A Variational Approach to Robust Bayesian Filtering

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*—A major challenge of applied Bayesian filtering is deriving estimates that are robust to misspecifications in the underlying statistical models, particularly when Bayes' rule does not directly yield analytical posterior probability densities. Variational approaches present a promising alternative to "traditional," closed-form Bayesian inference, wherein an approximate posterior is defined by minimizing the statistical dissimilarity to the conventional Bayesian posterior. There are, however, numerous realistic hurdles to defining an ideal dissimilarity measure, from which posteriors are derived using calculus of variations. This work utilizes a recently proposed framework, known as generalized variational inference (GVI), to define robust and tractable approximations of Bayes' rule. The GVI framework is presented and accompanied by a novel sensitivity analysis and gradient-based solution method that is applicable for particle, Gaussian, and Gaussian mixture posterior representations. The proposed GVI filter is applied to a dynamic state estimation scenario with inaccurate measurement modeling and, via Monte Carlo analysis, demonstrates both statistical consistency and improvements over conventional robust filtering methods.**

***Index Terms*—Bayesian inference, variational inference, nonlinear estimation, information theory**

I. Introduction

The most fundamental problem in applied statistical inference is the accurate, precise, and consistent fusion of existing information for an uncertain "state" variable, $\boldsymbol{x}$, with external, noisy observations, $\boldsymbol{z}$. Provided the prior and observation can be described statistically by their respective probability density functions (pdfs), the inference problem is conventionally solved using Bayes' rule [1],

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

where $p(\boldsymbol{x})$ is the prior state pdf, $p(\boldsymbol{z}|\boldsymbol{x})$ is the observation pdf conditioned on the state, commonly referred to as the likelihood function, and $p(\boldsymbol{x}|\boldsymbol{z})$ is the posterior state pdf. The denominator of (1), $p(\boldsymbol{z}) = \int_{\mathbb{X}} p(\boldsymbol{z}|\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$, ensures the posterior is a proper pdf (i.e., integrates to unity over $\boldsymbol{x} \in \mathbb{X} \subseteq \mathbb{R}^n$). Analytical solutions to (1) are notoriously difficult to derive for real-world applications, particularly for estimating the state of nonlinear dynamical systems. Numerical approximations are similarly limited as integration of the normalization constant, $p(\boldsymbol{z})$, grows exponentially computationally prohibitive with n. Practical implementations of (1) are further challenged when inconsistencies arise in the underlying statistical models. To that end, though Bayes' rule represents an undoubtedly powerful inference framework, (1) is only ideal provided the following three criteria are satisfied [2]: (i) the prior is correctly specified, (ii) the likelihood function, derived from the measurement model, is statistically accurate, and (iii) the posterior can be computed tractably.

In practice, (i), (ii), and (iii) can be difficult, if not impossible, to satisfy simultaneously for dynamic systems. Oftentimes, one elects to sacrifice measurement model fidelity in exchange for analytical closure of Bayes' rule. The most prolific examples of this trade-off are the "linear-Gaussian" Bayes' filters, wherein the underlying models are approximated as a linear combination of the state variables and Gaussian white noise. If the prior is further assumed to be Gaussian, or a weighted mixture of Gaussians, Bayes' rule closes analytically, recovering a conjugate Gaussian [3], or Gaussian mixture (GM) [4], posterior. Linear-Gaussian filters are immensely popular for both their computational simplicity and algorithmic similarity to the Kalman filter; however, the resulting estimators inherently risk incurring linearization errors and potential filter inconsistency. Statistical linearization, e.g., the unscented transform [5] and Gauss-quadrature [6], can mitigate linearization errors, but remain generally susceptible to modeling inaccuracies inevitably present in real-world systems. Modern filtering implementations tend to address off-nominal dynamics and observations through various *ad hoc* techniques, such as tuning noise, measurement underweighting, and residual editing [7]. Though useful, these approaches remain susceptible to errant modeling.

This critique is by no means intended to denigrate traditional Bayesian filtering methods or their Kalman filtering counterparts, both of which have enjoyed decades of inarguable success, but rather to motivate an alternative framework when (i), (ii), and/or (iii) cannot be satisfied by traditional estimators. The approach considered in this work, known as generalized variational inference (GVI) [2], recasts statistical inference as the optimization of a robust information-theoretic functional over a family of computationally favorable pdfs, e.g., Gaussians, GMs, and particle ensembles.

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

II. Generalized Variational Inference

A. Variational Bayesian Inference

The GVI formulation starts by considering the traditional variational form of Bayesian inference [8], or variational inference (VI) for short. Rather than artificially manipulating models to ensure analytical closure of (1), VI defines an optimal approximation to Bayes' rule of the form

$$q^*(\boldsymbol{x}) = \underset{q\in\mathcal{Q}(\mathbb{X})}{\operatorname{argmin}}\ D_{KL}\left[q(\boldsymbol{x})||p(\boldsymbol{x}|\boldsymbol{z})\right]\ , \tag{2}$$

where $D_{KL}\left[\cdot||\cdot\right]$ is the Kullback-Leibler (KL) divergence [9] and $q \in \mathcal{Q}(\mathbb{X})$ is an arbitrary candidate density belonging to a variational family of tractable pdfs over the state space, $\mathcal{Q}(\mathbb{X})$. VI offers several alluring simplifications for approximate inference [8], e.g., when $p(\boldsymbol{x}|\boldsymbol{z}) \in \mathcal{Q}(\mathbb{X})$, the KL divergence ensures Bayes' rule is recovered exactly [9].

It is common to operate with a simplified, but equivalent, form of (2),

$$q^*(\boldsymbol{x}) = \underset{q\in\mathcal{Q}(\mathbb{X})}{\operatorname{argmin}}\ \mathscr{F}_B[q(\boldsymbol{x})]\ , \tag{3}$$

where, denoting the expectation with respect to q as $\mathbb{E}_q\left\{\cdot\right\}$,

$$\mathscr{F}_B[q] = -\mathbb{E}_q\left\{\log p(\boldsymbol{z}|\boldsymbol{x})\right\} + D_{KL}\left[q(\boldsymbol{x})||p(\boldsymbol{x})\right]\ , \tag{4}$$

is the negative evidence lower bound (ELBO) [8]. The VI cost functional in (4) is comprised of two components: the negative expected logarithm of the likelihood function (log-likelihood), which encourages posterior probability mass to align with the observed measurement, and the KL divergence from the variational density to the prior, which penalizes posteriors that stray from the prior. Simultaneous minimization of both terms ensures that prior and measurement information are balanced in the posterior. Note that (3) also does not require explicit calculation of the Bayesian normalization constant, $p(\boldsymbol{z})$, which is oftentimes intractable.

An additional information-theoretic insight can be gleaned from (3) by manipulating (4) slightly to

$$\mathscr{F}_B[q] = -\mathbb{E}_q\left\{\log\left(p(\boldsymbol{z}|\boldsymbol{x})p(\boldsymbol{x})\right)\right\} - \mathrm{H}_S[q(\boldsymbol{x})]\ ,$$

where $\mathrm{H}_S[\cdot]$ represents the differential Shannon entropy [9]. As shown in [10], $\mathscr{F}_B$ bears a striking resemblance to the Helmholtz free energy functional in statistical mechanics. The expected value term operates as a potential function (or "energy"), the negative gradient of which attracts probability mass to the posterior modes, while the negative entropy term promotes posterior uncertainty. Subsequently, $\mathscr{F}_B$ is often referred to as the variational free energy (VFE), and, analogous to mechanical free energy, can be viewed as a heuristic measure for the maximum information content (or energy) available to conduct inference. This relation is succinctly summarized in [11], "Behind many *nonequilibrium* equations of statistical mechanics, there is a variational principle involving entropy and energy, or functionals alike..." This analogy aligns with the original variational form of Bayes' rule in [12] as an optimal information processing law, as well as more contemporary interpretations of free energies [13].

The presented interpretation of VI motivates the following proposition. Inherent in any estimator is a commitment, whether explicit or implicit, to an underlying variational principle involving the utilization of information.

B. A Generalized Variational Formulation

The *prima facie* question for practical applications of Bayesian inference is how to optimally utilize, or even measure, information when the underlying prior and likelihood are subject to unmodeled disturbances. Just as alternative free energies exist in statistical mechanics, the information landscape can be altered by generalizing the VFE, such as [2]

$$\mathscr{F}[q] = \mathbb{E}_q\left\{\ell(\boldsymbol{x},\boldsymbol{z})\right\} + D\left[q(\boldsymbol{x})||p(\boldsymbol{x})\right]\ , \tag{5}$$

where $\ell(\boldsymbol{x},\boldsymbol{z})$ is a loss function incorporating information from the measurement and $D\left[q(\boldsymbol{x})||p(\boldsymbol{x})\right]$ is a statistical divergence measure between the variational density and the prior (see [14] for the definition of a divergence). By strategically selecting the loss and divergence terms, one can (ideally) form a variational principle via (5) that measures observation and prior information in a more robust fashion (with respect to Bayes' rule). When $\ell = -\log p(\boldsymbol{z}|\boldsymbol{x})$ and $D = D_{KL}$, information is measured in the traditional Shannon sense [9] and (5) recovers the VI cost in (4).

Though direct structural similarity to the Helmholtz free energy is lost in (5), the underlying information utilization analogy is retained, provided the following are satisfied.

Assumption 1: The loss function is absolutely continuous and integrable with respect to q, such that its expectation is properly defined, $|\mathbb{E}_q\left\{\ell(\boldsymbol{x},\boldsymbol{z})\right\}| < \infty$.

Assumption 2: The divergence measure, $D\left[q(\boldsymbol{x})||p(\boldsymbol{x})\right]$, is defined for $q, p \in \mathcal{P}(\mathbb{X})$ and convex with respect to q.

Definition 1: Given Assumptions 1 and 2 hold, the generalized free energy is also convex (by linearity of the expected value operation) and defines the statistical divergence $D_{\mathscr{F}}\left[q(\boldsymbol{x})||\bar{q}(\boldsymbol{x})\right] \triangleq \mathscr{F}[q(\boldsymbol{x})] - \mathscr{F}[\bar{q}(\boldsymbol{x})]$, where $\bar{q}$ is the $\mathscr{F}$-optimal posterior over the set of all valid pdfs on $\mathbb{X}$, $\mathcal{P}(\mathbb{X})$.

Remark 1: As with Bayes' rule and VI, $\bar{q}$ is typically intractable, but represents the maximum utilization of information as defined by $\mathscr{F}$. Because $\mathscr{F}[\bar{q}]$ is constant, minimizing $\mathscr{F}[q]$ is equivalent to minimizing $D_{\mathscr{F}}$ between q and $\bar{q}$, as in traditional VI, where $D_{\mathscr{F}}\left[q||\bar{q}\right] \to D_{KL}\left[q(\boldsymbol{x})||p(\boldsymbol{x}|\boldsymbol{z})\right]$.

Remark 2: Note that Assumption 2 is of functional convexity, rather than convexity with respect to specific parameterizations of $q \in \mathcal{Q}$, as in information geometry [14]. Common convex divergences include the KL, Rényi, and f-divergences, among others. There are, however, pragmatic choices for D, such as the robust power divergence [15], that reduce sensitivity to prior misspecification, but are not, in general, convex. It is important to recognize that non-convex divergences, while often useful in robust inference, remove a layer of theoretical formalism in the GVI problem, just as in non-convex function minimization. Furthering the analogy to function minimization, when $\mathscr{F}$ is (locally) convex, a functional form of gradient descent, presented in Sec. III, can be used to solve GVI problems.

Attention is now turned to the concept of "information" as measured and optimized in GVI. For brevity, only the loss function is discussed as there exists a plethora of literature on divergences and their connection to information theory (see, for instance, [9] and [14]). Consider, initially, the scalar case for x and z. In the standard VI formulation, the loss function is the negative log-likelihood (NLL), or Shannon information [9], $\ell = -\log p(z|x)$. The Shannon information is heuristically viewed as a measure of surprisal for instantiations of z given x. Assuming the measurement takes the form $z = h(x) + \nu$, where $h(x)$ is the model and $\nu \sim p_g\left(\nu; 0, \sigma_\nu^2\right)$ is Gaussian noise, the observation information grows quadratically with respect to $z - h(x)$, as shown in Fig. 1. For ease of viewing, the losses in Fig. 1 are shifted vertically so $\ell(z - h(x) = 0) = 0$ (only the curve shapes are important for optimization). Conventional Bayes' filters and VI subsequently treat outlying measurements as highly informative, providing strong evidence against prior belief. In reality, however, an outlier may be more indicative of unmodeled anomalies than purely aleatoric phenomena.

To combat "over-fitting" to spurious data, it is potentially beneficial to systematically reduce measurement information during inference, e.g.,

$$\ell_\zeta[p(\boldsymbol{z}|\boldsymbol{x})] = -\log p^\zeta(\boldsymbol{z}|\boldsymbol{x}) = -\zeta \log p(\boldsymbol{z}|\boldsymbol{x}) \ , \qquad (6)$$

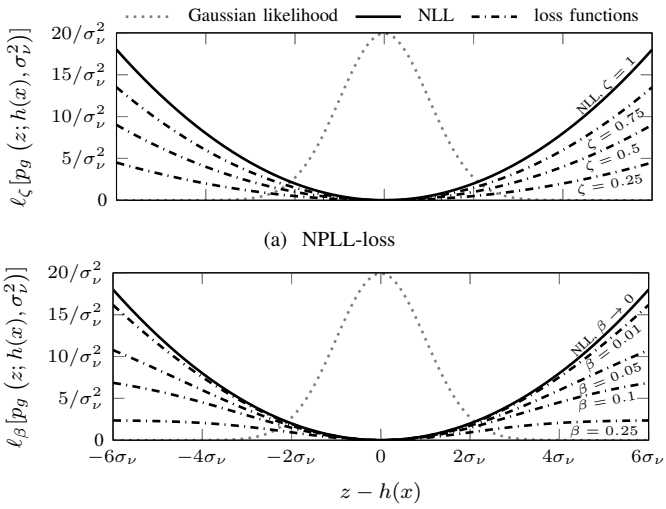
where $0 \leq \zeta \leq 1$ controls the dilution of information via the negative power log-likelihood (NPLL), ℓ_ζ. The result is a Bayesian analog to measurement underweighting common in sequential estimation [16]. For the univariate Gaussian case, where $z = h(x) + \nu$, the down-weighted information from (6) is plotted in Fig. 1a, with $\zeta = 1$ recovering the NLL. Losses of the form of (6) are common in VI and correspond to a power-likelihood approach for robust Bayesian inference [17].

Unfortunately, for the Gaussian case, (6) still results in quadratic surprisal at the tails while simultaneously diluting information around the mode. A potentially more desirable loss is one that preserves the information content when z aligns with the nominal model and diminishes its influence in the tails. One such loss, defined from the robust power divergence in [15], is given by [18]

$$\ell_\beta = -\frac{(\beta+1)}{\beta} p^\beta(\boldsymbol{z}|\boldsymbol{x}) + \int_{\mathbb{Z}} p^{\beta+1}(\boldsymbol{z}|\boldsymbol{x}) \mathrm{d}\boldsymbol{z} \ , \qquad (7)$$

where $\beta > 0$ with $\beta \to 0$ recovering the standard NLL. The univariate Gaussian case is again plotted for (7) in Fig. 1b. Rather than growing unbounded, information is asymptotically flattened at the tails, with β controlling the rate at which information is curtailed. Heuristically, the resulting filter views ever growing outlier behavior as equally unsurprising and, eventually, uninformative, diminishing the influence of outlying/errant measurements over the posterior.

Examining loss functions can provide a view of the information-theoretic landscape within the generalized VFE, $\mathscr{F}$; however, it is difficult to directly intuit performance and robustness of the resulting GVI filter. Consider the univariate Gaussian case with linear measurements, where $z = Hx + \nu$



(a) NPLL-loss

(b) β-loss

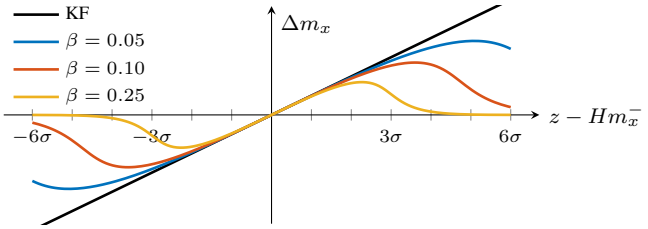
Fig. 1: NPLL and β-loss functions for a univariate Gaussian likelihood with standard deviation σ_ν (Gaussian pdfs not drawn to scale).

and H is a deterministic constant, and Gaussian prior. The linear-Gaussian Bayes' corrector is known analytically and aligns with the Kalman filter (KF) [3], allowing direct comparison between Bayesian posteriors from (1) and the proposed GVI corrector using a β-loss function. To remain congruent with the KF, $\mathcal{Q}$ in (3) is taken to be the set of all Gaussians on $\mathbb{R}$, such that each posterior is defined by its mean and covariance updates, $\Delta m_x = m_x^+ - m_x^-$ and $\Delta P_{xx} = P_{xx}^+ - P_{xx}^-$, respectively. The $(\cdot)^+$ and $(\cdot)^-$ notation is used to indicate *a posteriori* and *a priori* quantities, respectively, throughout this work. The resulting corrections are plotted as a function of the measurement residual, $z - Hm_x^-$, for various values of β in Fig. 2 (where $D = D_{KL}$ for the considered case).

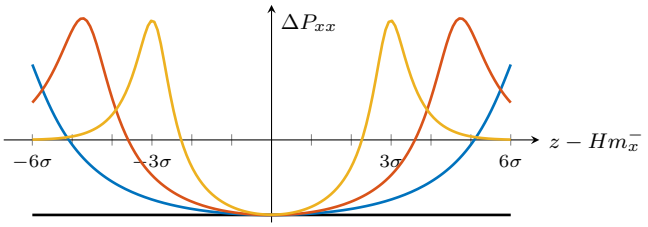
For nominal behavior (residuals near zero), the ideal, linear nature of the Kalman update is preserved. As the residual drifts into the likelihood tails, the asymptotic behavior of ℓ_β yields mean updates, plotted in Fig. 2a, that are curtailed, with the value of β controlling the degree of diminution. For $\beta = 0.1$ and a residual of $\pm 4\sigma$ ($\sigma^2 = H^2 P_{xx}^- + \sigma_\nu^2$), approximately half of the Kalman update is preserved. Similar behavior is observed in Fig. 2b, where the covariance correction approaches 0 as the residual becomes increasingly large. However, when the residual resides in a region of low, but not negligible, probability, there exists evidence that either the prior or likelihood model is incorrect. As such, a partial mean update is observed and the covariance increases to capture both the prior and nominal Bayesian posterior hypotheses.

C. Linear Sensitivity Analysis: Loss and Divergence Selection

While Figs. 1 and 2 can aid in filter design and hyperparameter (e.g., β) selection, they do not provide a complete picture of robustness with respect to specific modeling inadequacies. Pursuant to a more quantitative understanding of



(a) mean update



(b) covariance update

Fig. 2: Mean and covariance update comparison for Kalman filter and Gaussian GVI filter with β-loss and KL divergence.

robustness, a first-order sensitivity analysis inspired by influence functions popular in robust statistics [19] is developed. Let $q = q(\boldsymbol{x};\boldsymbol{\theta}) \in \mathcal{Q}(\mathbb{X})$ be a state pdf parameterized by $\boldsymbol{\theta}$. For instance, if $\mathcal{Q}$ is taken as the family of multivariate Gaussian densities, $\boldsymbol{\theta}$ is a concatenation of the mean vector and covariance matrix. The GVI problem then reduces to

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\ \mathscr{F}[q(\boldsymbol{x};\boldsymbol{\theta})]\ ,$$

where each point $\boldsymbol{\theta}$ defines a pdf in $\mathcal{Q}(\mathbb{X})$. As a result, minimization of the free energy is equivalent to a "maximum-likelihood-type" ρ-estimator considered in [19]. The connections to M- and ρ-estimators, and their subsequent asymptotic normal behavior, is a promising, but unexplored, attribute of GVI. Consistency has been explored in [20]; however, the presented analysis is not derived with the sequential estimation problem in mind.

Consider misspecified likelihood and prior pdfs of the form

$$\begin{aligned} p(\boldsymbol{x}) &= (1-\lambda)\rho(\boldsymbol{x}) + \lambda\Delta\rho(\boldsymbol{x}) \\ p(\boldsymbol{z}|\boldsymbol{x}) &= (1-\varepsilon)\pi(\boldsymbol{x},\boldsymbol{z}) + \varepsilon\Delta\pi(\boldsymbol{x},\boldsymbol{z})\ , \end{aligned}$$

where, respectively, ρ and π are the nominal prior and likelihood densities, $\Delta\pi$ and $\Delta\rho$ are contaminating, deviatoric pdfs, and λ and ε are, ideally small, mixture probabilities. It is then of interest to evaluate the sensitivity, or preferably insensitivity, of the resulting parametric estimate, $\hat{\boldsymbol{\theta}}$, to the misspecification probabilities, λ and ε. This is accomplished by examining the first-order necessary condition, $\boldsymbol{f}(\boldsymbol{\theta},\varepsilon,\lambda) \triangleq \frac{\partial\mathscr{F}[q(\boldsymbol{x};\boldsymbol{\theta})]}{\partial\boldsymbol{\theta}^T} = \boldsymbol{0}$, where, by virtue of the prior and likelihood dependencies on λ and ε, $\mathscr{F}$ in (5) is a function of the misspecification probabilities. Let $\boldsymbol{\theta}^*$ be $\mathscr{F}$-optimal under the true models and $\hat{\boldsymbol{\theta}}$ be "nominally-optimal," i.e., when $\lambda = \varepsilon = 0$.

From a Taylor series expansion of $\boldsymbol{f}$, one can then ascertain, to the first order, the stability of GVI estimates to modeling disparities. Truncating such an expansion yields

$$\boldsymbol{f}(\boldsymbol{\theta}^*,\lambda,\varepsilon) \approx \boldsymbol{f}(\hat{\boldsymbol{\theta}},0,0) + \frac{\partial\boldsymbol{f}}{\partial\boldsymbol{\theta}}\Big|_{(\star)}\Delta\boldsymbol{\theta} + \frac{\partial\boldsymbol{f}}{\partial\lambda}\Big|_{(\star)}\lambda + \frac{\partial\boldsymbol{f}}{\partial\varepsilon}\Big|_{(\star)}\varepsilon\ ,$$

where $\Delta\boldsymbol{\theta} = \boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}$ and $(\star)$ indicates a partial derivative evaluated under the nominal conditions, $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ and $\varepsilon = \lambda = 0$. By definition, $\boldsymbol{f}(\boldsymbol{\theta}^*,\lambda,\varepsilon) = \boldsymbol{f}(\hat{\boldsymbol{\theta}},0,0) = \boldsymbol{0}$, facilitating an estimate discrepancy of

$$\Delta\boldsymbol{\theta} \approx -\left[\frac{\partial\boldsymbol{f}}{\partial\boldsymbol{\theta}}\Big|_{(\star)}\right]^{-1}\left[\left(\frac{\partial\boldsymbol{f}}{\partial\lambda}\Big|_{(\star)}\right)\lambda + \left(\frac{\partial\boldsymbol{f}}{\partial\varepsilon}\Big|_{(\star)}\right)\varepsilon\right]\ . \quad (8)$$

This approximation is $\mathcal{O}(\|\Delta\boldsymbol{\theta}\|^2,\lambda^2,\varepsilon^2)$, where, for small λ and ε, first-order terms govern posterior parameter dissimilarities. Assuming the free energy takes the form

$$\mathscr{F}[q] = \mathbb{E}_q\left\{\ell[p(\boldsymbol{z}|\boldsymbol{x})]\right\} + g\left(\int_{\mathbb{X}} G[q,p]\mathrm{d}\boldsymbol{x}\right)\ , \quad (9)$$

where the function, $g(\cdot)$, and operator, $G[\cdot,\cdot]$, are defined by the utilized divergence, the derivatives of $\boldsymbol{f}$ with respect to λ and ε may be written as

$$\begin{aligned} \frac{\partial\boldsymbol{f}}{\partial\lambda}\Big|_{(\star)} &= g'(\hat{\omega})\int_{\mathbb{X}}\left(\frac{\partial q}{\partial\boldsymbol{\theta}^T}\Big|_{(\star)}\right)\left(\frac{\delta^2 G[q,\rho]}{\delta\rho\delta q}\Big|_{(\star)}\right)(\Delta\rho-\rho)\mathrm{d}\boldsymbol{x} \\ &\quad + g''(\hat{\omega})\left[\int_{\mathbb{X}}\left(\frac{\delta G[q,\rho]}{\delta\rho}\Big|_{(\star)}\right)(\Delta\rho-\rho)\mathrm{d}\boldsymbol{x}\right] \\ &\quad \times\left[\int_{\mathbb{X}}\left(\frac{\partial q}{\partial\boldsymbol{\theta}^T}\Big|_{(\star)}\right)\left(\frac{\delta G[q,\rho]}{\delta q}\Big|_{(\star)}\right)\mathrm{d}\boldsymbol{x}\right] \quad (10a) \\ \frac{\partial\boldsymbol{f}}{\partial\varepsilon}\Big|_{(\star)} &= \int_{\mathbb{X}}\left(\frac{\partial q}{\partial\boldsymbol{\theta}^T}\Big|_{(\star)}\right)\left(\frac{\delta\ell[\pi]}{\delta\pi}\right)(\Delta\pi-\pi)\mathrm{d}\boldsymbol{x}\ , \quad (10b) \end{aligned}$$

where $\delta/\delta q$ is the variation with respect to q, $\hat{\omega} = \int_{\mathbb{X}} G[q,\rho]\mathrm{d}\boldsymbol{x}$, $g'(\omega) = \partial g(\omega)/\partial\omega$, and so forth. If models are available for the deviatoric densities, $\Delta\rho$ and $\Delta\pi$, sensitivities follow from (8) directly: $\Delta\boldsymbol{\theta} = \boldsymbol{s}_\lambda\lambda + \boldsymbol{s}_\varepsilon\varepsilon$, where

$$\boldsymbol{s}_{(\cdot)} = -\left[\frac{\partial\boldsymbol{f}}{\partial\boldsymbol{\theta}}\Big|_{(\star)}\right]^{-1}\left(\frac{\partial\boldsymbol{f}}{\partial(\cdot)}\Big|_{(\star)}\right)\ .$$

An extension is considered when models for $\Delta\rho$ and $\Delta\pi$ are not available. Just as $\boldsymbol{\theta}$ parameterizes q, assume parameterizations exists for ρ and π, $\boldsymbol{\psi}_\rho$ and $\boldsymbol{\psi}_\pi$, respectively. The misspecification densities may be written to the first-order as

$$\begin{aligned} \Delta\rho(\boldsymbol{x}) - \rho(\boldsymbol{x}) &= (\partial\rho/\partial\boldsymbol{\psi}_\rho)\Delta\boldsymbol{\psi}_\rho \\ \Delta\pi(\boldsymbol{x},\boldsymbol{z}) - \pi(\boldsymbol{x},\boldsymbol{z}) &= (\partial\pi/\partial\boldsymbol{\psi}_\pi)\Delta\boldsymbol{\psi}_\pi\ , \end{aligned}$$

with $\rho(\boldsymbol{x}) = \rho(\boldsymbol{x};\boldsymbol{\psi}_\rho)$, $\Delta\rho(\boldsymbol{x}) = \rho(\boldsymbol{x};\boldsymbol{\psi}_\rho + \Delta\boldsymbol{\psi}_\rho)$, etc. The sensitivity of GVI estimates to model misspecifications then becomes $\Delta\boldsymbol{\theta} = \boldsymbol{S}_\rho\Delta\boldsymbol{\psi}_\rho\lambda + \boldsymbol{S}_\pi\Delta\boldsymbol{\psi}_\pi\varepsilon$, where $\boldsymbol{S}_\rho$ and $\boldsymbol{S}_\pi$ are defined by exchanging $\Delta\rho - \rho$ and $\Delta\pi - \pi$ for $\partial\rho/\partial\boldsymbol{\psi}_\rho$ and $\partial\pi/\partial\boldsymbol{\psi}_\pi$, respectively, in (10).

To illustrate the proposed approach, consider, again, the linear-Gaussian case where $\rho = p_g\left(x;m_x^-,P_{xx}^-\right)$ and $\pi = p_g\left(z;Hx+m_\nu,P_{\nu\nu}\right)$. For brevity, only a β-loss function sensitivity analysis is evaluated. The KL divergence is used in defining $\mathscr{F}$ with $g(\omega) = \omega$ and $G[q,p] = q\log(q/p)$ [9]. The

likelihood, being Gaussian, is parameterized by its mean, m_ν (nominally unbiased, $m_\nu = 0$), and covariance. Fig. 3 displays the subsequent sensitivities of the linear-Gaussian case for various values of β and the KF, where

$$\boldsymbol{\psi}_\pi^T = \begin{bmatrix} m_\nu & P_{\nu\nu} \end{bmatrix} \text{ and } \boldsymbol{S}_\pi = \begin{bmatrix} \Delta m_x/\Delta m_\nu & \Delta m_x/\Delta P_{\nu\nu} \\ \Delta P_{xx}/\Delta m_\nu & \Delta P_{xx}/\Delta P_{\nu\nu} \end{bmatrix}$$

for a given residual $z - Hm_x^-$. The KF, representing a linear update, has linear sensitivity across the entire space of residuals. GVI with a β-loss demonstrates reduced sensitivity (with respect to the KF) to both unmodeled measurement biases and misspecified noise covariances, with the exception of $\Delta P_{xx}/\Delta m_\nu$. For the KF case, the covariance update is independent of the measurement and, as such, is insensitive to uncharacterized biases in the model. This comes at the expense of increased posterior mean sensitivity to m_ν.

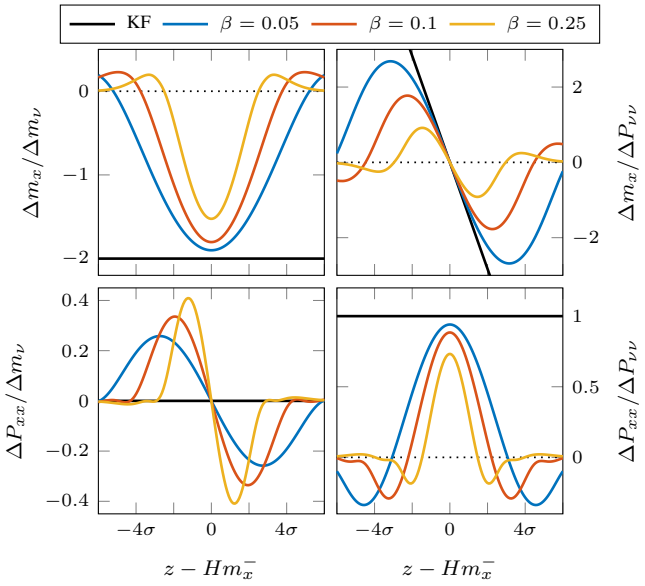


Fig. 3: Sensitivity values for the state mean, m_x, and covariance, P_{xx}, with respect to the likelihood mean, m_ν, and covariance, $P_{\nu\nu}$, as a function of the measurement residual.

For the multivariate parameterizations, it is convenient to analyze a single sensitivity "volume," given by the product of the singular values of $\boldsymbol{S}_{(\cdot)}$, as in Fig. 4. The proposed GVI update achieves improved sensitivity volumes across the entire span of residuals (with respect to the KF).

III. Variational Gradient Descent

Once a choice of loss and divergence functionals has been made, the challenge of minimizing (5) can be undertaken. For a given formulation of the generalized VFE, a target $\mathscr{F}$-optimal posterior is implicitly defined, which, ideally, can be sampled or represented as a parametric distribution. Unfortunately, the first variation of (5) often does not beget analytical minima. In lieu of closed-form solutions to (3), analogous to traditional VI, the $\mathscr{F}$-divergence is used to minimize dissimilarities between the approximating variational

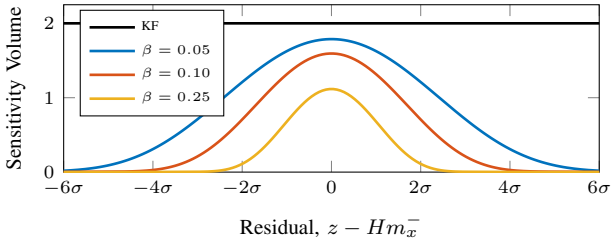


Fig. 4: Linear-Gaussian sensitivity volumes.

posterior, $q^* \in \mathcal{Q}$, and the globally-optimal pdf. Analogous to function minimization, convexity of $\mathscr{F}$ is leveraged to define a gradient-based solution method. As the following approach is an extension of traditional gradient descent to functional space, it is referred to as variational gradient descent (VGD). For ease of exposition, assume no model discrepancies exist, i.e., $p(\boldsymbol{z}|\boldsymbol{x}) = \pi(\boldsymbol{x}, \boldsymbol{z})$, etc.

First, a new pseudo-variable, τ, is introduced, often referred to as the homotopy parameter or pseudo-time. As motivated by [21], assume that through τ the state evolves according to the ordinary differential equation $\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}\tau} = \boldsymbol{\phi}(\boldsymbol{x})$, such that an artificial pseudo-time rate-of-change of the variational density is induced via the Liouville equation [22]

$$\frac{\partial q(\boldsymbol{x};\tau)}{\partial \tau} = -\nabla_{\boldsymbol{x}} \cdot \big(q(\boldsymbol{x};\tau)\boldsymbol{\phi}(\boldsymbol{x})\big) \,, \tag{11}$$

where $\boldsymbol{\phi}$ is referred to as the pseudo-dynamics and $\nabla_{\boldsymbol{x}} \cdot$ is the divergence operator. As in traditional gradient descent, the objective is to evolve the variational pdf in the direction of "steepest descent" of the free energy, and subsequently $D_{\mathscr{F}}$,

$$\boldsymbol{\phi}^*(\boldsymbol{x}) = \underset{\boldsymbol{\phi}}{\operatorname{argmin}} \left\{ \frac{\mathrm{d}\mathscr{F}[q(\boldsymbol{x};\tau)]}{\mathrm{d}\tau} \right\} \,. \tag{12}$$

Assuming $\mathscr{F}$ takes the form of (9), the pseudo-time rate-of-change for the free energy can be written as

$$\frac{\mathrm{d}\mathscr{F}}{\mathrm{d}\tau} = \int_{\mathbb{X}} q(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{x}) \cdot \nabla_{\boldsymbol{x}} \left[\ell[p(\boldsymbol{z}|\boldsymbol{x})] + g'(\omega) \left(\frac{\delta G}{\delta q} \right) \right] \mathrm{d}\boldsymbol{x} \,,$$

where $\omega = \int_{\mathbb{X}} G[q,p]\mathrm{d}\boldsymbol{x}$ and $\nabla_{\boldsymbol{x}}$ is the gradient operator (dependence of q on τ is suppressed for compactness). As shown in Appendix A, (12) can be solved by pseudo-dynamics functions of the form

$$\boldsymbol{\phi}^*(\boldsymbol{x}) \propto -\nabla_{\boldsymbol{x}} \left[\ell[p(\boldsymbol{z}|\boldsymbol{x})] + g'(\omega) \left(\frac{\delta G}{\delta q} \right) \right] \,. \tag{13}$$

As with conventional gradient descent, only the steepest descent direction is of interest, such that proportionality in (13) is sufficient.

The optimal pseudo-dynamics, $\boldsymbol{\phi}^*$, may then be heuristically viewed as evolving the variation pdf along the "path of least resistance" (or "principle of least action") for the given free energy, as in traditional mechanics. Subsequently, VGD is closely tied to the field of optimal transport [11] and can readily be shown to be a gradient flow of $\mathscr{F}$ in the space of the 2-Wasserstein metric on $\mathcal{P}(\mathbb{X})$ (c.f. [23]). If a

solution to the partial differential equation in (11), or discrete optimal transport scheme proposed in [10], exists, the globally-optimal pdf $\bar{q}$ is recovered in the limit as $\tau \to \infty$. In reality, both approaches are generally intractable. The alternative is to select a tractable variational family $q \in \mathcal{Q}(\mathbb{X})$, as in (3), and approximate (11), analogous to the prediction stage of filtering in physical time. The pseudo-dynamics then induce a rate-of-change for the underlying parameterization, e.g., mean and covariance, with respect to τ that can be integrated until convergence, yielding $q^*(\boldsymbol{x}; \hat{\boldsymbol{\theta}}) = q(\boldsymbol{x}; \boldsymbol{\theta}, \tau \to \infty)$.

It is important to note that parametric implementations of (13) often result in stiff, numerically inefficient systems. Convergence stability can be improved by pre-multiplying the pseudo-dynamics by a positive-definite operator, such as the Hessian of $\mathscr{F}$, without sacrificing convergence guarantees [24]. When the free energy Hessian is used, VGD is analogous to conventional Newton-based gradient descent. Convergence of parametric implementations is established by evaluating the change between discrete propagation steps, τ_k and τ_{k+1}, of either the free energy, $\mathscr{F}[q(\boldsymbol{x}; \boldsymbol{\theta}(\tau_k))] - \mathscr{F}[q(\boldsymbol{x}; \boldsymbol{\theta}(\tau_{k+1}))] <$ tolerance, or the parameterization, $|\boldsymbol{\theta}(\tau_{k+1}) - \boldsymbol{\theta}(\tau_k)| <$ tolerance.

A. Particle VGD, SVGD, and MCMC

The simplest parameterization to consider is the particle ensemble, wherein the variational density is approximated by a series of discrete points. The particles are then easily propagated through the pseudo-dynamics according to $\mathrm{d}\boldsymbol{x}_i/\mathrm{d}\tau = \boldsymbol{\phi}^*(\boldsymbol{x}_i)$, where $\boldsymbol{x}_i$ is the i^{th} particle of the ensemble. However, the derivation of (13) often necessitates the variational density be differentiable over the state space for many formulations of the free energy, limiting exact particle flow solutions. This restriction can be relaxed by fitting a continuous density to the particle ensemble at each evaluation of $\boldsymbol{\phi}^*$, e.g., clustering [25], kernel mean embedding [26], or moment matching. When the KL divergence is used in (5), Stein VGD (SVGD) [27] can be used to avoid continuous density approximations [28]. Additionally, free energies utilizing the KL divergence have a known globally-optimal solution, given by the Gibbs posterior [10], that is tractable up to a normalizing constant, facilitating Markov chain Monte Carlo (MCMC) methods [29].

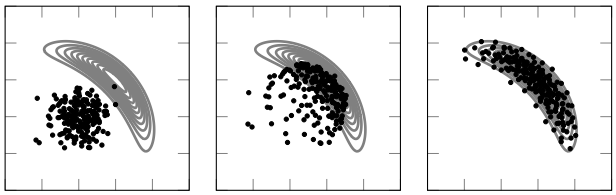
The process of particle-based VGD, visualized in Fig. 5, results in a posterior ensemble that mimics (or is equivalent to in the MCMC case) independent, identically distributed samples from the globally-optimal posterior [24], [30].

B. Gaussian VGD

Perhaps the most prolific and useful approximating density is the multivariate Gaussian pdf. Gaussians are convenient for both their ease of parameterization and link to the central limit theorem. For Gaussian VGD, approximating (11) is given by moment matching [1]

$$\frac{\mathrm{d}\boldsymbol{m}_x}{\mathrm{d}\tau} = \mathbb{E}_{q(\boldsymbol{x};\tau)}\left\{\boldsymbol{\phi}(\boldsymbol{x})\right\} \tag{14a}$$

$$\frac{\mathrm{d}\boldsymbol{P}_{xx}}{\mathrm{d}\tau} = \mathbb{E}_{q(\boldsymbol{x};\tau)}\left\{(\boldsymbol{\phi}(\boldsymbol{x}) - \dot{\boldsymbol{m}}_x(\tau))(\boldsymbol{x} - \boldsymbol{m}_x(\tau))^T\right\}$$

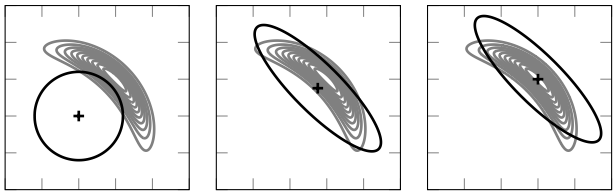


(a) initial (b) intermediate (c) converged

Fig. 5: Illustration of particle VGD (SVGD) with the optimal target pdf plotted as gray-scale contours.

$$+ \mathbb{E}_{q(\boldsymbol{x};\tau)}\left\{(\boldsymbol{x} - \boldsymbol{m}_x(\tau))(\boldsymbol{\phi}(\boldsymbol{x}) - \dot{\boldsymbol{m}}_x(\tau))^T\right\} \ , \tag{14b}$$

where $\dot{\boldsymbol{m}}_x = \mathrm{d}\boldsymbol{m}_x/\mathrm{d}\tau$ and $q(\boldsymbol{x};\tau) = p_g\left(\boldsymbol{x}; \boldsymbol{m}_x(\tau), \boldsymbol{P}_{xx}(\tau)\right)$. The implementation of (14) is no different than the prediction stage of Gaussian filters and can be approximated in numerous fashions, including a first-order Taylor series expansion (à la the extended KF), the unscented transform [5], or by quadrature [6]. For the Taylor-series approach, VGD has the added benefit of asymptotically approaching the posterior mean, continuously improving linearizations. Such an implementation for propagating the mean and covariance is depicted in Fig. 6. Once the moment propagation converges, the GVI posterior is taken as $q^*(\boldsymbol{x}) = \lim_{\tau\to\infty} p_g\left(\boldsymbol{x}; \boldsymbol{m}_x(\tau), \boldsymbol{P}_{xx}(\tau)\right)$.

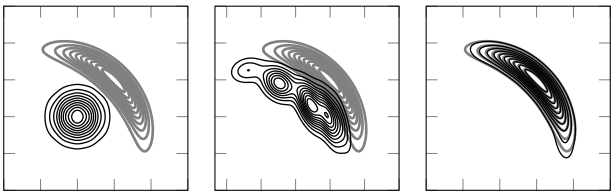


(a) initial (b) intermediate (c) converged

Fig. 6: Illustration of Gaussian VGD as the variational mean ($+$) and $\pm 3\sigma$ (standard deviation) contour (—) with the optimal target pdf plotted as gray-scale contours.

C. Gaussian Mixture (GM) VGD

The convenience of Gaussian approximations comes with the innate sacrifice of capturing complex target densities, as in Fig. 6. A computationally efficient means of improving the diversity of posteriors captured by VGD is through the use of GM approximations. The same processes used in GM filtering for nonlinear dynamic systems and their many algorithmic adaptations (cf [31], [32]) can be leveraged. The simplest method for propagating the means and covariances is by leveraging the Gaussian solution, where the i^{th} mean and covariance rates of change are given by evaluating the expectations in (14) over the i^{th} mixture component while holding the weights constant. Such an implementation is illustrated in Fig. 7.



(a) initial (b) intermediate (c) converged

Fig. 7: Illustration of GM VGD with the variational and optimal target pdfs plotted as black and gray-scale contours, respectively.

IV. Robust Filtering Example

To exemplify the proposed GVI framework, the following dynamic state estimation scenario is considered [33]. Let the state be comprised of the altitude, h, velocity, $\dot{h}$, and the ballistic coefficient, c, of an object in free fall, such that $\boldsymbol{x}^T = \begin{bmatrix} h & \dot{h} & c \end{bmatrix}$. The initial state elements are uncorrelated and Gaussian distributed with means and variances

$$\begin{aligned} m_{h,0} &= 10^5 \text{ [ft]} & P_{hh,0} &= 500 \text{ [ft}^2\text{]} \\ m_{\dot{h},0} &= -6 \times 10^3 \text{ [ft/s]} & P_{\dot{h}\dot{h},0} &= 2 \times 10^4 \text{ [ft}^2\text{/s}^2\text{]} \\ m_{c,0} &= 2 \times 10^3 \text{ [lb/ft}^2\text{]} & P_{cc,0} &= 9 \times 10^4 \text{ [lb}^2\text{/ft}^4\text{]} \ . \end{aligned}$$

The state evolves (in physical time) according to a standard atmospheric flight model, where $\dot{\boldsymbol{x}}^T = \begin{bmatrix} \dot{h} & \frac{1}{2}\rho(h)c^{-1}\dot{h}^2 - g & 0 \end{bmatrix}$, $\rho(h) = \rho_0 e^{-h/k_\rho}$, $\rho_0 = 3.4\times10^{-3}$ slug/ft^3, $k_\rho = 22 \times 10^3$ ft, and $g = 32.2$ ft/s^2. Range measurements are taken at 10 Hz from an observer $L = 15 \times 10^3$ ft downrange, such that $z_k = \sqrt{h_k^2 + L^2} + \nu_k$, where $\nu_k \sim p_c(\nu_k; 0, 10)$ is zero-mode Cauchy distributed white noise with scale parameter 10. Note that noise values are "zero-mode" as a Cauchy density possesses no finite statistical moments (beyond the 0^{th} moment). This property, stemming from its extreme "heavy-tailed" nature, makes Cauchy densities particularly interesting for applications in robust estimation. The true likelihood function is then given by $p(z_k|\boldsymbol{x}_k) = p_c(z_k; \sqrt{h_k^2 + L^2}, 10)$; however, assume the estimator is only privy to a misspecified Gaussian noise model, with nominal likelihood of $p_g(z_k; \sqrt{h_k^2 + L^2}, 10^2)$.

The described scenario is simulated over 15 seconds of motion for two filters, both assuming a Gaussian state pdf, with the resulting performance plotted in Figs. 8 and 9 (let $\delta h_k = h_k - m_{h,k}$, etc.). The first filter, Fig. 8, is an unscented Kalman filter (UKF), or linearized-Gaussian Bayes' filter, with a residual-editing scheme to reject measurements whose probabilities are less than 99.99%, with respect to the nominal Gaussian likelihood [7]. The second filter, Fig. 9, plots the performance of a GVI filter employing a β-loss function ($\beta = 0.01$) and KL divergence. An additional scaling term of 0.25 is included in the loss to match confidence intervals (i.e., $\ell = 0.25\ell_\beta$). Performance is analyzed in a 1,000 trial Monte Carlo (MC) analysis, where results are displayed as individual trial errors, mean (across all trials) filter errors, average filter $\pm3\sigma$ (standard deviation) confidence intervals, and Monte Carlo $\pm3\sigma$ confidence intervals.

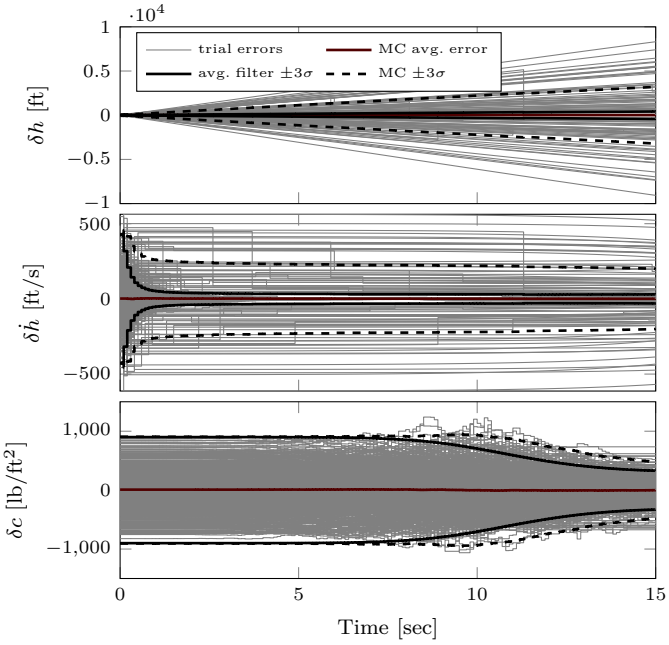


Fig. 8: Residual-edited UKF MC performance.

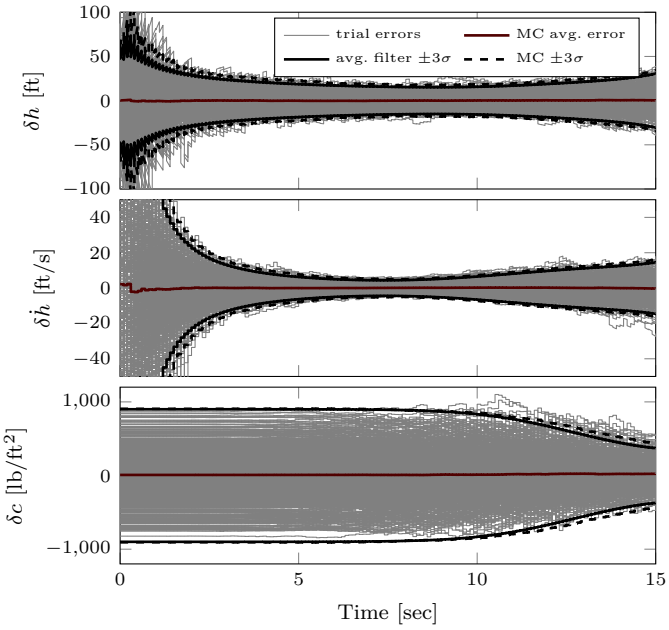


Fig. 9: Gaussian GVI-β-loss filter MC performance.

Ideal performance is an estimator that is unbiased (zero-mean error across all trials) and statistically consistent (the average filter $\pm3\sigma$ aligns with the experimental MC variance). As demonstrated in Fig. 9, the proposed GVI filter is able to accomplish both without access to the true noise model. The UKF, while unbiased, is unable to converge, with a significant number of trials outside the prescribed $\pm3\sigma$ confidence intervals in Fig. 8. This is due to sampling outlying

measurements before convergence, leading to measurements that are continuously rejected. Additionally, the UKF is significantly overconfident, i.e., the average filter $\pm 3\sigma$ interval is substantially less than that observed across all MC trials.

V. Conclusion

A significant hurdle to applied Bayesian inference is overcoming inaccuracies in the underlying statistical models. One approach that has demonstrated early success is generalized variational inference (GVI), which generalizes the popular variational Bayesian inference scheme to better accommodate modeling errors in the prior and likelihood probability densities. A novel quantitative sensitivity analysis was derived for GVI in addition to a general minimization procedure. The approach was demonstrated to overcome measurement modeling errors in a 1,000 trial Monte Carlo analysis for a dynamic estimation scenario.

Appendix

A. Optimal Pseudo-Dynamics

Starting from the VFE form in (9) and differentiating with respect to pseudo-time,

$$\frac{\mathrm{d}\mathscr{F}}{\mathrm{d}\tau} = \frac{\mathrm{d}}{\mathrm{d}\tau}\left\{\int_{\mathbb{X}} q\ell[p(\boldsymbol{z}|\boldsymbol{x})]\mathrm{d}\boldsymbol{x} + g\left(\int_{\mathbb{X}} G[q,p]\mathrm{d}\boldsymbol{x}\right)\right\} ,$$

and moving the differential inside the integral using the chain and Leibniz rules,

$$\frac{\mathrm{d}\mathscr{F}}{\mathrm{d}\tau} = \int_{\mathbb{X}}\left[\left(\frac{\partial q}{\partial \tau}\right)\ell[p(\boldsymbol{z}|\boldsymbol{x})] + g'(\omega)\left(\frac{\delta G}{\delta q}\right)\right]\mathrm{d}\boldsymbol{x} . \quad (15)$$

Substituting (11) into (15) results in

$$\frac{\mathrm{d}\mathscr{F}}{\mathrm{d}\tau} = -\int_{\mathbb{X}}\left[\nabla_{\boldsymbol{x}}\cdot\big(q\boldsymbol{\phi}\big)\ell[p(\boldsymbol{z}|\boldsymbol{x})] + g'(\xi)\left(\frac{\delta G}{\delta q}\right)\right]\mathrm{d}\boldsymbol{x} .$$

Assuming q is properly limited to zero at the boundaries of the state space, the divergence theorem may be applied to commute the gradient, yielding

$$\frac{\mathrm{d}\mathscr{F}}{\mathrm{d}\tau} = \int_{\mathbb{X}} q\boldsymbol{\phi}\cdot\nabla_{\boldsymbol{x}}\left[\ell[p(\boldsymbol{z}|\boldsymbol{x})] + g'(\xi)\left(\frac{\delta G}{\delta q}\right)\right]\mathrm{d}\boldsymbol{x} . \quad (16)$$

Equation (16) is a weighted (via q) L_2 inner product that is minimized when the two vectors are anti-colinear, as in (13).

References

[1] P. S. Maybeck, *Stochastic Models, Estimation, and Control*. Academic Press, 1979, vol. 1.

[2] J. Knoblauch, J. Jewson, and T. Damoulas, "An optimization-centric view on Bayes' rule: Reviewing and generalizing variational inference," *Journal of Machine Learning Research*, vol. 23, no. 132, pp. 1–109, 2022. [Online]. Available: http://jmlr.org/papers/v23/19-1047.html

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

[4] 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.

[5] S. Julier and J. Uhlmann, "Unscented filtering and nonlinear estimation," *Proceedings of the IEEE*, vol. 92, no. 3, pp. 401–422, 2004.

[6] I. Arasaratnam and S. Haykin, "Square-root quadrature Kalman filtering," *IEEE Transactions on Signal Processing*, vol. 56, no. 6, pp. 2589–2593, 2008.

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

[8] 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.

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

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

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

[12] A. Zellner, "Optimal information processing and Bayes's theorem," *The American Statistician*, vol. 42, no. 4, pp. 278–280, 1988.

[13] S. Gottwald and D. A. Braun, "The two kinds of free energy and the Bayesian revolution," *PLOS Computational Biology*, vol. 16, no. 12, pp. 1–32, 12 2020.

[14] S. Amari, *Information Geometry and Its Applications*. Springer, 2016.

[15] A. Basu, I. R. Harris, N. L. Hjort, and M. C. Jones, "Robust and efficient estimation by minimising a density power divergence," *Biometrika*, vol. 85, no. 3, pp. 549–559, 1998.

[16] 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.

[17] C. C. Holmes and S. G. Walker, "Assigning a value to a power likelihood in a general Bayesian model," *Biometrika*, vol. 104, no. 2, pp. 497–503, 2017.

[18] F. Futami, I. Sato, and M. Sugiyama, "Variational inference based on robust divergences," in *Proceedings of the 21st International Conference on Artificial Inteligence and Statistics (AISTATS)*, vol. 84, 2018.

[19] P. J. Huber and E. M. Ronchetti, *Robust Statistics*, 2nd ed. Wiley, 2009.

[20] J. Knoblauch, "Frequentist consistency of generalized variational inference," 2019, arXiv:1912.04946v1.

[21] 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. SPIE, 2009, pp. 733 603 1–12.

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

[23] F. Santambrogio, *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Birkhäuser Cham, 2015, ch. 8.

[24] A. N. Subrahmanya, A. A. Popov, and A. Sandu, "An ensemble variational Fokker-Planck method for data assimilation," Computational Science Laboratory, Department of Computer Science, Virginia Tech., Tech. Rep. CSL-TR-21-10, Apr. 2023, arXiv:2111.13926v4.

[25] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[26] 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, p. 1–141, 2017.

[27] Q. Liu and D. Wang, "Stein variational gradient descent: A general purpose Bayesian inference algorithm," in *Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS)*, 2016.

[28] K. J. Craft and K. J. DeMars, "Stein variational gradient descent for non-Bayesian particle flow," in *26th International Conference on Information Fusion (FUSION)*, 2023, pp. 1–8.

[29] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, 2nd ed. Springer, 2004.

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

[31] K. J. DeMars, R. H. Bishop, and M. K. Jah, "Entropy-based approach for uncertainty propagation of nonlinear dynamical systems," *Journal of Guidance, Control, and Dynamics*, vol. 36, no. 4, pp. 1047–1057, 2013.

[32] G. Terejanu, P. Singla, T. Singh, and P. D. Scott, "Uncertainty propagation for nonlinear dynamic systems using Gaussian mixture models," *Journal of Guidance, Control, and Dynamics*, vol. 31, no. 6, pp. 1623–1633, 2008.

[33] A. Gelb *et al.*, *Applied Optimal Estimation*. The MIT Press, 1974, ch. 6, pp. 194–199.