$

where  $h: \mathbb{R} \rightarrow \mathbb{R}$  is differentiable. For a solution  $g_h$  to this differential equation and any random variable  $Z$  we have

$$\mathbb{E}[h(X)] - \mathbb{E}[h(Z)] = \mathbb{E}[g'_h(Z)] - \mathbb{E}[Zg_h(Z)],$$

where the right-hand side vanishes by (7) if  $Z \sim \mathcal{N}(0, 1)$ . In the univariate case, Equation (5) can be thus written as

$$\begin{aligned} d_{\mathcal{H}}(X, Z) &= \sup_{h \in \mathcal{H}} |\mathbb{E}[h(X)] - \mathbb{E}[h(Z)]| \\ &= \sup_{h \in \mathcal{H}} |\mathbb{E}[g'_h(Z)] - \mathbb{E}[Zg_h(Z)]|, \end{aligned}$$

which depends on  $X$  only via  $g_h$ . Stein's method has its origins in the work of Stein [25]. See [24], [26], and [27] for reviews.

First of the inequalities we are interested in is variously known as *Poincaré inequality* [24] or *Chernoff's inequality* [28] and provides variance bounds for transformed Gaussian random variables. This inequality originates in the works of Nash [29, Part II] and Chernoff [30], with proofs based on a variance expansion in terms of Hermite polynomials. The multivariate version below is due to Cacoullos [28].

**Theorem 3** (Poincaré inequality; [28]). *If  $\mathbf{g}: \mathbb{R}^{d_1} \rightarrow \mathbb{R}^{d_2}$  is differentiable and  $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , then*

$$\mathbb{E}[\mathbf{J}_{\mathbf{g}}(\mathbf{X})] \boldsymbol{\Sigma} \mathbb{E}[\mathbf{J}_{\mathbf{g}}(\mathbf{X})]^\top \leq \text{Cov}[\mathbf{g}(\mathbf{X})] \leq \mathbb{E}[\mathbf{J}_{\mathbf{g}}(\mathbf{X}) \boldsymbol{\Sigma} \mathbf{J}_{\mathbf{g}}(\mathbf{X})^\top].$$

The bounds coincide if and only if  $\mathbf{g}$  is affine, which is to say that  $\mathbf{g}(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$  for some  $\mathbf{A} \in \mathbb{R}^{d_2 \times d_1}$  and  $\mathbf{b} \in \mathbb{R}^{d_2}$ .

Recent *second order Poincaré inequalities* are essential for our purposes. These inequalities bound distances between  $\mathbf{g}(\mathbf{X})$  and  $\text{pr} \mathbf{g}(\mathbf{X})$  when  $\mathbf{X}$  is Gaussian. The first such inequality, proved by Chatterjee [31], was for the total variation distance but required that  $\mathbf{g}$  be a mapping to the real line. Multivariate extensions for the Wasserstein distance appear in [32] and [33]. The version in Theorem 5 below is a corollary of [32, Thm. 7.1].

**Definition 4.** Let  $\mathbf{X} \in \mathbb{R}^{d_1}$  be a random variable. The set of twice-differentiable functions  $\mathbf{g}: \mathbb{R}^{d_1} \rightarrow \mathbb{R}^{d_2}$  such that  $\mathbf{g}(\mathbf{X})$ ,  $\nabla g_i(\mathbf{X})$  and  $\mathbf{H}_{g_i}(\mathbf{X})$  have finite  $L^4$ -norms for all  $i = 1, \dots, d_2$  is denoted by  $\mathcal{L}(\mathbf{X})$ .

For notational convenience we define

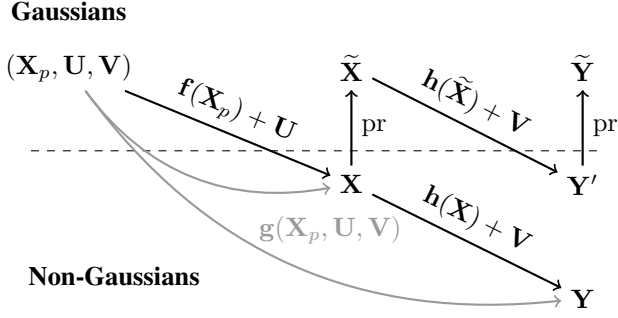
$$\begin{aligned} \|\nabla \mathbf{g}(\mathbf{X})\|_{4'} &= \sum_{i=1}^{d_2} (\mathbb{E}[\|\nabla g_i(\mathbf{X})\|^4])^{1/4}, \\ \|\mathbf{H}_{\mathbf{g}}(\mathbf{X})\|_{4'} &= \sum_{i=1}^{d_2} (\mathbb{E}[\|\mathbf{H}_{g_i}(\mathbf{X})\|^4])^{1/4} \end{aligned}$$

when  $\mathbf{g}: \mathbb{R}^{d_1} \rightarrow \mathbb{R}^{d_2}$ . Note that due to the placement of the root these are not conventional  $L^4$ -norms but sums of  $L^4$ -norms. These quantities are finite if  $\mathbf{g} \in \mathcal{L}(\mathbf{X})$ .

**Theorem 5** (2nd order Poincaré inequality; [32]). *Let  $\mathbf{X} \in \mathbb{R}^{d_1}$  be a Gaussian random variable and  $\mathbf{g}: \mathbb{R}^{d_1} \rightarrow \mathbb{R}^{d_2}$  for  $d_1 \geq d_2$  a twice differentiable function. Suppose that  $\mathbf{g} \in \mathcal{L}(\mathbf{X})$  and that  $\mathbf{g}(\mathbf{X})$  has mean  $\boldsymbol{\mu}$  and positive-definite covariance  $\boldsymbol{\Sigma}$ . Then*

$$\begin{aligned} d_W(\mathbf{g}(\mathbf{X}), \text{pr} \mathbf{g}(\mathbf{X})) &\leq \frac{3}{\sqrt{2}} \|\boldsymbol{\Sigma}^{-1}\| \|\boldsymbol{\Sigma}\|^{1/2} \\ &\quad \times \|\mathbf{H}_{\mathbf{g}}(\mathbf{X})\|_{4'} \|\nabla \mathbf{g}(\mathbf{X})\|_{4'}. \end{aligned}$$

Interpreting this inequality is straightforward: if the Hessians are non-zero, the distance is positive since  $\mathbf{g}(\mathbf{X})$  is not Gaussian; if the Hessians vanish,  $\mathbf{g}$  is affine, so that  $\mathbf{g}(\mathbf{X})$  is Gaussian and  $\text{pr} \mathbf{g}(\mathbf{X}) = \mathbf{g}(\mathbf{X})$ .



**Fig. 1:** Structure and relations of different random variables and their Gaussian approximations in Sections II and V. The random variables on the upper part of the figure are Gaussian while the ones on the lower part are not unless  $\mathbf{f}$  and  $\mathbf{h}$  are affine. The Gaussian projection  $\text{pr}$  is defined in (4) and the function  $\mathbf{g}$  in Equation (8).

## V. WASSERSTEIN BOUNDS FOR GAUSSIAN FILTERS

This section contains the main results of the article. We establish upper bounds on the Wasserstein distance  $d_W(\mathbf{Z}, \tilde{\mathbf{Z}})$  between the true joint distribution  $\mathbf{Z}$  of  $\mathbf{X}$  and  $\mathbf{Y}$  and the Gaussian filter approximation  $\tilde{\mathbf{Z}} = (\tilde{\mathbf{X}}, \tilde{\mathbf{Y}})$  constructed as described in Section II. To use Theorem 5 we need to write  $\mathbf{Z}$  as a transformation of a Gaussian random variable. This can be done by defining a transformation  $\mathbf{g}: \mathbb{R}^{2n+m} \rightarrow \mathbb{R}^{n+m}$  as

$$\mathbf{g}(\mathbf{x}) = \mathbf{g} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \end{pmatrix} = \begin{pmatrix} \mathbf{f}(\mathbf{x}_1) + \mathbf{x}_2 \\ \mathbf{q}(\mathbf{x}_1, \mathbf{x}_2) + \mathbf{x}_3 \end{pmatrix} \quad (8)$$

where  $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$ ,  $\mathbf{x}_3 \in \mathbb{R}^m$  and

$$\mathbf{q}(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{h}[\mathbf{f}(\mathbf{x}_1) + \mathbf{x}_2]. \quad (9)$$

It then follows that

$$\mathbf{Z} = (\mathbf{X}, \mathbf{Y}) = \mathbf{g}(\mathbf{W}) = \mathbf{g}(\mathbf{X}_p, \mathbf{U}, \mathbf{V}),$$

where  $\mathbf{W} = (\mathbf{X}_p, \mathbf{U}, \mathbf{V}) \in \mathbb{R}^{2n+m}$  is Gaussian.

As is apparent from the structure of a Gaussian filter depicted in Figure 1, the mean and covariance of  $\mathbf{Z}$  do not match those of  $\tilde{\mathbf{Z}}$ , which means that Theorem 5 cannot be applied directly.<sup>2</sup> The triangle inequality and (6) tackle this problem:

$$d_W(\mathbf{Z}, \tilde{\mathbf{Z}}) \leq d_W(\mathbf{Z}, \text{pr } \mathbf{Z}) + d_W(\text{pr } \mathbf{Z}, \tilde{\mathbf{Z}}). \quad (10)$$

Theorem 5 is now applicable to the first term and the bound (6) to the second. We begin by bounding the true joint covariance  $\text{Cov}(\mathbf{Z})$  with the Poincaré variance inequality of Theorem 3.

**Lemma 6.** *If  $\mathbf{f}$  and  $\mathbf{h}$  are differentiable, then the joint covariance  $\text{Cov}(\mathbf{Z})$  obeys the Löwner bounds*

$$\mathbf{A}_1 + \mathbf{B}_1 + \mathbf{C}_V \leq \text{Cov}(\mathbf{Z}) \leq \mathbf{A}_2 + \mathbf{C}_V, \quad (11)$$

where

$$\mathbf{A}_1 = \begin{pmatrix} \mathbb{E}[\mathbf{J}_f(\mathbf{X}_p)] \\ \mathbb{E}[\mathbf{J}_h[\mathbf{f}(\mathbf{X}_p) + \mathbf{U}]] \mathbf{J}_f(\mathbf{X}_p) \end{pmatrix} \mathbf{P}_p$$

<sup>2</sup>We note that one can construct, in a manner reminiscent of the *augmented* unscented transform [34], a filter that is based on a direct Gaussian approximation of  $\mathbf{Z} = \mathbf{g}(\mathbf{X}_p, \mathbf{U}, \mathbf{V})$  instead of an approximation derived from  $\text{pr } \mathbf{X}$  (see Figure 1). This is rarely done in practice. However, in the augmented case it suffices to bound  $d_W(\mathbf{Z}, \text{pr } \mathbf{Z})$ , which is done in Lemma 7.

$$\times \left( \mathbb{E}[\mathbf{J}_f(\mathbf{X}_p)]^\top \quad \mathbb{E}[\mathbf{J}_h[\mathbf{f}(\mathbf{X}_p) + \mathbf{U}]] \mathbf{J}_f(\mathbf{X}_p)^\top \right)^\top,$$

$$\begin{aligned} \mathbf{B}_1 &= \begin{pmatrix} \mathbf{I}_{n \times n} \\ \mathbb{E}[\mathbf{J}_h[\mathbf{f}(\mathbf{X}_p) + \mathbf{U}]] \end{pmatrix} \Sigma_U \\ &\times \begin{pmatrix} \mathbf{I}_{n \times n} & \mathbb{E}[\mathbf{J}_h[\mathbf{f}(\mathbf{X}_p) + \mathbf{U}]]^\top \end{pmatrix}, \\ \mathbf{A}_2 &= \mathbb{E} \left[ \begin{pmatrix} \mathbf{I}_{n \times n} \\ \mathbf{J}_h[\mathbf{f}(\mathbf{X}_p) + \mathbf{U}] \end{pmatrix} [\mathbf{J}_f(\mathbf{X}_p) \mathbf{P}_p \mathbf{J}_f(\mathbf{X}_p)^\top + \Sigma_U] \right. \\ &\quad \left. \times \begin{pmatrix} \mathbf{I}_{n \times n} & \mathbf{J}_h[\mathbf{f}(\mathbf{X}_p) + \mathbf{U}]^\top \end{pmatrix} \right], \\ \mathbf{C}_V &= \begin{pmatrix} \mathbf{0}_{n \times n} & \mathbf{0}_{n \times m} \\ \mathbf{0}_{m \times n} & \Sigma_V \end{pmatrix}. \end{aligned} \quad (12)$$

*Proof.* The Jacobian of  $\mathbf{g}$  in Equation (8) is

$$\mathbf{J}_g(\mathbf{x}) = \begin{pmatrix} \mathbf{J}_f(\mathbf{x}_1) & \mathbf{I}_{n \times n} & \mathbf{0}_{n \times m} \\ \mathbf{J}_h[\mathbf{f}(\mathbf{x}_1) + \mathbf{x}_2] \mathbf{J}_f(\mathbf{x}_1) & \mathbf{J}_h[\mathbf{f}(\mathbf{x}_1) + \mathbf{x}_2] & \mathbf{I}_{m \times m} \end{pmatrix}$$

by the chain rule. We can now use Theorem 3 with  $\mathbf{X} = \mathbf{W}$ . Because the prior and noise terms are assumed independent, the covariance matrix  $\text{Cov}(\mathbf{W}) = \text{diag}(\mathbf{P}_p, \Sigma_U, \Sigma_V)$  is block-diagonal. Using the expression for  $\mathbf{J}_g$  above, we compute the block matrices  $\mathbb{E}[\mathbf{J}_g(\mathbf{W})] \text{Cov}(\mathbf{W}) \mathbb{E}[\mathbf{J}_g(\mathbf{W})]^\top$  and  $\mathbb{E}[\mathbf{J}_g(\mathbf{W}) \text{Cov}(\mathbf{W}) \mathbf{J}_g(\mathbf{W})^\top]$  needed in Theorem 3 and, after some factorization, arrive at (11).  $\square$

The following lemma takes care of  $d_W(\mathbf{Z}, \text{pr } \mathbf{Z})$ .

**Lemma 7.** *Suppose that  $\mathbf{g}$  in Equation (8) is in  $\mathcal{L}(\mathbf{W})$  from Definition 4. Let  $\mathbf{P}_Z$  denote the covariance of  $\mathbf{Z}$ . Then*

$$\begin{aligned} d_W(\mathbf{Z}, \text{pr } \mathbf{Z}) &\leq \frac{3}{\sqrt{2}} \|\mathbf{P}_Z^{-1}\| \|\mathbf{P}_Z\|^{1/2} (\|\mathbf{H}_f(\mathbf{X}_p)\|_{4'} + \|\mathbf{H}_q(\mathbf{X}_p, \mathbf{U})\|_{4'}) \\ &\times \left( \sum_{i=1}^n \|1 + \|\nabla f_i(\mathbf{X}_p)\|_2\|^{1/2} \right. \\ &\quad \left. + \sum_{i=1}^m \|1 + \|\mathbf{J}_f^T(\mathbf{X}_p)(\nabla h_i)[\mathbf{f}(\mathbf{X}_p) + \mathbf{U}]\|_2\|^{1/2} \right), \end{aligned}$$

where the matrix  $\mathbf{J}_f^T(\mathbf{X}_p)$  is defined in Equation (14).

*Proof.* Because the mean and covariance of  $\text{pr } \mathbf{Z}$  are equal to those of  $\mathbf{Z}$ , we can apply Theorem 5 to  $\mathbf{W}$  and  $\mathbf{g}$ . This yields

$$\begin{aligned} d_W(\mathbf{Z}, \mathbf{Z}_G) &\leq \frac{3\sqrt{2}}{2} \|\mathbf{P}_Z^{-1}\| \|\mathbf{P}_Z\|^{1/2} \sum_{i=1}^{n+m} (\mathbb{E}[\|\mathbf{H}_{g_i}(\mathbf{W})\|^4])^{1/4} \\ &\times \sum_{i=1}^{n+m} (\mathbb{E}[\|\nabla g_i(\mathbf{W})\|^4])^{1/4}. \end{aligned}$$

The gradients and Hessians can be expanded. For  $i \leq n$ ,  $\mathbf{H}_{g_i}$  is zero except for  $\mathbf{H}_{f_i}$  on its upper left corner so that  $\|\mathbf{H}_{g_i}(\mathbf{W})\| = \|\mathbf{H}_{f_i}(\mathbf{X}_p)\|$ . For  $i > n$ , the Hessian has on its upper left corner the  $2n \times 2n$  Hessian  $\mathbf{H}_{q_i}$  of  $\mathbf{q}$  in (9) and hence  $\|\mathbf{H}_{g_i}(\mathbf{W})\| = \|\mathbf{H}_{q_i}(\mathbf{X}_p, \mathbf{U})\|$ . For the first  $n$  gradients,

$$\|\nabla g_i(\mathbf{W})\|^2 = 1 + \|\nabla f_i(\mathbf{X}_p)\|^2$$

and for the rest

$$\|\nabla g_i(\mathbf{W})\|^2 = 1 + \|\nabla q_i(\mathbf{X}_p, \mathbf{U})\|^2.$$

By the chain rule, the second gradient is

$$\begin{aligned}\nabla q_i(\mathbf{X}_p, \mathbf{U}) &= \begin{pmatrix} \mathbf{J}_f(\mathbf{X}_p)^\top (\nabla h_i)[\mathbf{f}(\mathbf{X}_p) + \mathbf{U}] \\ (\nabla h_i)[\mathbf{f}(\mathbf{X}_p) + \mathbf{U}] \end{pmatrix} \\ &= \mathbf{J}_f^I(\mathbf{X}_p) (\nabla h_i)[\mathbf{f}(\mathbf{X}_p) + \mathbf{U}],\end{aligned}\quad (13)$$

where

$$\mathbf{J}_f^I(\mathbf{X}_p) = \begin{pmatrix} \mathbf{J}_f(\mathbf{X}_p)^\top \\ \mathbf{I}_{n \times n} \end{pmatrix} \in \mathbb{R}^{2n \times n}.\quad (14)$$

Recall the  $L^p$ -norm notation  $\|\mathbf{X}\|_p = (\mathbb{E}[\|\mathbf{X}\|^p])^{1/p}$ . Therefore  $(\mathbb{E}[X^4])^{1/4} = \|X^2\|_2^{1/2}$  for any univariate random variable  $X$ . With this in mind, the bounds and identities above yield

$$\begin{aligned}d_W(\mathbf{Z}, \text{pr } \mathbf{Z}) &\leq \frac{3}{\sqrt{2}} \|\mathbf{P}_Z^{-1}\| \|\mathbf{P}_Z\|^{1/2} (\|\mathbf{H}_f(\mathbf{X}_p)\|_{4'} + \|\mathbf{H}_q(\mathbf{X}_p, \mathbf{U})\|_{4'}) \\ &\quad \times \left( \sum_{i=1}^n \|1 + \|\nabla f_i(\mathbf{X}_p)\|_2\|_2^{1/2} \right. \\ &\quad \left. + \sum_{i=1}^m \|1 + \|\nabla q_i(\mathbf{X}_p, \mathbf{U})\|_2\|_2^{1/2} \right),\end{aligned}$$

so that inserting (13) yields the claimed upper bound.  $\square$

We can now combine these lemmas with (10) and (6) to obtain the main result of this article.

**Theorem 8.** *If  $g$  in Equation (8) is in  $\mathcal{L}(\mathbf{W})$ , then*

$$d_W(\mathbf{Z}, \tilde{\mathbf{Z}}) \leq C_1 + C_2,$$

where the two constants  $C_1$  and  $C_2$ , which give the upper bounds  $d_W(\mathbf{Z}, \text{pr } \mathbf{Z}) \leq C_1$  and  $d_W(\text{pr } \mathbf{Z}, \tilde{\mathbf{Z}}) \leq C_2$ , are

$$\begin{aligned}C_1 &= \frac{3}{\sqrt{2}} \|(\mathbf{A}_1 + \mathbf{B}_1 + \mathbf{C}_V)^{-1}\| \|\mathbf{A}_2 + \mathbf{C}_V\|^{1/2} \\ &\quad \times (\|\mathbf{H}_f(\mathbf{X}_p)\|_{4'} + \|\mathbf{H}_q(\mathbf{X}_p, \mathbf{U})\|_{4'}) \\ &\quad \times \left( \sum_{i=1}^n \|1 + \|\nabla f_i(\mathbf{X}_p)\|_2\|_2^{1/2} \right. \\ &\quad \left. + \sum_{i=1}^m \|1 + \|\mathbf{J}_f^I(\mathbf{X}_p) (\nabla h_i)[\mathbf{f}(\mathbf{X}_p) + \mathbf{U}]\|_2\|_2^{1/2} \right), \\ C_2 &= \left[ \|\mathbb{E}(\mathbf{Y}) - \mathbf{m}_{\tilde{\mathbf{Y}}}\|^2 + \text{tr } \mathbf{A}_2 + \text{tr } \boldsymbol{\Sigma}_V + \text{tr } \mathbf{P}_X + \text{tr } \mathbf{P}_{\tilde{\mathbf{Y}}} \right. \\ &\quad \left. - 2 \text{tr} (\sqrt{\mathbf{L}(\mathbf{A}_1 + \mathbf{B}_1 + \mathbf{C}_V)\mathbf{L}}) \right]^{1/2}.\end{aligned}$$

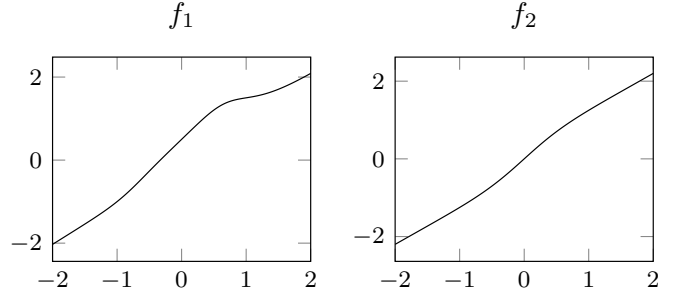
Here  $\mathbf{A}_1, \mathbf{B}_1, \mathbf{C}_V$  and  $\mathbf{A}_2$  are defined in Lemma 6,  $\mathbf{J}_f^I(\mathbf{X}_p)$  is defined in Equation (14), and  $\mathbf{L} = \text{Cov}(\tilde{\mathbf{Z}})^{1/2}$ .

*Proof.* The constant  $C_1$  is obtained from Lemmas 6 and 7. The constant  $C_2$  is obtained by estimating  $d_W(\text{pr } \mathbf{Z}, \tilde{\mathbf{Z}})$  through (6) and Lemma 6 and noting that  $\text{tr } \text{Cov}(\mathbf{Z}) \leq \text{tr}(\mathbf{A}_2 + \mathbf{C}_V)$ ,

$$\mathbb{E}(\mathbf{Z}) - \mathbb{E}(\tilde{\mathbf{Z}}) = \begin{pmatrix} \mathbf{0}_{n \times 1} \\ \mathbb{E}(\mathbf{Y}) - \mathbf{m}_{\tilde{\mathbf{Y}}} \end{pmatrix},$$

and

$$\text{tr } \text{Cov}(\tilde{\mathbf{Z}}) = \text{tr} \begin{pmatrix} \mathbf{P}_X & \mathbf{C}_{\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}} \\ \mathbf{C}_{\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}}^\top & \mathbf{P}_{\tilde{\mathbf{Y}}} \end{pmatrix} = \text{tr}(\mathbf{P}_X + \mathbf{P}_{\tilde{\mathbf{Y}}}). \quad \square$$



**Fig. 2:** The non-linear functions  $f_1$  and  $f_2$  in (16).

Let us examine the upper bound of Theorem 8 when  $\mathbf{f}$  and  $\mathbf{h}$  are affine. In this case the function  $\mathbf{q}$  is also affine, so that  $C_1 = 0$  because the Hessians are zero. We are left with

$$d_W(\mathbf{Z}, \tilde{\mathbf{Z}}) \leq C_2.$$

Note that  $C_1$  is positive if only either  $\mathbf{f}$  or  $\mathbf{h}$  is affine. Since the model and measurement functions are affine,  $\mathbf{X} = \tilde{\mathbf{X}}$  and  $\mathbf{Y} = \tilde{\mathbf{Y}}$  are Gaussian random variables. The bounds in Lemma 6 coincide because Jacobians are constant matrices. Hence  $\text{Cov}(\mathbf{Z}) = \text{Cov}(\tilde{\mathbf{Z}}) = \mathbf{A}_1 + \mathbf{B}_1 + \mathbf{C}_V = \mathbf{A}_2 + \mathbf{C}_V$  and  $\text{tr } \text{Cov}(\mathbf{Z}) = \text{tr } \mathbf{P}_X + \text{tr } \mathbf{P}_{\tilde{\mathbf{Y}}}$ . Also recall from (12) that  $\text{tr } \mathbf{C}_V = \text{tr } \boldsymbol{\Sigma}_V$ . These equations yield (note the square)

$$\begin{aligned}d_W(\mathbf{Z}, \tilde{\mathbf{Z}})^2 &\leq \|\mathbb{E}(\mathbf{Y}) - \mathbb{E}(\tilde{\mathbf{Y}})\|^2 + \text{tr } \mathbf{A}_2 + \text{tr } \boldsymbol{\Sigma}_V + \text{tr } \mathbf{P}_X + \text{tr } \mathbf{P}_{\tilde{\mathbf{Y}}} \\ &\quad - 2 \text{tr} (\sqrt{\mathbf{L}(\mathbf{A}_1 + \mathbf{B}_1 + \mathbf{C}_V)\mathbf{L}}) \\ &= \|\mathbb{E}(\mathbf{Y}) - \mathbb{E}(\mathbf{Y})\|^2 + \text{tr } \text{Cov}(\mathbf{Z}) + \text{tr } \text{Cov}(\mathbf{Z}) \\ &\quad - 2 \text{tr} (\sqrt{\text{Cov}(\mathbf{Z})^{1/2} \text{Cov}(\mathbf{Z}) \text{Cov}(\mathbf{Z})^{1/2}}),\end{aligned}$$

which is zero since  $\sqrt{\mathbf{A}^{1/2} \mathbf{A} \mathbf{A}^{1/2}} = \mathbf{A}$ .

## VI. A NUMERICAL EXAMPLE

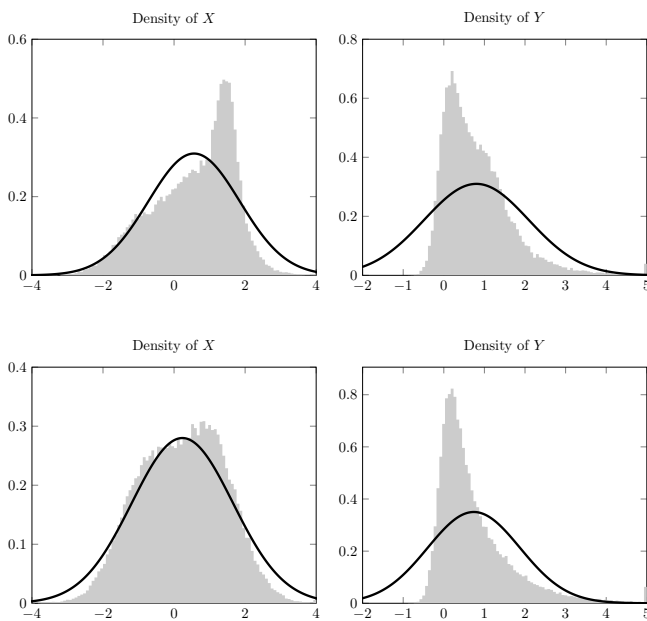
We consider two variants of the univariate non-stationary growth model (e.g., [35]) commonly used to assess and compare non-linear Kalman filters. The models we consider are

$$X = f_i(X_p) + U, \quad Y = \frac{1}{2}X^2 + V \quad (15)$$

for the dynamic model functions (plotted in Figure 2)

$$f_1(x) = x + \frac{1+x}{2(1+x^4)} \quad \text{and} \quad f_2(x) = x + \frac{x}{2(1+x^2)}. \quad (16)$$

The function  $f_1$  is, at least visually, more non-linear than  $f_2$ . We set  $U, V \sim \mathcal{N}(0, 0.05)$  and  $X_p \sim \mathcal{N}(0.2, 1)$ . Figure 3 shows the distributions of  $X, Y, \tilde{X}$ , and  $\tilde{Y}$ . With the expectations computed using Monte Carlo with  $10^6$  samples, Theorem 8 gives the bounds  $d_W(\mathbf{Z}, \tilde{\mathbf{Z}}) \leq 2267$  and  $d_W(\text{pr } \mathbf{Z}, \tilde{\mathbf{Z}}) \leq 663$  for  $f_1$  and  $f_2$ , respectively. As expected, the bound is larger for the more non-linear function (i.e.,  $f_1$ ). We note that the term  $C_1$  of Theorem 8 dominates as  $C_2 \approx \sqrt{2}$  for both functions. The bounds are very conservative: the true distances are  $d_W(X, \tilde{X}) \approx 0.21$  and  $d_W(Y, \tilde{Y}) \approx 0.45$  for  $f_1$ , and  $d_W(X, \tilde{X}) \approx 0.09$  and  $d_W(Y, \tilde{Y}) \approx 0.48$  for  $f_2$ . We have not found a convenient way to compute  $d_W(\mathbf{Z}, \tilde{\mathbf{Z}})$  from samples.



**Fig. 3:** Distributions of  $X$  and  $Y$  for the model in (15) with  $f_1$  (above) and  $f_2$  (below). Densities of the Gaussian approximations  $\tilde{X}$  and  $\tilde{Y}$  are in black.

## VII. CONCLUSION

Theorem 8 gives an upper bound on the Wasserstein distance between the true joint distribution  $\mathbf{Z}$  of the prediction  $\mathbf{X}$  and measurement  $\mathbf{Y}$  and its Gaussian approximation  $\tilde{\mathbf{Z}} = (\tilde{\mathbf{X}}, \tilde{\mathbf{Y}})$ . The bound vanishes when the model and measurement functions are affine. Understanding how the conditional filtering distribution  $\mathbf{X} | \mathbf{Y} = \mathbf{y}$  relates to its Gaussian approximation  $\tilde{\mathbf{X}} | \tilde{\mathbf{Y}} = \mathbf{y}$  is much more challenging. The key difficulty is in quantifying how perturbing the prior affects the posterior [36, Lem. 3.6]. Unfortunately, such stability bounds are available only under very restrictive assumptions (e.g., [23] and [37]). This is a pertinent problem throughout Bayesian statistics.

## REFERENCES

- [1] M. Li, B. H. Kim, and A. I. Mourikis, "Real-time motion tracking on a cellphone using inertial sensing and a rolling-shutter camera," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2013.
- [2] E. Foxlin, "Pedestrian tracking with shoe-mounted inertial sensors," *IEEE Computer Graphics and Applications*, vol. 25, no. 6, pp. 38–46, 2005.
- [3] S. Särkkä, A. Solin, A. Nummenmaa, A. Vehtari, T. Auranen, and F.-H. Lin, "Dynamic retrospective filtering of physiological noise in BOLD fMRI: DRIFTER," *NeuroImage*, vol. 60, no. 2, pp. 1517–1527, 2012.
- [4] M. W. M. G. Dissanayake, P. Newman, S. Clark, H. F. Durrant-Whyte, and M. Csorba, "A solution to the simultaneous localization and map building (SLAM) problem," *IEEE Transactions on Robotics and Automation*, vol. 17, no. 3, pp. 229–241, 2002.
- [5] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*, ser. Information and System Sciences Series. Prentice-Hall, 1979.
- [6] A. H. Jazwinski, *Stochastic Processes and Filtering Theory*. Academic Press, 1970.
- [7] B. D. O. Anderson and J. B. Moore, "Detectability and stabilizability of time-varying discrete-time linear systems," *SIAM Journal on Control and Optimization*, vol. 19, no. 1, pp. 20–32, 1981.
- [8] E. W. Kamen and J. K. Su, *Introduction to Optimal Estimation*, ser. Advanced Textbooks in Control and Signal Processing. Springer, 1999.
- [9] K. Reif, S. Günther, E. Yaz, and R. Unbehauen, "Stochastic stability of the discrete-time extended Kalman filter," *IEEE Transactions on Automatic Control*, vol. 44, no. 4, pp. 714–728, 1999.
- [10] K. Xiong, H. Zhang, and C. W. Chan, "Performance evaluation of UKF-based nonlinear filtering," *Automatica*, vol. 42, no. 2, pp. 261–270, 2006.
- [11] W. He, X. Yang, and X. Qiu, "Convergence analysis and improvement of progressive Gaussian approximation filters," *Signal Processing*, vol. 200, no. 108643, 2022.
- [12] T. Karvonen, S. Bonnabel, E. Moulines, and S. Särkkä, "On stability of a class of filters for nonlinear stochastic systems," *SIAM Journal on Control and Optimization*, vol. 58, no. 4, pp. 2023–2049, 2020.
- [13] Y. Wu, D. Hu, M. Wu, and X. Hu, "A numerical-integration perspective on Gaussian filters," *IEEE Transactions on Signal Processing*, vol. 54, no. 8, pp. 2910–2921, 2006.
- [14] S. Särkkä and L. Svensson, *Bayesian Filtering and Smoothing*, 2nd ed. Cambridge University Press, 2023.
- [15] M. R. Morelande and Á. F. García-Fernández, "Analysis of Kalman filter approximations for nonlinear measurements," *IEEE Transactions on Signal Processing*, vol. 61, no. 22, pp. 5477–5484, 2013.
- [16] D. S. Bernstein, *Matrix Mathematics: Theory, Facts, and Formulas*, 2nd ed. Princeton University Press, 2009.
- [17] J. McNamee and F. Stenger, "Construction of fully symmetric numerical integration formulas," *Numerische Mathematik*, vol. 10, no. 4, pp. 327–344, 1967.
- [18] S. J. Julier, J. K. Uhlmann, and H. F. Durrant-Whyte, "A new approach for filtering nonlinear systems," in *Proceedings of the 1995 American Control Conference*, 1995, pp. 1628–1632.
- [19] A. Farina, B. Ristic, and D. Benvenuti, "Tracking a ballistic target: comparison of several nonlinear filters," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 38, no. 3, pp. 854–867, 2002.
- [20] A. L. Gibbs and F. E. Su, "On choosing and bounding probability metrics," *International Statistical Review*, vol. 70, no. 3, pp. 419–435, 2002.
- [21] C. Villani, *Optimal Transport: Old and New*, ser. Grundlehren der mathematischen Wissenschaften. Springer, 2008, no. 338.
- [22] C. R. Givens and R. M. Shortt, "A class of Wasserstein metrics for probability distributions," *The Michigan Mathematical Journal*, vol. 31, no. 2, pp. 231–240, 1984.
- [23] C. Ley, G. Reinert, and Y. Swan, "Distances between nested densities and a measure of the impact of the prior in Bayesian statistics," *The Annals of Applied Probability*, vol. 27, no. 1, pp. 216–241, 2017.
- [24] I. Nourdin and G. Peccati, *Normal Approximations with Malliavin Calculus: From Stein's Method to Universality*, ser. Cambridge Tracts in Mathematics. Cambridge University Press, 2012, no. 192.
- [25] C. Stein, "A bound for the error in the normal approximation to the distribution of a sum of dependent random variables," in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 2, 1972, pp. 582–602.
- [26] L. H. Y. Chen, L. Goldstein, and Q. Shao, *Normal Approximation by Stein's Method*, ser. Probability and Its Applications. Springer, 2011.
- [27] C. Ley, G. Reinert, and Y. Swan, "Stein's method for comparison of univariate distributions," *Probability Surveys*, vol. 14, pp. 1–52, 2017.
- [28] T. Cacoullos, "On upper and lower bounds for the variance of a function of a random variable," *The Annals of Probability*, vol. 10, no. 3, pp. 799–809, 1982.
- [29] J. Nash, "Continuity of solutions of parabolic and elliptic equations," *American Journal of Mathematics*, vol. 80, no. 4, pp. 931–954, 1958.
- [30] H. Chernoff, "A note on an inequality involving the normal distribution," *The Annals of Probability*, vol. 9, no. 3, pp. 533–535, 1981.
- [31] S. Chatterjee, "Fluctuations of eigenvalues and second order Poincaré inequalities," *Probability Theory and Related Fields*, vol. 143, no. 1, pp. 1–40, 2009.
- [32] I. Nourdin, G. Peccati, and G. Reinert, "Second order Poincaré inequalities and CLTs on Wiener space," *Journal of Functional Analysis*, vol. 257, no. 2, pp. 593–509, 2009.
- [33] I. Nourdin, G. Peccati, and A. Réveillac, "Multivariate normal approximation using Stein's method and Malliavin calculus," *Annales de l'institut Henri Poincaré - Probabilités et Statistique*, vol. 46, no. 1, pp. 45–58, 2010.
- [34] Y. Wu, D. Hu, M. Wu, and X. Hu, "Unscented Kalman filtering for additive noise case: augmented vs. non-augmented," in *Proceedings of the 2005 American Control Conference*, 2005, pp. 4051–4055.
- [35] N. J. Gordon, D. J. Salmond, and A. F. M. Smith, "Novel approach to nonlinear/non-Gaussian Bayesian state estimation," *IEEE Proceedings F (Radar and Signal Processing)*, vol. 140, no. 2, pp. 107–113, 1993.
- [36] H. R. Künsch, "State space and hidden Markov models," in *Complex Stochastic Systems*, O. E. Barndorff-Nielsen, D. R. Cox, and C. Klüppelberg, Eds. Chapman & Hall, 2001.
- [37] A. Garbuno-Inigo, T. Helin, F. Hoffman, and B. Hosseini, "Bayesian posterior perturbation analysis with integral probability metrics," *arXiv:2303.01512*, 2023, preprint.