Nonlinear Kalman Filtering with Semi-Parametric Biscay Distributions

Steven Reece

Abstract—The problem of nonlinear estimation is reexamined, and a new semi-parametric representation of uncertainty called the Biscay distribution is presented. The Biscay distribution is combined with the extended Kalman filter (EKF) and a new filtering paradigm called the Biscay distribution filter (BDF) is developed. The BDF is provably optimal for linear estimation and generalizes naturally to nonlinear estimation. Further, the BDF is of the same computational order of complexity as the EKF. The BDF is compared with the EKF through an application in re-entry vehicle tracking.

Index Terms—Kalman filter, multisensor data fusion, nonlinear estimation, target tracking.

I. INTRODUCTION

WELL-known sufficient statistic approaches to nonlinear model-based parameter estimation, such as the extended Kalman filter (EKF) [1], [2], endeavor to estimate the mean and covariance of a probability distribution. These statistics are assumed to hold some relationship with the actual true parameter value, namely, that they are unbiased and contain good “quality” information. An estimate \( \hat{x} \) of a nonrandom parameter \( x \) is said to be unbiased if \( E[\hat{x}] = x \), where \( x \) is the true value of the parameter [2]. The quality of the estimate is determined from estimates of the variance of \( \hat{x} \).

Often, parameter estimates are obtained indirectly by first measuring an observable parameter and then transforming estimates of this parameter into “inferred” parameter state spaces using measurement and process models. Nonlinear transformations rarely produce estimates of the true state that are unbiased and of predictable quality since the mean of the transformed distribution need not necessarily coincide with the transformed true state. Thus, bias is evident even for optimal infinite dimensional filters [3] when applied to nonlinear systems.

This paper introduces a novel approach to nonlinear parameter estimation that uses mean and variance statistics that are in a sense “orthogonal” to the parameter state space. The new approach assigns mean and variance statistics individually to each point in the state space. The state space can then be transformed while leaving the orthogonal statistics unaffected. We show that the new approach goes some way towards alleviating nonlinear parameter estimation problems. The new nonlinear estimator yields performance that is equivalent to the EKF for linear systems, yet generalizes elegantly to nonlinear systems with a computational cost of the same order as the EKF.

The key contributions of the paper are to
1) introduce the Biscay distribution, which has zero mean and unit variance;
2) show how to combine many such distributions to yield another Biscay distribution from which parameter estimates can be drawn;
3) show that estimate uncertainties from combined distributions are no greater than those of the prior distributions;
4) examine the effect of a nonlinear variable transformation \( y = f(x) \) on estimation;
5) present an application of the Biscay distribution filter on radar tracking of a missile;
6) show that the Biscay distribution filter (BDF) outperforms the EKF for minimal additional computational effort.

Section II introduces the Biscay distribution and demonstrates its properties for one-dimensional (1–D) nonlinear inference. Section III develops exact distribution fusion methods and develops an approximate fusion method that utilizes distribution “innovation.” Section IV extends the theory to multivariate inference, and Section V discusses efficient BDF implementations. Finally, Section VI demonstrates the differences in performance of the BDF and EKF with an example application: the re-entry vehicle tracking problem.

II. BISCAY DISTRIBUTION

We introduce a sufficient statistic \( I_x(x; Z) \colon \mathbb{R}^2 \to \mathbb{R} \) for estimating the true value \( x_T \) of a parameter \( x \), where \( Z \) is a random measurement of \( x_T \). \( I_x(x; Z) \) is called the Biscay distribution and is a random variable since the measurement \( Z \) is random. When the true state \( x_T \) coincides with the measurement mean \( x_T = E[Z] \), and \( \sigma_e^2 \) is the variance of \( Z \), then the appropriate form for \( I_x(x; Z) \) is the normalized random variable

\[
I_x(x; Z) \triangleq \frac{x - Z}{\sigma_e},
\]

(1)

In general, a Biscay statistic is designed to satisfy three conditions.

1) The parameter estimate \( \hat{x} \) coincides with a zero of the statistic [i.e., \( I_x(\hat{x}; Z) = 0 \)].
2) The Biscay distribution is unbiased. The statistic is deemed to be unbiased when \( E[I_x(x_T; Z)] = 0 \).
3) The variance of \( I_x(x_T; Z) \) is unity. We regard \( I_x(x_T; Z) \) as a distribution over \( x \), which is called the Biscay distribution. The Biscay distribution can take both
negative as well as positive values. Generally, the larger the absolute value of \( I_x(x; Z) \), the less likely the true state has value \( x \).

It is straightforward to show that for the mean estimator, \( \Pr[|I_x(x_T; Z)| \geq 2] = 0.05 \) when \( Z \) is normally distributed and using the Chebyshev inequality \( \Pr(|I_x(x_T; Z)| \geq C) \leq C^{-2} \) for all distributions of \( Z \). The quality of an estimate is often represented by a confidence region \( J_C \) around \( \hat{x} \) that contains the true state with a “confidence” \( \Pr(x_T \in J_C) \). Traditionally, the EKF \( J_{EKF} = (\hat{x} - C\sigma_x, \hat{x} + C\sigma_x) \) and \( \Pr(x_T \in J_{EKF}) = 0.95 \) when \( Z \) is normally distributed. In a similar manner, a Biscay distribution filter confidence region is defined to be

\[
J_C^{\text{BDF}} \Delta \{ x : |I_x(x; Z)| < C \}.
\]

Again, \( \Pr(x_T \in J_C^{\text{BDF}}) = 0.95 \) when \( Z \) is normally distributed.

**Theorem 1:** In general, when \( J_C^{\text{BDF}} = \{ x : |I_x(x; Z)| < C \} \), for any distribution of \( Z \)

\[
\Pr(x_T \in J_C^{\text{BDF}}) \geq 1 - \frac{1}{C^2}. \tag{2}
\]

**Proof:**

\[
\Pr(x_T \in J_C^{\text{BDF}}) = 1 - \Pr(x_T \notin J_C^{\text{BDF}}) = 1 - \Pr(|I_x(x_T; Z)| > C),
\]

By the Chebyshev inequality

\[
\Pr(|I_x(x_T; Z)| > C) \leq \frac{\text{E}[|I_x(x_T; Z)|^2]}{C^2}.
\]

Since \( \text{E}[I_x(x_T; Z)] = 0 \), then

\[
\Pr(x_T \in J_C^{\text{BDF}}) \geq 1 - \frac{\text{Var}[I_x(x_T; Z)]}{C^2} = 1 - \frac{1}{C^2}.
\]

The key idea of this paper is the application of Biscay distributions to nonlinear inference \( y = f(x) \). The Biscay distribution \( I_y(y_T; Z) \) over the image space \( y \) is derived by a simple mapping from \( I_x(x_T; Z) \) over the domain space. For a 1-D, possibly nonlinear, injective mapping \( y = f(x) \)

\[
I_y(y_T; Z) = I_y(f(x_T); Z) \equiv I_x(x_T; Z).
\]

It follows immediately that \( I_y(\cdot; Z) \) is unbiased whenever \( I_x(\cdot; Z) \) is unbiased. Further, \( \text{Var}(I_y(\cdot; Z)) = 1; \) therefore, \( I_y(\cdot; Z) \) is a Biscay distribution.

To illustrate Biscay distribution inference, Fig. 1 shows the Biscay distributions \( I_y(y_T; Z) \) and \( I_x(x_T; Z) \) for a measurement \( Z = 0.25 \) of \( x \) (with standard deviation \( \sigma_Z = 0.33 \)) and the inferred parameter \( y \), where \( y = x^3, x_T = 0.7, \) and \( y_T = 0.343 \). The figure shows the BDF and EKF confidence regions \( J_x \) for \( x \) and \( y \). We note that although \( J_{EKF}^{\text{BDF}} \) and \( J_{EKF}^{\text{BDF}} \) are identical and both contain the true parameter value \( x_T, y_T \in J_{EKF}^{\text{BDF}}, \) but \( y_T \notin J_{EKF}^{\text{BDF}} \). This example illustrates the following key property of the Biscay distribution.

**Remark 1:** When \( y = f(x) \) and \( f \) is injective, the image confidence region \( J_y(y_T, C) \) contains the true state \( y_T \) if and only if the domain confidence region \( J_x(x_T, C) \) contains the true state \( x_T \).

Thus, \( \Pr(y_T \in J_y(y_T, C)) = \Pr(x_T \in J_x(x_T, C)) \), and therefore, (3) guarantees a measure of \( y \) estimate quality when the quality of the \( x \) estimate is known. A similar remark cannot be made for the EKF when \( f \) is nonlinear.

When \( f \) is not subjective, then for those members of the image space \( y \) that are not values of some member of the domain space \( x \), \( I_y(y_T; \cdot) = \pm \infty \). When \( f(\cdot) \) is many-one, a unique “conservative” value is assigned to \( I_y(\cdot; \cdot) \) so that \( [I_y(y_T; \cdot)] \leq [I_x(x_T; \cdot)] \). The following is a realization of this rule, which was used to obtain the results in Section VI and is particularly useful when \( I_y(\cdot) \) is determined by finite sampling of \( I_x(\cdot) \). For all \( C \geq 0 \), let \( J_C \) be the \( C \) confidence region \( J_C = \{ y : y = f(x), |I_x(x_T; \cdot)| \leq C \} \). Then

\[
I_y(y_T; \cdot) = \begin{cases} -\min\{C; y = \inf J_C\} & \text{if } I_y(\cdot; \cdot) \leq C, \\
\min\{C; y = \sup J_C\} & \text{otherwise}. \end{cases}
\]

Although for a many-one mapping \( I_y(\cdot) \) is not guaranteed to be unbiased, in practice, filters that use Biscay distributions exhibit less bias than EKFs.

Sections III and IV extend the idea of the Biscay distribution to the development of the Biscay distribution filter (BDF).
Since the BDF and EKF cycles are very similar, it is perhaps useful to see how the BDF equations derived in these sections relate to the EKF. The following are the stages of the EKF cycle corresponding to those in [2] and are listed along with equivalent BDF equations, which we will derive in the following sections. The only difference between the BDF described here and the EKF is that state estimates are inferred from measurements prior to fusion. These extra steps are indicated by an asterisk.

1) \* Observation state estimate … (1).
2) \* Observation covariance … (10).
3) \* Inferred state from observation … (17) and (20).
4) \* Inferred covariance update … (20) and (26).
5) State prediction … (17) and (20).
6) Prediction covariance update … (20) and (26).
7) Updated state estimate … (11), (14) and (16).
8) Updated state covariance … (15).

III. OPTIMAL FUSION

We now consider the fusion of uncorrelated Biscay distributions, each of which offer an estimate for the true value of $x$. These Biscay distributions can originate from direct (noisy) measurements of $x_T$ or from Biscay distributions mapped onto $x$ by (3).

The general form for the fusion of two prior Biscay distributions $L_{\alpha}(x; \cdot)$ and $L_{\beta}(x; \cdot)$ is a weighted linear combination

$$I(x; \cdot) = k_\alpha L_{\alpha}(x; \cdot) + k_\beta L_{\beta}(x; \cdot)$$

(4)

where $k_\alpha$ and $k_\beta$ satisfy a condition to guarantee unit variance

$$k_\alpha^2 + k_\beta^2 + 2\rho_{\alpha\beta}k_\alpha k_\beta = 1$$

where $\rho_{\alpha\beta} = \text{Cov}[L_{\alpha}, L_{\beta}]$. Of course, $I(\cdot)$ is unbiased whenever both $L_{\alpha}(\cdot)$ and $L_{\beta}(\cdot)$ are unbiased.

Values for the weights $k_\alpha$ and $k_\beta$ are chosen to minimize (i.e., optimize) the width $W$ of the posterior confidence region. In the following, we determine weights that are optimal for linear prior Biscay distributions. These weights, although valid for nonlinear distributions, may not be optimal for nonlinear distributions.

For any two parameter values $x_1$ and $x_2$, a linear Biscay distribution and an arbitrary parameter value $x_0$:

$$I(x_1; \cdot) = I(x_0; \cdot) + (x_1 - x_0)P(x_0; \cdot)$$

(5)

$$I(x_2; \cdot) = I(x_0; \cdot) + (x_2 - x_0)P(x_0; \cdot)$$

(6)

where $P(\cdot)$ is the derivative of $I$ with respect to $x$. The confidence region $J_C(x_1, x_2)$ has a width $W = x_2 - x_1$, where $-I(x_1; \cdot) = I(x_2; \cdot) = C$ (see Fig. 1). Subtracting (6) from (5) and substituting $-I(x_1; \cdot) = I(x_2; \cdot) = C$

$$2C = I(x_2; \cdot) - I(x_1; \cdot) = (x_2 - x_1)P(x_0; \cdot) = WP(x_0).$$

Thus

$$W = \frac{2C}{P(x_0; \cdot)}.$$  

(7)

Differentiating (4), dividing throughout by $2C$, and substituting for $I'$ from (7)

$$W = \left(\frac{k_\alpha}{W_\alpha} + \frac{k_\beta}{W_\beta}\right)^{-1}.$$  

(8)

Substituting for $k_\beta$ from $k_\alpha^2 + k_\beta^2 + 2\rho_{\alpha\beta}k_\alpha k_\beta = 1$ and then differentiating by $k_\alpha$

$$\frac{dW}{dk_\alpha} \propto \left[ \frac{1}{W_\alpha} + \frac{1}{W_\beta} \left( -\rho_{\alpha\beta} \pm \frac{k_\alpha[p_{\alpha\beta} - 1]}{\sqrt{k_\alpha^2[\rho_{\alpha\beta} - 1]} + 1} \right) \right] = 0.$$  

Solving for $k_\alpha$ and then for $k_\beta$ using (8)

$$k_\alpha = \frac{W_\alpha - \rho_{\alpha\beta}W_\beta}{\sqrt{1 - \rho_{\alpha\beta}^2} X}, \quad k_\beta = \frac{W_\alpha - \rho_{\alpha\beta}W_\beta}{\sqrt{1 - \rho_{\alpha\beta}^2} X}$$

where $X = W_\alpha^2 + W_\beta^2 - 2\rho_{\alpha\beta}W_\alpha W_\beta$.

The following theorem guarantees that the uncertainty in the combined Biscay distribution is no greater than either of the prior distributions.

**Theorem 2:** When two estimates of width $W_\alpha$ and $W_\beta$ are combined using the BDF, then the width $W$ of the fused distribution obeys $W \leq \min\{W_\alpha, W_\beta\}$.

**Proof:** Without loss of generality, assume that $W_\alpha \leq W_\beta$. Then, by (8)

$$W = \frac{W_\alpha W_\beta}{k_\alpha W_\beta + k_\beta W_\alpha}$$

$$= W_\alpha \sqrt{\frac{W_\alpha^2(1 - \rho_{\alpha\beta}^2)}{W_\alpha^2 + W_\beta^2 - 2\rho_{\alpha\beta}W_\alpha W_\beta}}$$

$$= W_\alpha \sqrt{1 - \frac{(W_\alpha - \rho_{\alpha\beta}W_\beta)^2}{W_\alpha^2 + W_\beta^2 - 2\rho_{\alpha\beta}W_\alpha W_\beta}}$$

$$\leq W_\alpha \quad \text{[since } |\rho_{\alpha\beta}| \leq 1\]$$

$$= \min\{W_\alpha, W_\beta\}. \quad \square$$

Substituting for $k_\alpha$ and $k_\beta$ in (8) and assuming uncorrelated measurements (i.e., $\rho_{\alpha\beta} = 0$)

$$W = \frac{W_\alpha W_\beta}{\sqrt{W_\alpha^2 + W_\beta^2}}.$$  

This equation for $W$ has a strong structural analogy with the 1-D EKF covariance update equation. Equating $W_\alpha$ and $W_\beta$ with measurement standard deviations $W_\alpha = \sigma_\alpha$ and $W_\beta = \sigma_\beta$ yields

$$\frac{1}{\sigma_\alpha^2} = \frac{1}{\sigma_\alpha^2} + \frac{1}{\sigma_\beta^2}.$$  

Thus, the linear limit of the BDF covariance update is equivalent to that of the 1-D EKF.

**A. Fusion Covariance Update**

Suppose two Biscay distributions, labeled $\alpha_1$ and $\beta_1$, are fused to produce a new distribution for parameter $\alpha_1$ and
similarly, distributions $\alpha_2$ and $\beta_2$ are fused to yield a new distribution for parameter $\beta$. Assume uncorrelated measurements (i.e., $\rho_{\alpha_1,\beta} = 0$) although this is not a necessary requirement. Then, by (4)

$$I_\alpha(x; \cdot) = \frac{W_\beta I_{\alpha_1}(x; \cdot) + W_\alpha I_{\beta_1}(x; \cdot)}{\sqrt{W_\alpha^2 + W_\beta^2}}.$$  

$$I_\beta(x; \cdot) = \frac{W_\beta I_{\alpha_2}(x; \cdot) + W_\alpha I_{\beta_2}(x; \cdot)}{\sqrt{W_\alpha^2 + W_\beta^2}}.$$  

The covariance of $I_\alpha(x; \cdot)$ with $I_\beta(x; \cdot)$ is then

$$\rho_{\alpha\beta} = \frac{\text{Cov}[Z_\alpha, Z_\beta]}{\text{Var}[Z_\alpha] \text{Var}[Z_\beta]}.$$ (9)

The BDF covariance matrix can be initialized from the measurement statistics. By (4)

$$\rho_{\alpha\beta} = \frac{\text{Cov}[Z_\alpha, Z_\beta]}{\text{Var}[Z_\alpha] \text{Var}[Z_\beta]}.$$ (10)

B. Incorporating Innovation

When $I_{x_\alpha}(x; \cdot)$ and $I_{x_\beta}(x; \cdot)$ are two Biscay distributions over parameter space $x$ with confidence region widths $W_{x_\alpha}$ and $W_{x_\beta}$, respectively, the innovation $n_x$ for parameter $x$ is defined to be

$$n_x = W_{x_\beta} I_{x_\beta}(x; \cdot) - W_{x_\alpha} I_{x_\alpha}(x; \cdot).$$

When $I_{x_\alpha}(x; \cdot)$ and $I_{x_\beta}(x; \cdot)$ are linear and mean unbiased and yield estimates $\hat{x}_\alpha$ and $\hat{x}_\beta$, respectively [where $I_{x_\alpha}(\hat{x}_\alpha; \cdot) = 0$ and $I_{x_\beta}(\hat{x}_\beta; \cdot) = 0$], then $n_x = \hat{x}_\beta - \hat{x}_\alpha$ is the familiar EKF innovation. The innovation is independent of the unknown value $x_T$ for linear distributions. Extending (4) to include innovations

$$I_x(x; \cdot) = K_{x_\alpha} I_{x_\alpha}(x; \cdot) + K_{x_\beta} I_{x_\beta}(x; \cdot) + \sum_{y \neq x} B_{xy} k_{xy} y$$ (11)

where $B_{xy}$ are constants chosen to minimize the confidence region widths of $I_x(\cdot)$. In order to find optimal values for $B_{xy}$ and at the same time expedite the relationship between the EKF and the innovation form of the BDF, we transform (11) into a more familiar form. Substituting (without loss of generality)

$$K_{x_\alpha} = \frac{W_{x_\alpha}}{W_{x_T}}(1 - k_{x_\alpha}), K_{x_\beta} = \frac{k_{x_\alpha}}{W_{x_\beta}},$$

and

$$B_{xy} = k_{xy}/W_{x},$$

we obtain

$$I_x(x; \cdot) = \frac{1}{W_{x}} \left( W_{x_\alpha} I_{x_\alpha}(x; \cdot) + \sum_y k_{xy} y \right).$$ (12)

We note that when the prior distributions are linear, then $I_x(x; \cdot)$ is also linear and has a confidence region width $W_x$, where

$$W_x = \sqrt{\text{Var} \left[ W_{x_\alpha} I_{x_\alpha}(x; \cdot) + \sum_y k_{xy} y \right]}.$$  

Now, define $X_\alpha$, $X_\beta$, $X$, and $k$

$$X_\alpha = \begin{pmatrix} W_{x_\alpha} I_{x_\alpha}(x_T) \\ W_{y_\alpha} I_{y_\alpha}(y_T) \\ \vdots \\ W_{x_\beta} I_{x_\beta}(x_T) \end{pmatrix}, \quad X_\beta = \begin{pmatrix} W_{x_\beta} I_{x_\beta}(x_T) \\ W_{y_\beta} I_{y_\beta}(y_T) \\ \vdots \\ W_{x_\beta} I_{x_\beta}(x_T) \end{pmatrix}$$

$$k = \begin{pmatrix} k_{x_\alpha} & k_{x_\beta} & \cdots & k_{x_\alpha} \\ k_{y_\alpha} & k_{y_\beta} & \cdots & k_{y_\beta} \\ \vdots & \vdots & \ddots & \vdots \\ k_{x_\alpha} & k_{x_\beta} & \cdots & k_{x_\beta} \end{pmatrix}, \quad X = \begin{pmatrix} W_{x_\alpha} I_{x_\alpha}(x_T) \\ W_{y_\beta} I_{y_\beta}(y_T) \\ \vdots \\ W_{x_\alpha} I_{x_\alpha}(x_T) \end{pmatrix}$$

and

$$\nu = X_\beta - X_\alpha.$$  

Then, (12) becomes

$$X = X_\alpha + k \nu.$$ (13)

We will determine optimal values for $k$ that minimize the sum of the squares of the posterior confidence region widths $\sum_i W_i^2$. Since Biscay distributions have unit variance, then

$$\sum_i W_i^2 = \text{trace} \left[ \text{Cov}[X, X^T] \right].$$

Equation (13) is in the form of the Kalman filter with updated state estimate $X$ and filter gain $k$. Further, the EKF gain is always chosen to minimize the trace of the updated state covariance matrix, which is equivalent to the BDF optimality condition to minimize $\sum_i W_i^2$. Therefore, we can adapt the EKF method for calculating gain directly to the BDF.

Let $P = \text{Cov}[X, X^T]$ and $P_\alpha = \text{Cov}[X_\alpha, X_\alpha^T]$ be $n \times n$ covariance matrices over $n$ parameters and $P_\beta = \text{Cov}[X_\beta, X_\beta^T]$ be an $m \times m$ covariance matrix over $m$ parameters. The prior Biscay distributions are assumed to be over state variables that may include distributions transformed from observation spaces. Let $H$ be an $m \times n$ matrix such that

$$[H]_{x,y} = \begin{cases} 1, & \text{if } I_{x_\alpha} \text{ and } I_{y_\beta} \text{ are distributions over the same parameter space} \\ 0, & \text{otherwise} \end{cases}$$

Then, with reference to [2]

$$k = P_\alpha H^T (P_\alpha + P_\beta)^{-1}$$ (14)

and

$$P = P_\alpha - P_\alpha (P_\alpha + P_\beta)^{-1} P_\alpha.$$ (15)

Of course, when it comes to measuring the innovation $\nu_x$, in practice, $x_T$ is unknown. When $I_{x_\alpha}(\cdot)$ and $I_{x_\beta}(\cdot)$ are linear, then the innovation is independent of $x_T$. However, when the distributions are nonlinear, then an approximation $\tilde{x}$ for $x_T$ can be used instead. Suppose $W_{x_\alpha} < W_{x_\beta}$; then, the best estimate for $x_T$ satisfies $I_{x_\alpha}(\tilde{x}) = 0$. The innovation is then

$$\nu_x = W_{x_\beta} I_{x_\beta}(\tilde{x}; \cdot).$$ (16)

After obtaining values for $k$ and $\nu$, the Biscay distributions are combined using (11). Distributions over the same parameter
space, \( x \), say, are combined by weighted averaging at each parameter space value as described at the beginning of Section III [but this time using \( k \) values calculated using (14)]. This will yield a new distribution over the parameter space. This new distribution is then uniformly adjusted by adding \( \sum_{y \in V} B y \).

IV. MULTIVARIATE INFERENCE

Consider the mapping \( f : \mathbb{R}^n \rightarrow \mathbb{R} \)

\[
y = f(x_1, \ldots, x_n).
\]

The Biscay distribution for \( y \) is obtained from the prior distributions \( I_i(\cdot) \) for the parameters \( x_i \)

\[
I(y_{\cdot}, \cdot) \triangleq \frac{1}{\sigma} \sum_{i=1}^{n} S_i I_i(x_{i\cdot}, \cdot) \tag{17}
\]

where \( \sigma^2 = \sum_{i,j} s_{ij} \sigma_i \sigma_j \) and \( s_{ij} \in \mathbb{R} \) are constants. \( I(y_{\cdot}, \cdot) \) is clearly unbiased when each of \( I_i(\cdot) \) are unbiased. In addition, \( \text{Var}(I(y_{\cdot}, \cdot)) = 1 \); therefore, \( I(y_{\cdot}, \cdot) \) is a Biscay distribution.

Values for \( S_i \) are chosen so that (17) is injective when \( f(\cdot) \) is linear. The Appendix shows that this choice of \( S_i \) also minimizes the confidence region widths in image space \( y \) when \( f(\cdot) \) is linear. Expanding \( f(\cdot) \) about the estimate \( \hat{x}_i \) [where \( I(\hat{x}_i, \cdot) = 0 \)]

\[
y \approx f(\hat{x}_1, \ldots, \hat{x}_n) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} [x_i - \hat{x}_i]. \tag{18}
\]

Assuming near linear Biscay distributions, \( I_i(x_i) \approx (dU_i/dx_i) \delta_i [x_i - \hat{x}_i] \) and

\[
I(y_{\cdot}, \cdot) = \frac{1}{\sigma} \sum_{i=1}^{n} \frac{\partial L_i}{\partial x_i} [x_i - \hat{x}_i]. \tag{19}
\]

Comparing (18) and (19), we see that a linear model applied to linear Biscay distributions is injective when

\[
S_i = \frac{\partial f}{\partial x_i} / \frac{\partial L_i}{\partial x_i}, \tag{20}
\]

since, for this \( S_i \)

\[
I(y_{\cdot}, \cdot) = \frac{1}{\sigma} [y - f(\hat{x}_1, \ldots, \hat{x}_n)] \tag{21}
\]

which is independent of \( x_1, \ldots, x_n \).

When \( f(\cdot) \) or the prior Biscay distributions are nonlinear, (20) still applies, although the resultant image space confidence region may not be minimal. When \( f(\cdot) \) is extremely nonlinear and is a many-one mapping, unique conservative values are assigned to \( I_i(\cdot) \), as described in Section II.

Example: This simple example illustrates the relationship between the BDF and traditional approaches to statistical inference for linear systems. Consider the stochastic process

\[
x(k+1) = x(k) + \omega \tag{22}
\]

where \( \omega \) is zero mean white noise with variance \( \sigma^2 \). For traditional methods, \( \hat{x}(k+1) = \hat{x}(k) \) and \( \sigma(k+1)^2 = \sigma(k)^2 + \sigma^2 \). The extent of the confidence region \( J^{\text{BDF}}(k+1) \) is \( J^{\text{BDF}}(k+1) = 2 \sqrt{\sigma(k)^2 + \sigma^2} \). The Biscay distributions for \( x(k) \) and \( \omega \) are

\[
I_k(x; \cdot) = \frac{x - \hat{x}(k)}{\sigma(k)} \quad \text{and} \quad I_\omega(\cdot) = \frac{\omega}{\sigma_\omega}, \tag{23}
\]

where \( I_k(x; \cdot) \) is 0. By (22), \( S_x = \sigma \) and \( S_\omega = \sigma_\omega \).

Hence, by (17)

\[
I_{k+1}(x + \omega; \cdot) = \frac{\sigma(k)I_k(x; \cdot) + \sigma_\omega I_\omega(\cdot)}{\sqrt{\sigma(k)^2 + \sigma_\omega^2}} = \frac{x - \hat{x}(k) + \omega}{\sqrt{\sigma(k)^2 + \sigma_\omega^2}}. \tag{24}
\]

Therefore

\[
I_{k+1}(x) = \frac{x - \hat{x}(k)}{\sqrt{\sigma(k)^2 + \sigma_\omega^2}}. \tag{25}
\]

The BDF confidence region \( J^{\text{BDF}}(k+1) = (x_t, x_u) \), and by (23)

\[-1 = \frac{x_t - \hat{x}(k)}{\sqrt{\sigma(k)^2 + \sigma_\omega^2}} \quad \text{and} \quad 1 = \frac{x_u - \hat{x}(k)}{\sqrt{\sigma(k)^2 + \sigma_\omega^2}}. \tag{26}\]

Therefore, \( J^{\text{BDF}}(k+1) = 2 \sqrt{\sigma(k)^2 + \sigma_\omega^2} \), which is identical to the traditional method \( J^{\text{EKF}}(k+1) \). Further, the best estimate \( \hat{x}(k+1) \) corresponds to \( I_{k+1}(\hat{x}(k+1); \cdot) = 0 \), and by (23), \( \hat{x}(k+1) = \hat{x}(k) \).

A. Prediction Covariance Update

The covariances between posterior Biscay distributions are obtained iteratively from the covariances between the prior distributions. Suppose the value of parameters \( \alpha \) and \( \beta \) are inferred from \( n_\alpha \) parameters with distributions \( I_{i, \alpha} \) \( i = 1 \) \( n_\alpha \) and \( n_\beta \) parameters with distributions \( I_{j, \beta} \) \( j = 1 \) \( n_\beta \). The posterior distributions are

\[
I_\alpha(x_{i\cdot}, \cdot) = \frac{\sum_{i=1}^{n_\alpha} S_i I_{i, \alpha} (x_{i\cdot}, \cdot)}{\sqrt{\sum_{i=1}^{n_\alpha} S_i I_{i, \alpha} S_{i, \alpha} \rho_{i, \alpha}}} \tag{24}
\]

\[
I_\beta(x_{j\cdot}, \cdot) = \frac{\sum_{j=1}^{n_\beta} S_j I_{j, \beta} (x_{j\cdot}, \cdot)}{\sqrt{\sum_{j=1}^{n_\beta} S_j I_{j, \beta} S_{j, \beta} \rho_{j, \beta}}} \tag{25}
\]

where \( \rho_{i, \alpha} = \text{Cov}[I_{i, \alpha}(\cdot), I_{i, \alpha}(\cdot)] \) and \( \rho_{j, \beta} = \text{Cov}[I_{j, \beta}(\cdot), I_{j, \beta}(\cdot)] \). Thus

\[
\rho_{\alpha, \beta} = \frac{\sum_{i=1}^{n_\alpha} S_i I_{i, \alpha} S_{i, \beta} \rho_{i, \alpha}}{\sqrt{\sum_{i=1}^{n_\alpha} S_i I_{i, \alpha} S_{i, \alpha}}} \frac{\sum_{j=1}^{n_\beta} S_j I_{j, \beta} S_{j, \beta} \rho_{j, \beta}}{\sqrt{\sum_{j=1}^{n_\beta} S_j I_{j, \beta} S_{j, \beta}}} \tag{26}\]

where \( \rho_{ij} = \text{Cov}[I_{i, \alpha}(\cdot), I_{j, \beta}(\cdot)] \).
V. EFFICIENT BDF IMPLEMENTATIONS

For reasons of efficient computation and representation, the Biscay distribution should be described at only a finite set of state space values. The full distribution is obtained by interpolation or extrapolation. Supposing that the Biscay distribution is known at values \( l_1, \ldots, l_m \), then for all \( x \) such that \( l_i \leq x \leq l_{i+1} \)

\[
I(x; \cdot) \approx \frac{(x - l_{i+1})I(l_{i+1}; \cdot) + (l_i - x)I(l_i; \cdot)}{l_i - l_{i+1}}.
\]

We have found that linear extrapolation is a reasonable approximation when \( x < l_1 \) or \( x > l_m \). When \( x < l_1 \) or \( x > l_m \), replace \( l_i \) with \( l_1 \) and \( l_{i+1} \) with \( l_m \) in (27).

Potentially, the most computationally intensive part of the BDF cycle is multivariate inference since, in general, Biscay values in the posterior space are inferred from every combination of prior Biscay samples. However, there is a subsampling scheme that works for many problems that are not severely nonlinear. Consider the mapping \( f: \mathbb{R}^n \rightarrow \mathbb{R} \). In Fig. 2, the Biscay contours and functional contours are everywhere parallel when both \( f(\cdot) \) and the prior distributions \( I_1(\cdot), \ldots, I_n(\cdot) \) are linear. Thus, for a linear system, a complete description of \( I_p(\cdot) \) can be obtained from those sample points along \( \nabla f \) (represented by \( \mathbf{d}_a \) in the figure). When \( f(\cdot) \) is nonlinear, the contours are near parallel in a neighborhood of the estimate \( \hat{x}_1, \ldots, \hat{x}_n \) and, provided the \( J_2 \) confidence regions of the priors are less than or of the same order of magnitude as the curvature of \( f(\cdot) \), the following computationally efficient sub-sampling scheme can be used:

\[
x = \hat{x} + \kappa \nabla f
\]

where \( \kappa \in \mathbb{R} \). This subsampling scheme was used to obtain the results in Section VI. Three sample points were used at \( \kappa = 0 \) and \( \kappa = \pm 2/(\nabla \mathbf{L} \nabla f) \).

VI. RE-ENTRY VEHICLE TRACKING PROBLEM

The re-entry vehicle tracking problem (which is also called the ballistic missile tracking problem) is a benchmark problem in statistical estimation [4]–[7] and illustrates the superior accuracy of the BDF over the EKF. It is the problem of accurately tracking an incoming projectile that is described by nonlinear physical dynamics. Tracking is initialized when the missile enters the atmosphere at high altitude and at a very high speed. Three types of force act on the projectile:

1) aerodynamic drag, which is a function of its speed and has a substantial nonlinear variation in altitude;
2) gravitational force, which accelerates it toward the center of the Earth;
3) random buffeting terms.

The effect of these forces gives a trajectory that is initially ballistic but as the density of the atmosphere increases, drag effects become important, and the missile rapidly decelerates until its motion is almost vertical. The missile is tracked by a radar that is able to measure range, elevation, and azimuth. From the first sighting, a typical trajectory lasts about 200 s. We assume that the radar measures the missile’s progress every 5 s.
with $\mu$ zero mean, white, and $\text{Var}[\mu] = 0.25$, and

$$G(k) = \frac{Gm_0}{H(k)}$$

where the characteristic atmospheric depth $H_0 = 13.406$ km, the gravitational constant Earth mass product $Gm_0 = 3.986 \times 10^5$ km$^3$s$^{-2}$, and the equatorial radius $R_0 = 6378$ km. Estimates of the range $r$, elevation $\theta$, and azimuth $\phi$ of the missile (see Fig. 3) relative to a radar located at $(r_1, r_2, r_3) = (6378, 0, 0)$ are

$$\dot{r}(k) = \sqrt{(X_1(k) - r_1)^2 + (X_2(k) - r_2)^2 + (X_3(k) - r_3)^2}$$

$$\dot{\theta}(k) = \arctan \left( \frac{X_2(k) - r_2}{\sqrt{(X_2(k) - r_2)^2 + (X_3(k) - r_3)^2}} \right) + \omega_2(k),$$

$$\dot{\phi}(k) = \arctan \left( \frac{X_3(k) - r_3}{X_2(k) - r_2} \right) + \omega_3(k).$$

The terms $\omega_1$, $\omega_2$, and $\omega_3$ are zero mean, uncorrelated noise with standard deviations of 4.5 m, 15 mrad, and 15 mrad, respectively. State estimates are obtained from the measurements using the following transformations:

$$\dot{X}_1(k) = R_0 + \dot{r}(k) \sin \dot{\theta}(k)$$

$$\dot{X}_2(k) = \dot{r}(k) \cos \dot{\theta}(k) \cos \dot{\phi}(k)$$

$$\dot{X}_3(k) = \dot{r}(k) \cos \dot{\theta}(k) \sin \dot{\phi}(k).$$

The initial true state is normally distributed about $(6500.4, 349.14, 100.3, -1.81, -6.80, -3.00)$ with standard deviations $\sigma = (10^{-3}, 10^{-3}, 10^{-3}, 0.0, 0.0, 0.0)$. The BDF represents the initial state by six Biscay distributions $I_i(X_0(0))$ with zeros at $E[X_i(0)]$ and confidence region widths $\sigma$. Throughout each run, the BDF represents all distributions at $I(\cdot) = -2$. $I(\cdot) = 0$, and $I(\cdot) = 2$. The initial EKF state estimate is $E[X_i(0)]$ with a diagonal covariance matrix with diagonal values $\sigma_d$.

Fig. 4 shows that the BDF can be considerably more consistent than the EKF. The figure shows the probability that the true parameter value is contained within the second confidence region $J_2$. The high-quality BDF estimate confidence regions contain the true parameter values about 90% of the time, whereas the EKF estimates of $X_1$ and $X_2$ degrade significantly over time. Data for 100 Monte Carlo trajectories.

VII. CONCLUSIONS

This paper presents a new filter paradigm called the Biscay distribution filter for nonlinear estimation. The new method is statistically optimal for linear estimation, but for nonlinear systems, it is more consistent than, but of the same order of computational complexity as, the extended Kalman filter. The advantages of the method were demonstrated through an application in re-entry vehicle tracking.

APPENDIX

Theorem 3: For any linear mapping $f$: $\mathbb{R}^n \rightarrow \mathbb{R}$ and linear prior Biscay distributions $I_i$ ($i = 1, \ldots, n$), the image space confidence region widths are minimized when $S_i = (\partial f/\partial x_i)_{x_i}(\partial L_i/\partial x_i)_{x_i}^{-1}$ in (17).

Proof: Define a hyperplane of dimension $n - 1$ in $x_1, \ldots, x_n$ such that $I(f(x_1, \ldots, x_n); \cdot) = v$ (call this hyperplane 1). By (21), all points in hyperplane 1 map onto the same value in the image space, and by (19) and (20), the hypernormals to hyperplane 1 are $(1/\sigma)(\partial f/\partial x_i)_{x_i}$. Now,
consider $S^*_k = \beta_k\frac{\partial f}{\partial x_i}x_i/dL_i/dx_i$, for some $\beta_i \in \mathbb{R}$. 

Rewriting (19)

$$I^*(f(x_1, \ldots, x_n); \cdot) = \frac{1}{\sigma^*} \sum_{i=1}^{n} \beta_i \frac{\partial f}{\partial x_i} [x_i - \bar{x_i}]$$  \hspace{1cm} (28)

where

$$\sigma^* = \sqrt{\sum_{i=1}^{n} \beta_i \beta_j \frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j} \beta_i \beta_j}.$$  

Our claim is that $\beta_k = 1$ minimizes the confidence region width in the image space. When some or all of $\beta_k \neq 1$, (28) describes a hyperplane (call this hyperplane2) with hypernormals $(\beta_i/\sigma^*)(\partial f/\partial x_i)\bar{x_i}$. Both hyperplane1 and hyperplane2 are parallel only when all the $\beta_k$s take the same value, in which case, (28) reduces to (19). When at least two $\beta_k$s are different, hyperplane2 intersects hyperplane1 along some hyperplane of dimension $n = 2$. Thus, for any $y$ in the image space, there is at least one point $x_1, \ldots, x_n$ in the domain space such that $I(y, \cdot) = I^*(y, \cdot)$. Since $J_y$ contains the image subspace with $[\partial f(y, \cdot)] \leq v$, then if $y$ is included in $J_y$, $y$ must also be included in $J_y^*$.  

Fig. 5. Comparison of estimate confidence region widths (dash-dot line), standard deviations of the estimate about the true parameter value (thick solid line), and bias (thin solid line) between the BDF and EKF. Although the confidence region widths are comparable between the filters, the BDF estimate bias and standard deviations are better than those of the EKF.

References


**Steven Reece** received the B.Sc. degree in physics and mathematics in 1989 and Part III of the Cambridge Mathematics Tripos and Diploma in computer science in 1990 and 1991, respectively. In 1994, he joined the Robotics Research Group, Oxford University, Oxford, U.K., and in 1998, received the D.Phil. degree after doing research under the supervision of Prof. J. M. Brady.

From 1991 to 1994, he was a research assistant with the Computer Laboratory, Cambridge University, Cambridge, U.K. He is currently a Research Fellow at Corpus Christi College at Oxford. His research interests include nonstandard representations of uncertainty, machine learning, concept formation, and adaptive sensor systems.