EasyChair Preprint

№ 10348

Stein Variational Gradient Descent for Non-Bayesian Particle Flow

Kyle Craft and Kyle DeMars

June 7, 2023

Stein Variational Gradient Descent for Non-Bayesian Particle Flow

Kyle J. Craft
Department of Aerospace Engineering
Texas A&M University
College Station, TX, USA
kcraft@tamu.edu

Kyle J. DeMars
Department of Aerospace Engineering
Texas A&M University
College Station, TX, USA
demars@tamu.edu

***Abstract*—Bayes' rule provides an undoubtedly powerful framework for statistical inference; however, the assumptions inherent in Bayesian filtering often cannot be realized in physical systems. Oftentimes, the true Bayesian posterior probability density function (pdf) is infinite-dimensional and lacks tractable implementations, in addition to errors induced by inaccurate realizations of the prior and likelihood pdfs. Though particle-based methods can provide versatile and computationally efficient approximations of Bayes' rule, they lack the theoretical ability to mitigate estimation errors incurred by erroneous measurement modeling. This work merges Stein Variational Gradient Descent, a nonlinear particle flow update scheme, with generalized variational inference, a method for formulating optimal non-Bayesian posteriors, to produce tractable variational posterior pdfs that remain robust to modeling errors. The new framework is demonstrated to outperform conventional filtering approaches in a simplified relative spacecraft navigation scenario.**

***Index Terms*—statistical inference, nonlinear estimation, information theory, generalized variational inference, particle flow**

I. Introduction

The most fundamental problem in data fusion is the optimal combination of existing information for an uncertain state variable, $\boldsymbol{x}$, with an external and noisy observation, $\boldsymbol{z}$. When prior state knowledge is given by the underlying probability density function (pdf), the most statistically complete incorporation of an observation is dictated by Bayes' rule,

$$p(\boldsymbol{x}|\boldsymbol{z}) \propto p(\boldsymbol{z}|\boldsymbol{x})p(\boldsymbol{x}) \ , \tag{1}$$

where $p(\boldsymbol{x})$ is the prior state density, $p(\boldsymbol{z}|\boldsymbol{x})$ is the measurement pdf conditioned on the state, referred to as the likelihood function, and $p(\boldsymbol{x}|\boldsymbol{z})$ is the posterior state pdf. Proportionality in Eq. (1) is resolved by ensuring the posterior is a proper pdf, i.e., $p(\boldsymbol{x}|\boldsymbol{z})$ integrates to unity across its support $\boldsymbol{x} \in \mathbb{X}$, yielding the Bayes' normalization constant,

$$c = p(\boldsymbol{z}) = \int_{\mathbb{X}} p(\boldsymbol{z}|\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \ .$$

Unfortunately, as observed in [1], Bayes' rule is predicated on three fundamental assumptions: *i*) the prior is correctly specified, *ii*) the likelihood function associated with the measurement model is statistically accurate; and *iii*) the posterior can be tractably calculated. Though the Bayesian framework is an undoubtedly powerful inference tool, practical applications of Eq. (1) frequently demand violation of one or more of these assumptions. In many instances, the prior (*i*) and likelihood (*ii*) are intentionally misspecified to recover a computationally cheap analytical solution (*iii*). Perhaps the most prominent example of this violation are "linear-Gaussian" filters, wherein the measurement is approximated as a linear combination of the states and noise while both the prior and likelihood pdfs are represented as Gaussians [2] or Gaussian mixtures [3]. Linear-Gaussian schemes are popular for their algorithmic equivalence to the extended Kalman filter; however, errors incurred from the linearization and/or Gaussian assumptions can accumulate, leading to filter inconsistency or divergence. For most practical systems, when (*i*) and (*ii*) are retained, Eq. (1) does not beget an analytical solution. Even in instances where the Bayesian posterior can be derived analytically, the result is not guaranteed to be a conjugate (belonging to the same distributional family) of the prior or to possess a finite parameterization, preventing closed-form recursions. Furthermore, numerical implementations of Eq. (1) are often computationally prohibitive, as integration of the normalization constant grows exponentially expensive with the state dimension.

An alternative to parametric estimation is particle filtering [4], which leverages Monte Carlo methods to approximate solutions for Bayes' rule. However, traditional particle filtering is well-documented to suffer from both the "curse of dimensionality," wherein the number of particles required to accurately represent the state pdf grows exponentially with the dimension of the system, and particle degeneracy, where the majority of particle weights degenerate to (numerically) zero rapidly [5]. The primary culprit behind particle degeneracy is the stagnant nature of particle filtering updates. As the filter progresses through time, the prior distribution of particles may be insufficient to properly encapsulate posterior probability mass, particularly when uncertainty decreases rapidly. This necessitates sequential resampling from a new importance distribution [4], which can be difficult to implement recursively. To circumvent instances of particle degeneracy, a "particle flow" update can be utilized. Rather than updating the particle weights, particle flow, alternatively, transports, or "flows," the

This work was supported by a NASA Space Technology Graduate Research Opportunities Award.

particles through the state space from prior to posterior, such as in [6], [7].

Though traditional particle flow approaches may result in tractable recursions, Bayesian methods are inherently subject to errors induced by inaccurate modeling. For practical systems, where model fidelity is inevitably scarified in favor of computational expedience, model-based errors are mitigated by *ad hoc* methods, such as measurement underweighting [8] or residual editing [9]. Rather than limit the information available to the filter by altering Bayes' rule, it may be beneficial to define an alternative update framework, such as generalized variational inference (GVI) [1]. GVI is an information-theoretic approach that recasts statistical inference as an optimization problem over a loss function, representing information ingested by the filter from the observation, and a statistical divergence, which weights the filter's faithfulness to the prior realization of the state pdf. By making informed decisions on the loss and divergence functions over which the optimization occurs, it is possible to design optimal filters that are inherently robust to errors in system modeling.

This paper is a continuation of previous work by the authors in [10] that extends Stein variational gradient descent (SVGD) [11], a form of nonlinear particle flow, to determine optimal non-Bayesian posteriors formulated using GVI. The paper is organized as follows. Section II defines SVGD and demonstrates its application to both Bayesian and non-Bayesian inference. Section III provides an example of a GVI applied to the popular Gaussian filtering scheme. Section IV provides results for a test scenario involving relative space object tracking, and Section V provides concluding remarks and outlines future research avenues.

II. SVGD for Statistical Inference

The objective of any particle flow update scheme is to transport a collection of particles from some initial ensemble, whether that be samples from the prior pdf or an arbitrary reference, to some target measure. This approach can be viewed as a pseudo-Monte-Carlo method, where the posterior ensemble should imitate direct sampling of the target density. By leveraging Stein's identity to minimize a discrepancy measure, SVGD [11] provides a means for sampling target densities that are known up to a proportionality constant. Additionally, it does not restrict the posterior to specific distributional families, such as Gaussians in [6], [7].

A. SVGD Particle Flow

Let $q(\boldsymbol{x})$ be the variational density approximating some intractable, potentially infinite-dimensional, pdf $\rho(\boldsymbol{x})$. A natural selection for the discrepancy measure between q and ρ is the Kullback-Leibler (KL) divergence, defined as [12]

$$D_{KL}\left[q||\rho\right] = \int_{\mathbb{X}} q(\boldsymbol{x}) \ln\left(\frac{q(\boldsymbol{x})}{\rho(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x} \ . \tag{2}$$

From an information-theoretic perspective, the KL divergence as written in Eq. (2) can be viewed as a measure of information loss in approximating q as ρ, where "measure" indicates that D_{KL} is, in general, non-symmetric and not a proper metric. Additionally, the KL divergence possesses two favorable properties: it is convex in q and non-negative, with equality if and only if the two pdfs are identical (almost everywhere) [12]. Non-negativity for the KL divergence, known as the self-identifying property, allows the sampling problem to be formulated as

$$q^*(\boldsymbol{x}) = \underset{q(\boldsymbol{x})}{\operatorname{argmin}} \ D_{KL}\left[q(\boldsymbol{x})\,||\,\rho(\boldsymbol{x})\right] \ ,$$

where the global minimum, q^*, exactly recovers ρ. In many instances, such as Bayes' rule, it is only possible to define ρ tractably up to a normalizing constant. Let $\tilde{\rho}$ be proportional to the true target density, $\rho(\boldsymbol{x}) \propto \tilde{\rho}(\boldsymbol{x})$, and $c_{\tilde{\rho}}$ be the (possibly unknown) normalization constant, allowing Eq. (2) to be written as

$$D_{KL}\left[q||\rho\right] = \int_{\mathbb{X}} q(\boldsymbol{x}) \ln\left(\frac{q(\boldsymbol{x})}{(1/c_{\tilde{\rho}})\tilde{\rho}(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x} \ .$$

From the properties of logarithms and the linearity of integration, the previous equation can be separated to

$$\begin{aligned} D_{KL}\left[q||\rho\right] &= \int_{\mathbb{X}} q(\boldsymbol{x}) \ln\left(\frac{q(\boldsymbol{x})}{\tilde{\rho}(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x} + \int_{\mathbb{X}} q(\boldsymbol{x}) \ln c_{\tilde{\rho}} \mathrm{d}\boldsymbol{x} \\ &= \int_{\mathbb{X}} q(\boldsymbol{x}) \ln\left(\frac{q(\boldsymbol{x})}{\tilde{\rho}(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x} + \ln c_{\tilde{\rho}} \ , \end{aligned}$$

where the second simplification results from $c_{\tilde{\rho}}$ being deterministic. Because the minimization of Eq. (2) is not influenced by the addition of a constant, a new cost functional, $J[q||p]$, sharing the same minimum can be formulated as

$$J[q||\rho] = \int_{\mathbb{X}} q(\boldsymbol{x}) \ln\left(\frac{q(\boldsymbol{x})}{\tilde{\rho}(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x} \ , \tag{3}$$

where q^* is now derived from

$$q^*(\boldsymbol{x}) = \underset{q(\boldsymbol{x})}{\operatorname{argmin}} \ J[q(\boldsymbol{x})\,||\,\rho(\boldsymbol{x})] \ ,$$

without necessitating knowledge of the proportionality constant.

Analytical minimization of Eq. (3) is typically reserved to a limited number of cases that often align with closure of Bayes' rule. When no analytical minimum exists, the convexity of the KL divergence naturally lends itself to gradient-based methods. Similar to conventional gradient descent, the goal is to iteratively transport the variational density from an initial pdf, q_0, to the target, ρ. This is accomplished by defining a new pseudo-parameter, τ, referred to as pseudo-time (or the homotopy parameter as in [6], [7]) due to its strong connection to traditional dynamics. Let the distribution of states evolve according to the deterministic differential equation,

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}\tau} = \boldsymbol{\phi}(\boldsymbol{x}) \ , \tag{4}$$

where $\boldsymbol{\phi}(\boldsymbol{x})$ is the state pseudo-time rate of change, referred to as the pseudo-dynamics function. Equation (4) then induces a

pseudo-time rate of change in $q(\boldsymbol{x})$ governed by the Liouville equation [13],

$$\frac{\partial q(\boldsymbol{x};\tau)}{\partial \tau} = -\nabla_{\boldsymbol{x}} \cdot \left(q(\boldsymbol{x};\tau)\boldsymbol{\phi}(\boldsymbol{x})\right) \ , \tag{5}$$

where the variational density is now parameterized by the pseudo-time, $q(\boldsymbol{x};\tau=0)$, and $\nabla_{\boldsymbol{x}}\cdot$ is the divergence operator. Similar to traditional gradient descent, the resulting flow map for the variational density, i.e., the solution to Eq. (5), transports the initial pdf to the target density, according to

$$\lim_{\tau\to\infty} q(\boldsymbol{x};\tau) = q^*(\boldsymbol{x}) = \rho(\boldsymbol{x}) \ , \tag{6}$$

where a superscript asterisk, $(\cdot)^*$, indicates an optimal quantity. This limit is achieved by defining the optimal pseudo-dynamics according to

$$\boldsymbol{\phi}^*(\boldsymbol{x}) = \underset{\boldsymbol{\phi}(\boldsymbol{x})}{\operatorname{argmin}} \left\{ \frac{\mathrm{d}J[q(\boldsymbol{x};\tau)\,||\,\rho(\boldsymbol{x})]}{\mathrm{d}\tau} \right\} \ , \tag{7}$$

such that, heuristically, $\boldsymbol{\phi}^*(\boldsymbol{x})$ evolves the variational density through pseudo-time in the "steepest descent direction" of the KL divergence. Taking the pseudo-time derivative of Eq. (3) yields

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\tau}\Big(J[q\,||\,\rho]\Big) &= \frac{\mathrm{d}}{\mathrm{d}\tau}\int_{\mathbb{X}} q(\boldsymbol{x};\tau)\ln\left(\frac{q(\boldsymbol{x};\tau)}{\tilde{\rho}(\boldsymbol{x})}\right)\mathrm{d}\boldsymbol{x} \\ &= \int_{\mathbb{X}} \frac{\partial}{\partial\tau}\left[q(\boldsymbol{x};\tau)\ln\left(\frac{q(\boldsymbol{x};\tau)}{\tilde{\rho}(\boldsymbol{x})}\right)\right]\mathrm{d}\boldsymbol{x} \\ &= \int_{\mathbb{X}} \left(\frac{\partial q(\boldsymbol{x};\tau)}{\partial\tau}\right)\left[\ln\left(\frac{q(\boldsymbol{x};\tau)}{\tilde{\rho}(\boldsymbol{x})}\right)+1\right]\mathrm{d}\boldsymbol{x} \ ,\end{aligned}$$

where the derivative is moved inside the integral using Leibniz' rule and the integrand is simplified from the chain rule. Substituting Eq. (5) into the previous expression results in

$$\frac{\mathrm{d}J}{\mathrm{d}\tau} = -\int_{\mathbb{X}} \nabla_{\boldsymbol{x}}\cdot\left(q(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{x})\right)\left[\ln\left(\frac{q(\boldsymbol{x};\tau)}{\tilde{\rho}(\boldsymbol{x})}\right)+1\right]\mathrm{d}\boldsymbol{x} \ ,$$

facilitating an application of the divergence theorem [14], provided q is limited to zero at the boundaries of its support $\mathbb{X}$, to commute the derivative, such that

$$\begin{aligned}\frac{\mathrm{d}J}{\mathrm{d}\tau} &= \int_{\mathbb{X}} q(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{x})\cdot\nabla_{\boldsymbol{x}}\left[\ln\left(\frac{q(\boldsymbol{x};\tau)}{\tilde{\rho}(\boldsymbol{x})}\right)+1\right]\mathrm{d}\boldsymbol{x} \\ &= \int_{\mathbb{X}} q(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{x})\cdot\nabla_{\boldsymbol{x}}\ln\left(\frac{q(\boldsymbol{x};\tau)}{\tilde{\rho}(\boldsymbol{x})}\right)\mathrm{d}\boldsymbol{x} \ ,\end{aligned}$$

where the convention $\nabla_{\boldsymbol{x}} = \begin{bmatrix}\frac{\partial}{\partial x_1} & \frac{\partial}{\partial x_2} & \cdots & \frac{\partial}{\partial x_n}\end{bmatrix}^T$ is used, $\boldsymbol{x}$ is of dimension n, and $\cdot$ is the Cartesian inner product. Applying the properties of logarithms and the chain rule, it follows that

$$\begin{aligned}\frac{\mathrm{d}J}{\mathrm{d}\tau} &= \int_{\mathbb{X}} q(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{x})\cdot\nabla_{\boldsymbol{x}}\left[\ln q(\boldsymbol{x};\tau)-\ln\tilde{\rho}(\boldsymbol{x})\right]\mathrm{d}\boldsymbol{x} \\ &= \int_{\mathbb{X}} q(\boldsymbol{x};\tau)\boldsymbol{\phi}(\boldsymbol{x})\cdot\left[\nabla_{\boldsymbol{x}}\ln q(\boldsymbol{x};\tau)-\nabla_{\boldsymbol{x}}\ln\tilde{\rho}(\boldsymbol{x})\right]\mathrm{d}\boldsymbol{x} \\ &= \int_{\mathbb{X}} q(\boldsymbol{x};\tau)\boldsymbol{\phi}(\boldsymbol{x})\cdot\left[\frac{\nabla_{\boldsymbol{x}}q(\boldsymbol{x};\tau)}{q(\boldsymbol{x};\tau)}-\nabla_{\boldsymbol{x}}\ln\tilde{\rho}(\boldsymbol{x})\right]\mathrm{d}\boldsymbol{x} \\ &= \int_{\mathbb{X}} \Big[\boldsymbol{\phi}(\boldsymbol{x})\cdot\nabla_{\boldsymbol{x}}q(\boldsymbol{x};\tau)-q(\boldsymbol{x};\tau)\boldsymbol{\phi}(\boldsymbol{x})\cdot\nabla_{\boldsymbol{x}}\ln\tilde{\rho}(\boldsymbol{x})\Big]\mathrm{d}\boldsymbol{x}.\end{aligned}$$

A final application of the divergence theorem [14] allows

$$\begin{aligned}\frac{\mathrm{d}J}{\mathrm{d}\tau} &= -\int_{\mathbb{X}} q(\boldsymbol{x};\tau)\left[\nabla_{\boldsymbol{x}}\cdot\boldsymbol{\phi}(\boldsymbol{x})+\boldsymbol{\phi}(\boldsymbol{x})\cdot\nabla_{\boldsymbol{x}}\ln\tilde{\rho}(\boldsymbol{x})\right]\mathrm{d}\boldsymbol{x} \\ &= -\mathbb{E}_q\left\{\nabla_{\boldsymbol{x}}\cdot\boldsymbol{\phi}(\boldsymbol{x})+\boldsymbol{\phi}(\boldsymbol{x})\cdot\nabla_{\boldsymbol{x}}\ln\tilde{\rho}(\boldsymbol{x})\right\} \ ,\end{aligned} \tag{8}$$

where $\mathbb{E}_q\{\cdot\}$ is the statistical expectation operator with respect to the pdf q, defined for a generic function, $f(\boldsymbol{x})$, as

$$\mathbb{E}_q\{f(\boldsymbol{x})\} = \int_{\mathbb{X}} q(\boldsymbol{x})f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \ .$$

Substituting Eq. (8) into Eq. (7) and exchanging the minimization of a negative for maximization yields

$$\boldsymbol{\phi}^*(\boldsymbol{x}) = \underset{\boldsymbol{\phi}}{\operatorname{argmax}}\ \mathbb{E}_q\left\{\nabla_{\boldsymbol{x}}\cdot\boldsymbol{\phi}(\boldsymbol{x})+\boldsymbol{\phi}(\boldsymbol{x})\cdot\nabla_{\boldsymbol{x}}\ln\tilde{\rho}(\boldsymbol{x})\right\} \ . \tag{9}$$

Unfortunately, Eq. (9) still does not, in general, possess an analytical maximum. The key insight of [11] is that restricting the pseudo-dynamics to a reproducing kernel Hilbert space (RKHS), $\boldsymbol{\phi}(\boldsymbol{x}) \in \{\mathcal{H}^n \ s.t. \ \|\boldsymbol{\phi}(\boldsymbol{x})\|_{\mathcal{H}^n} \leq 1\}$, where $\mathcal{H}^n$ is an n-dimensional RKHS, results in the analytical solution

$$\boldsymbol{\phi}^*(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\xi}\sim q}\left\{\nabla_{\boldsymbol{\xi}}k(\boldsymbol{\xi},\boldsymbol{x})+k(\boldsymbol{\xi},\boldsymbol{x})\nabla_{\boldsymbol{\xi}}\ln\tilde{\rho}(\boldsymbol{\xi})\right\} \ , \tag{10}$$

where $k(\cdot,\cdot)$ is the kernel of the RKHS, $\boldsymbol{\xi}\in\mathbb{X}$ is an integration variable, and $\mathbb{E}_{\boldsymbol{\xi}\sim q}\{\cdot\}$ is the expectation with respect to $q(\boldsymbol{\xi})$. If one selects $\mathcal{H}^n$ to be "sufficiently dense" such that a diverse space of pseudo-dynamics is available, e.g., the radial basis function [10], [11], the restriction to an RKHS does not prevent achieving the limit in Eq. (6).

Approximating the variational density as an ensemble of N particles, where $q(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^N \delta(\boldsymbol{x}-\boldsymbol{x}_i)$, $\delta(\cdot)$ is the Dirac measure, and $\boldsymbol{x}_i$ is the i^{th} particle's state, Eq. (10) can be written simply as [11]

$$\boldsymbol{\phi}^*(\boldsymbol{x}_i) = \frac{1}{N}\sum_{j=1}^N \left[\nabla_{\boldsymbol{\xi}}k(\boldsymbol{\xi},\boldsymbol{x}_i)+k(\boldsymbol{\xi},\boldsymbol{x}_i)\nabla_{\boldsymbol{\xi}}\ln\tilde{\rho}(\boldsymbol{\xi})\right]\Big|_{\boldsymbol{\xi}=\boldsymbol{x}_j} \ . \tag{11}$$

Propagating the particles according to Eqs. (4) and (11), where $\mathrm{d}\boldsymbol{x}_i/\mathrm{d}\tau = \boldsymbol{\phi}^*(\boldsymbol{x}_i)$, until convergence results in an ensemble that approximates independent identically distributed samples of the underlying target density that (weekly) converges to ρ as $N\to\infty$ [11]. The two components in Eq. (11) produce distinct dynamical characteristics in the resulting flow. The second term, containing the gradient of the (proportional) log-density, attracts particles to regions of high probability in ρ. The first term, dependent solely on the gradient of the kernel function, acts as a repulsive force preventing over-convergence and encouraging particle trajectories that explore alternative modes of the target density. This process can be viewed as a deterministic analog to Markov-Chain Monte Carlo methods and the Langevin process induced by Eq. (3). It can also be viewed as a gradient flow of the KL divergence functional, but, for brevity, the reader is referred to [15], [16] for a more complete geometric analysis of SVGD.

B. SVGD for Bayesian Inference

The most frequent application of SVGD is particle approximations of Bayes' rule [11], wherein the proportional density is readily provided by Eq. (1). It is easy to develop recursive filtering equations from the method presented in the previous section [10]; however, there is an additional information theoretic intuition that can be gleaned by examining the structure of the Bayesian formulation. Substituting $\tilde{\rho}(\boldsymbol{x}) = p(\boldsymbol{z}|\boldsymbol{x})p(\boldsymbol{x})$ into Eq. (3) yields

$$J[q(\boldsymbol{x})||p(\boldsymbol{x}|\boldsymbol{z})] = \mathscr{F}_{VI}[q] = \mathbb{E}_q\left\{\ln\left(\frac{q(\boldsymbol{x})}{p(\boldsymbol{z}|\boldsymbol{x})p(\boldsymbol{x})}\right)\right\} \ ,$$

which, from the properties of logarithms and the definition of the KL divergence, can easily be manipulated to

$$\mathscr{F}_{VI}[q] = \underbrace{-\mathbb{E}_q\left\{\ln\left(p(\boldsymbol{z}|\boldsymbol{x})p(\boldsymbol{x})\right)\right\}}_{\text{potential "energy"}} + \underbrace{\mathbb{E}_q\left\{\ln q(\boldsymbol{x})\right\}}_{\text{neg. entropy}} \ . \quad (12)$$

The process of minimizing Eq. (12) is often referred to as variational inference [17], and the cost functional, $\mathscr{F}_{VI}$, can be separated into two main components: the negative expected log-proportional density and the negative differential Shannon entropy [18], both with respect to the variational density, q. The former can be viewed as a potential surface whose "valleys" are regions of state space corresponding to high Bayesian posterior probability, and the latter, as a negative entropy measure, promotes uncertainty in the variational density. Minimizing the potential surface, which can be viewed as an artificial potential energy, results in variational posteriors, q^*, that collapse to a single Dirac measure at the Bayesian maximum *a posteriori* state. Conversely, the global minimum of the negative Shannon entropy is a uniform distribution over the state space, $\mathbb{X}$ [18]. The inclusion of both terms ensures variational probability mass is attracted to the peaks of the Bayesian posterior pdf without over converging. Additionally, because Eq. (12) can be written as the sum of a potential functional and a negative entropy functional, the name variational free energy is typically ascribed to $\mathscr{F}_{VI}$ (also referred to as the negative evidence lower bound [17]), indicating its relation to the Helmholtz free energy in thermodynamics [19]. Continuing with the analogy to mechanical work, the variational free energy can be viewed as a heuristic measure of information energy freely available to conduct inference.

C. Extension to Generalized Variational Inference

The motivation behind the previous sections derivations is twofold. First, it provides the explicit definition for Bayesian SVGD particle flow. Second, it emphases the role information theory can play in formulating statistical inference laws. This concept is succinctly summarized by Villani, "Behind many *nonequilibrium* equations of statistical mechanics, there is a variational principle involving entropy and energy, or functionals alike... [19]." An application of this prospective can mitigate estimation errors and statistical inconsistencies stemming from misspecified prior and likelihood pdfs while remaining computationally tractable. This is similarly accomplished by reexamining the inference problem in an optimization-centric light, where variational posteriors are governed by [1]

$$q^*(\boldsymbol{x}) = \underset{q(\boldsymbol{x})}{\operatorname{argmin}}\Big(\mathbb{E}_q\left\{\ell(\boldsymbol{x},\boldsymbol{z})\right\} + D\left[q(\boldsymbol{x})||p(\boldsymbol{x})\right]\Big) \ , \quad (13)$$

where $\ell(\cdot,\cdot)$ is a user-defined loss function associated with the measurement and $D[\cdot||\cdot]$ is a generic statistical divergence measure. Deriving variational posteriors from Eq. (13) is referred to as generalized variational inference (GVI). When an analytical solution to Eq. (13) does not exist, SVGD can be extended to compute samples from the true GVI posterior.

In order to utilize SVGD, the underlying cost, or discrepancy, functional must be defined as an expectation and linear with respect to the pseudo-dynamics. This is facilitated most readily by selecting the divergence as the KL divergence while leaving the loss function generic. The result is a subset of GVI known as Gibbs variational inference with free energy functional, $\mathscr{F}_G$, given by

$$\mathscr{F}_G[q] = \mathbb{E}_q\left\{\ell(\boldsymbol{x},\boldsymbol{z})\right\} + D_{KL}\left[q(\boldsymbol{x})||p(\boldsymbol{x})\right] \ . \quad (14)$$

Though an equivalent procedure to Sec. II-A could be undertaken to derive the SVGD pseudo-dynamics in the "steepest descent direction" of Eq. (14), it is more concise to analytically derive a proportional optimal variational posterior and substitute the result for the target density in Equation (10). Pursuant to this approach, Eq. (14) is separated into the requisite energy and entropy expressions by first applying the definition of the KL divergence in Eq. (2), resulting in

$$\begin{aligned}\mathscr{F}_G[q] &= \mathbb{E}_q\left\{\ell(\boldsymbol{x},\boldsymbol{z})\right\} + \mathbb{E}_q\left\{\ln q(\boldsymbol{x})\right\} - \mathbb{E}_q\left\{\ln p(\boldsymbol{x})\right\} \\ &= \mathbb{E}_q\left\{\ell(\boldsymbol{x},\boldsymbol{z}) - \ln p(\boldsymbol{x})\right\} + \mathbb{E}_q\left\{\ln q(\boldsymbol{x})\right\} \ .\end{aligned}$$

It can easily be shown that variational costs of the form

$$\mathscr{F}_G[q] = \int_{\mathbb{X}} q(\boldsymbol{x})\Psi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_{\mathbb{X}} q(\boldsymbol{x})\ln q(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \quad (15)$$

are minimized, with respect to the density q, by the Gibbs distribution [20]

$$q^*(\boldsymbol{x}) \propto \tilde{\rho}(\boldsymbol{x}) = \exp\left\{-\Psi(\boldsymbol{x})\right\} \ ,$$

provided $\int_{\mathbb{X}}\exp\left\{-\Psi(\boldsymbol{x})\right\}\mathrm{d}\boldsymbol{x}$ exists. The resulting log-proportional density is then simply

$$\ln\tilde{\rho}(\boldsymbol{x}) = -\Psi(\boldsymbol{x}) \ ,$$

with gradient

$$\nabla_{\boldsymbol{x}}\ln\tilde{\rho}(\boldsymbol{x}) = -\nabla_{\boldsymbol{x}}\Psi(\boldsymbol{x}) \ .$$

Substituting this result into Eq. (10) yields the Gibbs variational inference SVGD pseudo-dynamics

$$\boldsymbol{\phi}^*(\boldsymbol{x}_i) = \frac{1}{N}\sum_{j=1}^{N}\Big[\nabla_{\boldsymbol{\xi}}k(\boldsymbol{\xi},\boldsymbol{x}_i) - k(\boldsymbol{\xi},\boldsymbol{x}_i)\nabla_{\boldsymbol{\xi}}\Psi(\boldsymbol{x})\Big]\Big|_{\boldsymbol{\xi}=\boldsymbol{x}_j} \ . \quad (16)$$

Letting $\Psi(\boldsymbol{x}) = \ell(\boldsymbol{x},\boldsymbol{z}) - \ln p(\boldsymbol{x})$ for a given observation $\boldsymbol{z}$, where

$$\nabla_{\boldsymbol{x}}\Psi(\boldsymbol{x}) = \nabla_{\boldsymbol{x}}\ell(\boldsymbol{x},\boldsymbol{z}) - \nabla_{\boldsymbol{x}}\ln p(\boldsymbol{x}) \ ,$$

Eq. (16) can be used to sample potentially infinite dimensional non-Bayesian posteriors that are optimal in the minimum Gibbs free energy sense. When the loss function is the negative log-likelihood, $\ell(\boldsymbol{x}, \boldsymbol{z}) = -\ln p(\boldsymbol{z}|\boldsymbol{x})$, it can easily be shown that the potential surface is $\Psi(\boldsymbol{x}) = -\ln[p(\boldsymbol{z}|\boldsymbol{x})p(\boldsymbol{x})]$, recovering the Bayesian SVGD case and the variational free energy in Equation (12).

III. Gaussian Filtering with Erroneous Measurement Modeling

Oftentimes in practical estimation, such as object tracking or onboard navigation, it is necessary to operate with incomplete and possibly inaccurate measurement modeling. To mitigate the effects of processing errant measurements, *ad hoc* methods, such as underweighting [8] and residual editing [9], are frequently employed. However, there are both theoretical and practical advantages to retaining optimality and robustness when filtering in the presence of measurement model errors. This section applies the previous theoretical developments to a single motivating scenario. To remain congruent with conventional estimators, such as the extended Kalman and linear-Gaussian Bayes' filters [2], the prior and likelihood functions are assumed to be Gaussian.

A convenient and robust loss function, derived from the γ-divergence for the assumed likelihood function, is the γ-loss given by [21]

$$\mathcal{L}_\gamma(\boldsymbol{x}, \boldsymbol{z}) = \frac{\gamma}{1-\gamma}\left(p(\boldsymbol{z}|\boldsymbol{x})\left[\int_{\mathbb{R}^m} p^\gamma(\boldsymbol{s}|\boldsymbol{x})\mathrm{d}\boldsymbol{s}\right]^{1/\gamma}\right)^{\gamma-1} , \tag{17}$$

where $\gamma > 1$. For numerical stability, the negative logarithm of Eq. (17) is used [20]

$$\ell(\boldsymbol{x}, \boldsymbol{z}) = -\ln\left(\frac{\gamma}{\gamma-1}\left[p(\boldsymbol{z}|\boldsymbol{x})\left\{\int_{\mathbb{R}^m} p^\gamma(\boldsymbol{s}|\boldsymbol{x})\mathrm{d}\boldsymbol{s}\right\}^{\frac{1}{\gamma}}\right]^{\gamma-1}\right) ,$$

where $\boldsymbol{z} \in \mathbb{R}^m$. Assuming a nominal, possibly incorrect, measurement model of the form

$$\boldsymbol{z} = \boldsymbol{h}(\boldsymbol{x}) + \boldsymbol{\nu} , \tag{18}$$

where $\boldsymbol{h}(\boldsymbol{x})$ is the deterministic nonlinear measurement function, $\boldsymbol{v} \sim p_g(\boldsymbol{v}; \boldsymbol{0}, \boldsymbol{R})$ is additive Gaussian noise, and

$$p_g(\boldsymbol{a}; \boldsymbol{m}, \boldsymbol{P}) = |2\pi\boldsymbol{P}|^{-1/2} \times \exp\left\{-\frac{1}{2}(\boldsymbol{a}-\boldsymbol{m})^T\boldsymbol{P}^{-1}(\boldsymbol{a}-\boldsymbol{m})\right\} ,$$

is a multivariate Gaussian pdf in $\boldsymbol{a}$ with mean $\boldsymbol{m}$ and covariance $\boldsymbol{P}$. The model in Eq. (18) yields a Gaussian likelihood function of the form

$$p(\boldsymbol{z}|\boldsymbol{x}) = p_g(\boldsymbol{z}; \boldsymbol{h}(\boldsymbol{x}), \boldsymbol{R}) .$$

Substituting the assumed likelihood into the γ-loss function results in

$$\ell(\boldsymbol{x}, \boldsymbol{z}) = -\ln\left(\frac{\gamma}{\gamma-1}\Big[p_g(\boldsymbol{z}; \boldsymbol{h}(\boldsymbol{x}), \boldsymbol{R}) \times \left\{\int_{\mathbb{R}^m}\big(p_g(\boldsymbol{s}; \boldsymbol{h}(\boldsymbol{x}), \boldsymbol{R})\big)^\gamma \mathrm{d}\boldsymbol{s}\right\}^{\frac{1}{\gamma}}\Big]^{\gamma-1}\right) , \tag{19}$$

which can be further simplified from the Gaussian structure of the likelihood. First, the integral can be evaluated as

$$\begin{aligned} C &= \int_{\mathbb{R}^m}\Big(p_g(\boldsymbol{s}; \boldsymbol{h}(\boldsymbol{x}), \boldsymbol{R})\Big)^\gamma \mathrm{d}\boldsymbol{s} \\ &= (1/\gamma)^{m/2}\,|2\pi\boldsymbol{R}|^{(1-\gamma)/2} . \end{aligned}$$

Substituting C into Eq. (19),

$$\ell(\boldsymbol{x}, \boldsymbol{z}) = -\ln\left(\frac{\gamma}{\gamma-1}\Big[p_g(\boldsymbol{z}; \boldsymbol{h}(\boldsymbol{x}), \boldsymbol{R})\, C^{1/\gamma}\Big]^{\gamma-1}\right) ,$$

and expanding the loss function from the properties of logarithms yields

$$\begin{aligned} \ell(\boldsymbol{x}, \boldsymbol{z}) = &-\ln\left(\frac{\gamma}{\gamma-1}\right) - \left(\frac{\gamma-1}{\gamma}\right)\ln C \\ &+ (1-\gamma)\ln p_g(\boldsymbol{z}; \boldsymbol{h}(\boldsymbol{x}), \boldsymbol{R}) . \end{aligned}$$

From the definition of the Gaussian pdf, the log-likelihood term can be expanded to

$$\begin{aligned} \ell(\boldsymbol{x}, \boldsymbol{z}) = &-\ln\left(\frac{\gamma}{\gamma-1}\right) - \left(\frac{\gamma-1}{\gamma}\right)\ln C + \frac{1}{2}(\gamma-1)\ln|2\pi\boldsymbol{R}| \\ &+ \frac{1}{2}(\gamma-1)\big(\boldsymbol{z}-\boldsymbol{h}(\boldsymbol{x})\big)^T\boldsymbol{R}^{-1}\big(\boldsymbol{z}-\boldsymbol{h}(\boldsymbol{x})\big) . \end{aligned}$$

Recognizing the preceding three terms are constant with respect to $\boldsymbol{x}$, the loss function can be further simplified without altering the underlying minimization to

$$\ell(\boldsymbol{x}, \boldsymbol{z}) = \frac{\gamma-1}{2}\big(\boldsymbol{z}-\boldsymbol{h}(\boldsymbol{x})\big)^T\boldsymbol{R}^{-1}\big(\boldsymbol{z}-\boldsymbol{h}(\boldsymbol{x})\big) .$$

The GVI potential function can then be written as

$$\Psi(\boldsymbol{x}) = \frac{\gamma-1}{2}\big(\boldsymbol{z}-\boldsymbol{h}(\boldsymbol{x})\big)^T\boldsymbol{R}^{-1}\big(\boldsymbol{z}-\boldsymbol{h}(\boldsymbol{x})\big) - \ln p(\boldsymbol{x}) .$$

If the prior is also Gaussian, $p(\boldsymbol{x}) = p_g\left(\boldsymbol{x}; \boldsymbol{m}^-, \boldsymbol{P}^-\right)$, with mean, $\boldsymbol{m}^-$, and covariance, $\boldsymbol{P}^-$, the potential function can be simplified to

$$\begin{aligned} \Psi(\boldsymbol{x}) = &\frac{\gamma-1}{2}\big(\boldsymbol{z}-\boldsymbol{h}(\boldsymbol{x})\big)^T\boldsymbol{R}^{-1}\big(\boldsymbol{z}-\boldsymbol{h}(\boldsymbol{x})\big) + \frac{1}{2}\ln|2\pi\boldsymbol{P}^-| \\ &+ \frac{1}{2}\big(\boldsymbol{x}-\boldsymbol{m}^-\big)^T(\boldsymbol{P}^-)^{-1}\big(\boldsymbol{x}-\boldsymbol{m}^-\big) , \end{aligned}$$

that, by again removing constant terms not dependent on $\boldsymbol{x}$, results in a potential of the form

$$\begin{aligned} \Psi(\boldsymbol{x}) = &\frac{\gamma-1}{2}\big(\boldsymbol{z}-\boldsymbol{h}(\boldsymbol{x})\big)^T\boldsymbol{R}^{-1}\big(\boldsymbol{z}-\boldsymbol{h}(\boldsymbol{x})\big) \\ &+ \frac{1}{2}\big(\boldsymbol{x}-\boldsymbol{m}^-\big)^T(\boldsymbol{P}^-)^{-1}\big(\boldsymbol{x}-\boldsymbol{m}^-\big) . \end{aligned}$$

Finally, taking the gradient of the previous equation gives

$$\nabla_{\boldsymbol{x}}\Psi(\boldsymbol{x}) = -(\gamma-1)\boldsymbol{H}(\boldsymbol{x})\boldsymbol{R}^{-1}\big(\boldsymbol{z}-\boldsymbol{h}(\boldsymbol{x})\big) + (\boldsymbol{P}^{-})^{-1}\big(\boldsymbol{x}-\boldsymbol{m}^{-}\big) \ , \tag{20}$$

where

$$\boldsymbol{H}(\boldsymbol{x}) = \frac{\partial \boldsymbol{h}(\boldsymbol{x})}{\partial \boldsymbol{x}} \ ,$$

is the measurement Jacobian matrix. A common choice for the kernel in Eq. (16) is the radial basis function (RBF) or Gaussian kernel, given by

$$k(\boldsymbol{\xi},\boldsymbol{x}) = \exp\left\{\frac{(\boldsymbol{\xi}-\boldsymbol{x})^T(\boldsymbol{\xi}-\boldsymbol{x})}{h^2}\right\} \ ,$$

where h is the user-defined bandwidth parameter.

IV. Results and Discussion

As a motivating test case, the performance of the proposed GVI filter is compared to standard estimation schemes for two simplified scenarios in which the measurement noise statistics are improperly characterized.

A. Two-Dimensional Example

Suppose that a hypothetical range sensor is employed in a navigation filter to estimate a two-dimensional position state, $\boldsymbol{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^T$, with nonlinear measurement model

$$h(\boldsymbol{x}) = \sqrt{x_1^2 + x_2^2} \ .$$

From "pre-flight" calibration the sensor noise is approximated as the zero-mean Gaussian $\nu \sim p_g\left(\nu; 0, 1\right)$ [m]. However, the true "in-flight" noise is both biased and non-Gaussian, with pdf $\nu \sim p_c(\nu; \frac{1}{2}, \frac{1}{2})$, where

$$p_c(\nu;\nu_0,\alpha) = \frac{1}{\pi}\left[\frac{\alpha}{(\nu-\bar{\nu})^2+\alpha^2}\right] \ ,$$

is a Cauchy pdf in ν with mode $\bar{\nu}$ and scale parameter α. The Cauchy density is notoriously difficult to implement, particularly for minimum mean-square error filters, as it possess no finite statistical moments [22]. The two noise pdfs are compared in Fig.1, with specific emphasis placed on the "heavy-tailed" behavior of the Cauchy, suggesting an extremely high probability of sampling outside three standard deviations with respect to the Gaussian.

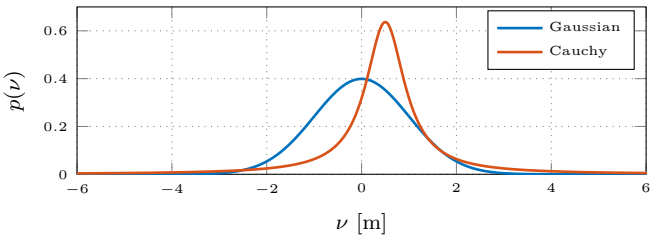


Fig. 1. Gaussian and Cauchy (true) measurement noise pdfs.

The result of a single range measurement, z, update with zero-mean Gaussian prior, $p(\boldsymbol{x}) = p_g\left(\boldsymbol{x}; \boldsymbol{0}, \boldsymbol{P}_0\right)$, where

$$\boldsymbol{P}_0 = \begin{bmatrix} 45 & -25 \\ -25 & 45 \end{bmatrix} \ [m^2] \ ,$$

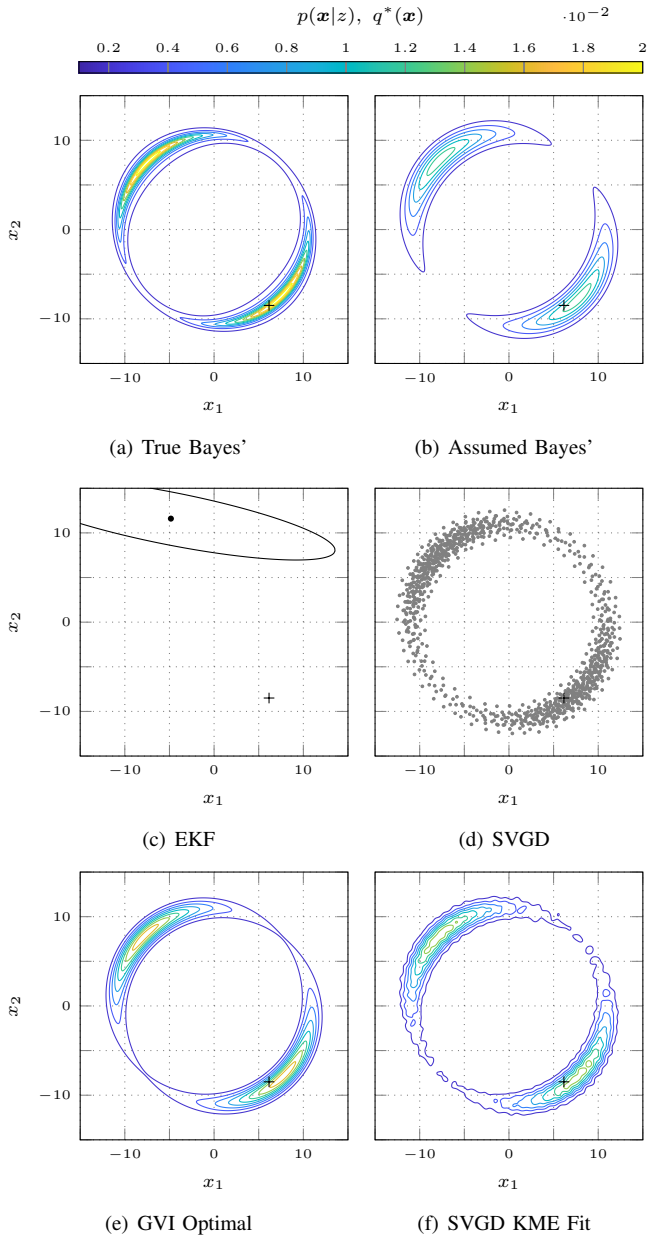
is shown in Fig. 2. Figures 2(a) and 2(b) compare the true Bayesian posterior, resulting from the Cauchy likelihood, against the assumed Bayesian posterior, derived from the "calibrated" Gaussian likelihood, $p_g\left(z; h(\boldsymbol{x}), 1\right)$. The assumed likelihood function does not prescribe the appropriate probability mass concentration at either mode compared to the true Bayesian posterior, nor does it adequately capture the posterior "tails" connecting both modes. Figures 2(c) and 2(d), respectively, contrast a standard approach, which can be interpreted as either an extended Kalman filter (EKF) or linearized Gaussian filter, to the proposed SVGD GVI particle flow update, with $\gamma = 2.75$, $N = 1000$ and an RBF kernel using the "mean trick" discussed in [11]. The EKF posterior is plotted as the ± 3 standard deviation (σ) ellipse. A continuous density can then be fit to the converged SVGD particle set, Fig. 2(f), using kernel mean embedding (KME) [23] (RBF kernel with bandwidth $h = 0.15$), and can be compared to the GVI optimal posterior in Fig. 2(e). The SVGD update not only preserves the bimodal nature of the posterior, compared to the linear EKF update, but by selecting a robust GVI loss function the proposed update scheme was able to better approximate the Bayesian posterior while remaining ignorant to the true likelihood function.

B. Relative Spacecraft Navigation Example

The proposed update scheme can be extended to sequential Gaussian filtering using the method in [10], where a continuous prior density is approximated at each time step from the mean and covariance of the *a priori* particle set. The same range sensor, with identical calibrated (assumed) noise characteristics, is applied to a relative spacecraft navigation scenario. Assuming the two spacecraft share an orbit plane, the state is comprised of the radial, $(\cdot)_r$, and along-track, $(\cdot)_a$, components of the relative position and velocity, $\boldsymbol{x} = \begin{bmatrix} r_r & r_a & v_r & v_a \end{bmatrix}^T$, where $\boldsymbol{r}$ and $\boldsymbol{v}$ are the position and velocity, respectively, of the chaser spacecraft with respect to the target spacecraft. Using a Clohessy-Wiltshire model, where the target is fixed at the origin, the discrete state transition matrix for motion about the target is analytically known as [24]

$$\boldsymbol{F}(\Delta t) = \begin{bmatrix} 4-3c & 0 & s/n & \frac{2}{n}(1-c) \\ 6(s-n\Delta t) & 1 & -\frac{2}{n}(1-c) & \frac{1}{n}(4s-3n\Delta t) \\ 3ns & 0 & c & 2s \\ -6n(1-c) & 0 & -2s & 4c-3 \end{bmatrix}$$

where $\boldsymbol{x}(t+\Delta t) = \boldsymbol{F}(\Delta t)\boldsymbol{x}(t)$, $n = 0.0011314$ [rad/s], $\Delta t = 30$ [s], $c = \cos(n\Delta t)$, and $s = \sin(n\Delta t)$ for the modeled scenario. The true trajectory is plotted in Fig. 3 for a one hour tracking arc. The prior at the initial epoch is Gaussian with mean $\boldsymbol{m}_0 = \begin{bmatrix} 0 & 1000 & -0.3 & 0.1 \end{bmatrix}^T$ (units are [m] and [m/s], respectively) and covariance



(a) True Bayes' (b) Assumed Bayes'

(c) EKF (d) SVGD

(e) GVI Optimal (f) SVGD KME Fit

Fig. 2. Various update performance for the true Bayesian posterior (a), Bayes' rule with the assumed Gaussian likelihood (b), extended Kalman filter (c), SVGD particle posterior (d), optimal GVI posterior (e), and approximated Gaussian mixture SVGD posterior using kernel mean embedding (f), along with the true state (+).

$\boldsymbol{P}_0 = \text{diag}\left\{20^2,\ \ 20^2,\ \ 0.05^2,\ \ 0.05^2\right\}$ (units are [m^2] and [m^2/s^2], respectively).

Two implementations of the EKF are compared to the proposed sequential GVI particle flow update in Fig. 4. The standard, or "vanilla" EKF, Fig. 4(a), quickly diverges when processing measurements with noise sampled outside the calibrated 3σ interval. These effects are mitigated by introducing a residual editing scheme into the EKF [9], where a measurement outside the calculated residual 3σ interval is not processed. Though filter divergence is not observed, 7.5% of the measurements for this case were rejected, restricting the amount of information available to the filter. The proposed GVI filter, with $\gamma = 1.5$, was able to produce equivalent estimates without rejecting measurements or linearizing the underlying models. For computational tractability, the SVGD scheme is implemented using the RBF kernel with the mean trick [11] and a covariance-based scaling, as in [25].

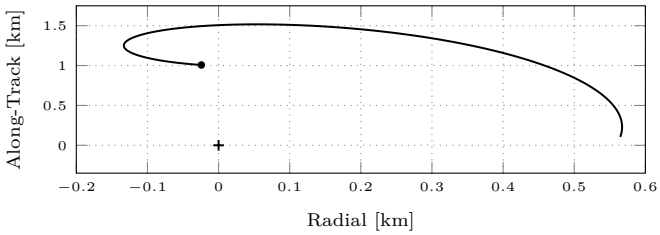


Fig. 3. Relative spacecraft trajectory (—), target position (+), and chaser position at t_0 (·).

V. Conclusion

Bayes' rule defines the most refined posterior probability density function provided complete statistical knowledge is available. Unfortunately, tractability and model accuracy are often in direct competition for practical implementations of Bayesian filters. One approach that can preserve the fidelity of nonlinear models while remaining computationally feasible is particle flow updates, where the *a posteriori* particle ensemble resembles independent sampling of the true posterior density. One form of nonlinear particle flow, Stein Variational Gradient Descent (SVGD), facilitates deterministic flow dynamics that are not limited to Bayes' rule. This flexibility of SVGD lends exceptionally well to non-Bayesian methods, such as generalized variational inference (GVI). GVI is an information-theoretic approach for formulating optimal statistical inference that remains cognizant and robust to model errors. GVI with the Kullback-Liebler divergence and an arbitrary loss function, known as Gibbs inference, can be implemented using SVGD. Such a filter is extended to both a two-dimensional and spacecraft navigation scenarios and demonstrated improved performance over traditional methods, such as the extended Kalman filter. Future work will focus of extension of SVGD filtering to full GVI with arbitrary statistical divergence functionals.

References

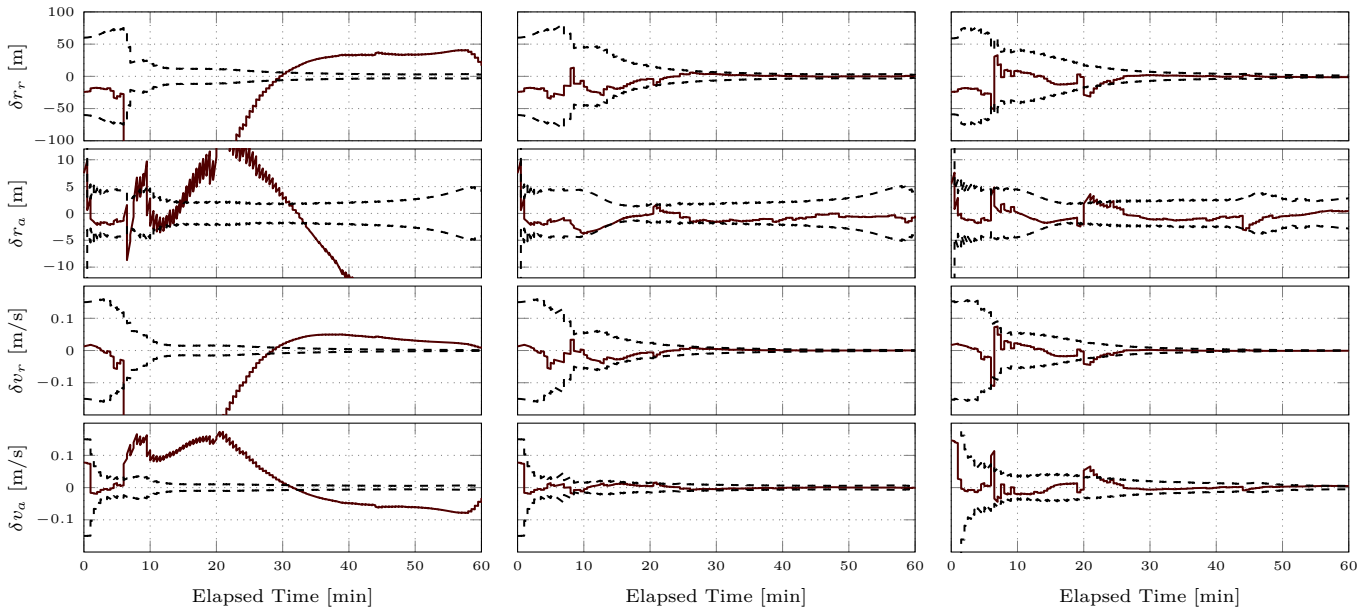
[1] J. Knoblauch, J. Jewson, and T. Damoulas, "Generalized variational inference: Three arguments for deriving new posteriors." arXiv:1904.02063, Dec. 2019.

[2] Y. Ho and R. Lee, "A Bayesian approach to problems in stochastic estimation and control," *IEEE Transactions on Automatic Control*, vol. 9, no. 4, pp. 333–339, 1964.

[3] D. Alspach and H. Sorenson, "Nonlinear Bayesian estimation using gaussian sum approximations," *IEEE Transactions on Automatic Control*, vol. 17, no. 4, pp. 439–448, 1972.

[4] S. Särkkä, *Bayesian Filtering and Smoothing*. Cambridge University Press, 2013.

[5] F. Daum and J. Huang, "Particle degeneracy: root cause and solution," in *Signal Processing, Sensor Fusion, and Target Recognition XX*, vol. 8050, p. 80500W, SPIE, 2011.



(a) Vanilla EKF (b) Residual-Edited EKF (c) GVI-SVGD

Fig. 4. Filter performance as estimation error (—) and $\pm 3\sigma$ confidence interval (**- -**) for the unaltered, or vanilla, EKF (a), the residual-edited EKF (b), and the proposed GVI-SVGD Gaussian particle flow filter (c).

[6] F. Daum and J. Huang, "Nonlinear filters with particle flow induced by log-homotopy," in *Signal Processing, Sensor Fusion, and Target Recognition XVIII*, vol. 7336, p. 733603, SPIE, 2009.

[7] K. C. Ward and K. J. DeMars, "Information-based particle flow with convergence control," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 58, no. 2, pp. 1377–1390, 2022.

[8] R. Zanetti, K. J. DeMars, and R. H. Bishop, "Underweighting nonlinear measurements," *Journal of Guidance, Control, and Dynamics*, vol. 33, no. 5, pp. 1670–1675, 2010.

[9] J. R. Carpenter and C. N. D'Souza, "Navigation filter best practices," Tech. Rep. NASA/TP–2018–219822, NASA, Apr. 2018.

[10] K. J. Craft and K. J. DeMars, "Optimal nonlinear particle flow using Stein variational gradient descent," in *Proceedings of the AAS/AIAA Astrodynamics Specialist Conference*, 2022.

[11] Q. Liu and D. Wang, "Stein variational gradient descent: A general purpose Bayesian inference algorithm," in *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[12] S. Kullback, *Information Theory and Statistics*. Mineola, NY: Dover Publications, Inc., 1968.

[13] H. Risken, *The Fokker-Planck Equation*. Springer, 2 ed., 1996.

[14] O. D. Kellogg, *Foundations of Potential Theory*, ch. 5. The Divergence Theorem, pp. 84–121. Berlin, Heidelberg: Springer, 1967.

[15] Q. Liu, "Stein variational gradient descent as gradient flow," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[16] A. Duncan, N. Nuesken, and L. Szpruch, "On the geometry of stein variational gradient descent." arXiv:1912.00894, Dec. 2019.

[17] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[18] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley, 2 ed., 2006.

[19] C. Villani, *Optimal Transport, Old and New*, p. 694. Springer, 2008.

[20] R. Jordan, D. Kinderlehrer, and F. Otto, "The variational formulation of the Fokker-Planck equation," *Journal on Mathematical Analysis*, vol. 29, no. 1, pp. 1–17, 1998.

[21] H. Hung, Z.-Y. Jou, and S.-Y. Huang, "Robust mislabel logistic regression without modeling mislabel probabilities," *Biometrics*, vol. 74, no. 1, pp. 145–154, 2018.

[22] J. L. Speyer, M. Idan, and J. Fernández, "The two-state estimator for linear systems with additive measurement and process Cauchy noise," in *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pp. 4107–4114, 2012.

[23] K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf, "Kernel mean embedding of distributions: A review and beyond," *Foundations and Trends® in Machine Learning*, vol. 10, no. 1-2, pp. 1–141, 2017.

[24] D. A. Vallado, *Fundementals of Astrodynamics and Applicaitons*, ch. 8. Microcosm Press, 4 ed., 2013.

[25] D. Wang, Z. Tang, C. Bajaj, and Q. Liu, "Stein variational gradient descent with matrix-valued kernels." arXiv:1910.12794, Nov. 2019.