1 Probability Basics Review

1.1 Random Variables

Definition 1.1 (Random Variable)**.** We call $x \in \mathcal{X}$ a random variable (RV) from the set of possible outcomes $\mathcal{X}$, with an associated probability density function (PDF) $p_x : \mathcal{X} \to \mathbb{R}$ that satisfies

- $p_x(\bar{x}) \geq 0$ for all $\bar{x} \in \mathcal{X}$, and
- if $\mathcal{X}$is countable (discrete random variable, DRV)

$$\sum_{\bar{x}\in\mathcal{X}} p_x(\bar{x}) = 1 \quad \text{or} \quad \int_{\mathcal{X}} p_x(\bar{x})d\bar{x} = 1,$$

in the case of a continuous random variable (CRV) ($\mathcal{X}$ is an interval).

The PDF is then used to define the notion of probability, i.e. the probability that a discrete RV x takes the value $\bar{x} \in \mathcal{X}$ is $p_x(\bar{x})$, and is written as $\Pr\{x = \bar{x}\}$.

For a continuous RV x the probability of any specific value is always 0, instead we can only can only refer to the RV being in some interval $[a, b] \subseteq \mathcal{X}$ and we write $\Pr\{x \in [a, b]\} = \int_a^b p_x(\bar{x})\, d\bar{x}$.

Definition 1.2 (Joint PDF)**.** Let $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ be RVs. The joint PDF satisfies

- $p_{xy}(\bar{x}, \bar{y}) \geq 0$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$,
- further

$$\sum_{\bar{x}\in\mathcal{X}} \sum_{\bar{y}\in\mathcal{Y}} p_{xy}(\bar{x}, \bar{y}) = 1 \text{ or } \iint_{\mathcal{X}\times\mathcal{Y}} p_{xy}(\bar{x}, \bar{y})\, d\bar{x}d\bar{y} = 1$$

for DRVs or CRVs, respectively.

The interpretation of the joint PDF is that both x equals $\bar{x}$ and y equals $\bar{y}$.

Definition 1.3 (Marginalization)**.** From a joint PDF of x and y we can redefine the PDF of either variable by going through each case of the other:

$$p_x(\bar{x}) = \sum_{\bar{y}\in\mathcal{Y}} p_{xy}(\bar{x}, \bar{y}) \text{ or } p_x(\bar{x}) = \int_{\mathcal{Y}} p_{xy}(\bar{x}, \bar{y})\, d\bar{y}$$

and similar for $p_y(\bar{y})$.

Definition 1.4 (Conditioning)**.** Given the RVs x, y and $p_{xy}(\bar{x}, \bar{y})$ define

$$p_{x|y}(\bar{x}|\bar{y}) = \frac{p_{xy}(\bar{x}, \bar{y})}{p_y(\bar{y})} \iff p_{x|y}(\bar{x}|\bar{y})p_y(\bar{y}) = p_{xy}(\bar{x}, \bar{y})$$

when $p_y(\bar{y}) \neq 0$.

The $p_{x|y}(\bar{x}|\bar{y})$ is reads x given y ans is to be understood the probability of x while keeping y fixed.

The above can generalized for more variables by having a vector of RVs $(x_1, \ldots, x_N) \in \mathcal{X}^N$. In particular $p : \mathcal{X}^N \to \mathbb{R}$ is still a scalar. Marginalization still applies

$$p(x_1, x_3, \ldots, x_{N-1}) = \sum_{(x_2,x_N)\in\mathcal{X}^2} p(x_1, x_2, \ldots, x_N),$$

and so does conditioning

$$p(x_1, \ldots |x_N)p(x_N) = p(x_1, \ldots x_N).$$

However, now there can be mixed cases of conditioning.

Proposition 1.1 (Conditioning)**.** *Given the RVs x, y, z:*

$$p_{x|yz}(\bar{x}|\bar{y}, \bar{z}) = \frac{p_{xy|z}(\bar{x}, \bar{y}|\bar{z})}{p_{y|x}(\bar{y}|\bar{z})}.$$

This generalizes to more variables.

For RVs there is a notion if independence, that is, they do not affect each other.

Definition 1.5 (Independence)**.** The RVs x and y are said to be independent if $p(x|y) = p(x)$.

From the definition it follows that $p(x, y) = p(x)p(y)$ and $p(y|x) = p(y)$.

1.2 Expectation and Moments

The expectation is to be understood as a statistical average, or as a weighted sum with the coefficients being the probability.

Definition 1.6 (Expectation)**.** For a RV $x \in \mathcal{X}$

$$\mathop{\mathrm{E}}_{x}\{x\} = \sum_{\bar{x}\in\mathcal{X}} \bar{x}p_x(\bar{x}) \text{ or } \int_{\mathcal{X}} \bar{x}p_x(\bar{x})\, d\bar{x}.$$

In the definition above p_x can be replaced with a conditional $p_{x|y}$, to obtain the conditional expectation

$$\mathop{\mathrm{E}}_{x|y}\{x|\bar{y}\} = \int_{\mathcal{X}} \bar{x}p_{x|y}(\bar{x}|\bar{y})\, d\bar{x}.$$

Theorem 1.1 (Law of the Unconscious Statistician)**.** *Let $y = g(x) \in \mathcal{Y} = g(\mathcal{X})$ where $x \in \mathcal{X}$ is a DRV or CRV. Then*

$$\mathop{\mathrm{E}}_{y}\{y\} = \sum_{\bar{x}\in\mathcal{X}} g(\bar{x})p_x(\bar{x}) \text{ or } \int_{\mathcal{X}} g(\bar{x})p_x(\bar{x})\, d\bar{x},$$

or more compactly $\mathrm{E}_y\{y\} = \mathrm{E}_x\{g(x)\}$.

Definition 1.7 (Variance)**.** For a RV $x \in \mathcal{X}$

$$\mathop{\mathrm{Var}}_{x}\{x\} = \mathop{\mathrm{E}}_{x}\left\{(x - \mathop{\mathrm{E}}_{x}\{x\})(x - \mathop{\mathrm{E}}_{x}\{x\})^{\mathsf{T}}\right\}.$$

If x is a vector the resulting matrix is sometimes called covariance.

1.3 Sampling Distributions

Most mathematical libraries offer a function to sample a RV that is uniformly distributed on $(0, 1)$ (in MATLAB `rand()`). In other words we have a PDF

$$p_u(\bar{u}) = \begin{cases} 1 & \bar{u} \in (0, 1) \\ 0 & \text{otherwise} \end{cases}.$$

We can use p_u to generate samples for any other desired PDF using the following algorithms.

Algorithm 1.1 (Sample a DRV)**.** Given a derided PDF $\hat{p}_x$ for a DRV $x \in \mathcal{X} = \mathbb{Z}$, its cumulative distribution function (CDF) is a nondecreasing function

$$\hat{F}_x(\bar{x}) = \sum_{i=-\infty}^{\bar{x}} \hat{p}_x(i), = \Pr\{x \leq \bar{x}\}$$

and has the property that $\hat{F}_x(-\infty) = 0$ and $\hat{F}_x(\infty) = 1$.

Let $\bar{u}$ be the samples of $u \sim \mathcal{U}(0, 1)$. To find a sample $\bar{x}$ of x we solve for a $\bar{x}$ such that $\hat{F}_x(\bar{x} - 1) < \bar{u}$ and $\bar{u} \leq \hat{F}_x(\bar{x})$.

Algorithm 1.2 (Sample multiple finite DRV)**.** Given a desired joint PDF $\hat{p}_{xy}$ for the scalar DRVs $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, where $N_x = |\mathcal{X}|$ and $N_y = |\mathcal{Y}|$ are both finite, let $\mathcal{Z} = \{1, 2, \ldots, N_xN_y\}$. Then define a new $\hat{p}_z$ such that $\hat{p}_z(1) = \hat{p}_{xy}(1, 1)$, $\hat{p}_z(2) = \hat{p}_{xy}(1, 2)$, ..., $\hat{p}_z(N_xN_y) = \hat{p}_{xy}(N_x, N_y)$, and apply algorithm 1.1 to $\hat{p}_z$.

If the constraint of having finite sets of outcome is a problem, the following algorithm also works for infinite sets $\mathcal{X}$ and $\mathcal{Y}$.

Algorithm 1.3 (Sample multiple DRVs)**.** Given a desired joint PDF $\hat{p}_{xy}$, decompose it into $\hat{p}_{x|y}(\bar{x}|\bar{y})\hat{p}_y(\bar{y})$. Apply algorithm 1.1 to get a sample $\bar{y}$ for y via $\hat{p}_y(\bar{y})$, then with $\bar{y}$ fixed apply algorithm 1.1 again to get $\bar{x}$ for x via $\hat{p}_{x|y}(\bar{x}|\bar{y})$.

Remark. The independence of the uniform number generator between successive calls is important. Further, both algorithms were described for 2 variables but they both generalize any number of DRVs.

Algorithm 1.4 (Sample a CRV)**.** Given a desired piecewise continuous and bounded PDF $\hat{p}_x$ for a CRV x, let

$$\hat{F}_x(\bar{x}) = \int_{-\infty}^{\bar{x}} \hat{p}_x(\lambda)\, d\lambda = \Pr\{x \leq \bar{x}\}$$

be the CDF of x. To find a sample of x let $\bar{x}$ be any solution to $\bar{u} = \hat{F}_x(\bar{x})$, then x has PDF $p_x = \hat{p}_x$.

Algorithm 1.5 (Sample multiple CRVs)**.** Analogously to algorithm 1.3 decompose the given desired joint PDF into $\hat{p}_{xy}(\bar{x}, \bar{y}) = \hat{p}_{x|y}(\bar{x}|\bar{y})\hat{p}_y(\bar{y})$. Then, apply algorithm 1.4 to get a $\bar{y}$ for y via $\hat{p}_y(\bar{y})$, and with $\bar{y}$ fixed apply it again to get a sample $\bar{x}$ of x with $\hat{p}_{x|y}(\bar{x}|\bar{y})$.

1.4 Change of Variables

When we work with functions of RVs we usually also wish to know the PDFs of the results.

Proposition 1.2 (Change of variables for DRVs)**.** *Let p_y be given for $y \in \mathcal{Y}$ and consider $x = g(y) \in \mathcal{X} = g(\mathcal{Y})$. For each $\bar{x} \in \mathcal{X}$ let*

$$\mathcal{Y}_{\bar{x}} = \{\bar{y}_i : \bar{y}_i \in \mathcal{Y}, g(\bar{y}_i) = \bar{x}\},$$

then

$$p_x(\bar{x}) = \sum_{\bar{y}\in\mathcal{Y}_{\bar{x}}} p_y(\bar{y}) = \sum_{\bar{y}\in\mathcal{Y}:g(\bar{y})=\bar{x}} p_y(\bar{y}).$$

Proposition 1.3 (Change of variables for CRVs)**.** *Consider a strictly monotonic differentiable continuous function $x = g(y)$, then*

$$p_x(\bar{x}) = \frac{p_y(\bar{y})}{g'(\bar{y})} = \frac{p_y \circ g^{-1}(\bar{x})}{g' \circ g^{-1}(\bar{x})}.$$

Proposition 1.4 (Multivariate change of variables for CRVs)**.** *Let $g : \mathbb{R}^m \to \mathbb{R}^m, w \mapsto g(w)$, be a map with nonsingular Jacobian for all w, i.e.*

$$\det \frac{\partial g}{\partial w} = \det \begin{bmatrix} \partial_{w_1} g_1 & \cdots & \partial_{w_m} g_1 \\ \vdots & \ddots & \vdots \\ \partial_{w_1} g_m & \cdots & \partial_{w_n} g_m \end{bmatrix} \neq 0, \quad \forall w.$$

Further, assume that $z = g(w)$ has a unique solution for w in terms of z, say $w = h(z)$. Then

$$p_z(\bar{z}) = p_w(h(\bar{z})) \left| \det \frac{\partial g}{\partial w}(h(\bar{z})) \right|^{-1}.$$

1.5 Bayes' Theorem

Theorem 1.2 (Bayes' theorem)**.** *For the RVs x and z*

$$p(x|z) = p(z|x)\frac{p(x)}{p(z)}.$$

Remark. The interpretation is as follows: x is the unknown quantity of interest (state); $p(x)$ is the prior belief of the state; z is an observation related to the state; $p(z|x)$ is, for a given state, what is the probability of observing z? $p(x|z)$ is the posterior belief, that is the observation what is the probability that the state is x?

Bayes' theorem is a systematic way of combining prior beliefs with observations. Since observing z is usually not enough to directly determine x. That is because usually $\dim z < \dim x$ and with noise $p(z|x)$ that is not "sharp".

Proposition 1.5 (Generalization of Bayes' theorem)**.** *Suppose there are N (vector or scalar) observations $z_1, \ldots, z_N$. Assuming conditional independence, i.e.*

$$p(z_1, \ldots, z_N|x) = p(z_1|x) \cdots p(z_N|x),$$

then

$$p(x|z_1, \ldots, z_N) = \frac{p(x)\prod_i p(z_i|x)}{p(z_1, \ldots, z_N)},$$

where the normalization

$$p(z_1, \ldots, z_N) = \sum_{x\in\mathcal{X}} p(x) \prod_i p(z_i|x)$$

by the total probability theorem.

A possible interpretation for the independence assumption is that a measurement of the state x is corrupted by noise which is independent at each time step.

1.6 Gaussian Random Variables

For Kalman filter we need to know the properties of Gaussian RVs.

Definition 1.8 (Gaussian RV (GRV))**.** The PDF of a Gaussian (normally) distributed D-dimensional CRV $y = (y_1, \ldots, y_D)$ is

$$p(y) = \frac{1}{\sqrt{(2\pi)^D \det \Sigma}} \exp\left(-\frac{1}{2}(y - \mu)^{\mathsf{T}}\Sigma^{-1}(y - \mu)\right),$$

where $\mu \in \mathbb{R}^D$ is the mean vector and $\Sigma \in \mathbb{R}^{D\times D}$ and $\Sigma \succ 0$ (is a positive definite matrix) and symmetric ($\Sigma^{\mathsf{T}} = \Sigma$).

Proposition 1.6. *In the special case where Σ is a diagonal matrix with entries σ_i^2*

$$p(y) = \prod_{i=1}^{D} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(y_i - \mu_i)^2}{2\sigma_i^2}\right).$$

Hence, the PDF is a product of scalar GRVs, and thus the variables are mutually independent (the converse is also true).

Remark. For a time-dependent GRV $y(k)$ we say that it is *spatially* independent for a fixed time k if $y_1(k)$, ..., $y_D(k)$ are mutually independent, and *temporally* independent if $y(1)$, ..., $y(k)$ are mutually independent.

Definition 1.9 (Jointly GRVs)**.** Two GRVs x and y are said to be jointly Gaussian if the vector RV (x, y) is also a GRV.

Remark. If two variables are GRVs this does not imply that they are jointly GRVs.

Proposition 1.7. *If two GRVs $x \sim \mathcal{N}(\mu_x, \Sigma_x)$ and $y \sim \mathcal{N}(\mu_y, \Sigma_y)$ are independent, i.e. $p(x, y) = p(x)p(y)$ then they are jointly Gaussian.*

Proof. We assume hat $p(x, y) = p(x)p(y)$. By computing $p(x)p(y)$ we see that it is still Gaussian

$$\begin{aligned} p(x)p(y) &\propto \exp\Big(-\frac{1}{2}\Big[(x - \mu_x)^{\mathsf{T}}\Sigma_x^{-1}(x - \mu_x) \\ &\qquad + (y - \mu_y)^{\mathsf{T}}\Sigma_y^{-1}(x - \mu_y)\Big]\Big) \\ &= \exp\left(-\frac{1}{2}\begin{bmatrix} x - \mu_x \\ y - \mu_y \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} \Sigma_x^{-1} & 0 \\ 0 & \Sigma_y^{-1} \end{bmatrix} \begin{bmatrix} x - \mu_x \\ x - \mu_y \end{bmatrix}\right). \end{aligned}$$ □

Lemma 1.1 (Affine transformation of a GRV is a GRV)**.** *Let y be a GRV, M a matrix and b a vector of appropriate size, then $x = My + b$ is a GRV.*

Lemma 1.2 (Linear combination of jointly GRVs is a GRV)**.** *Let x and y be jointly Gaussian GRVs, then $z = M_x x + M_y y$ where M_x and M_y are constant matrices of appropriate dimensions, is a GRV.*

2 Bayesian Tracking

2.1 Problem Statement

Let $x(k) \in \mathcal{X}$ be a vector valued state at time $k \in \mathbb{Z}^+$ we wish to estimate. Assume x is a DRV and $\mathcal{X}$ is finite. Let $z(k)$ be a vector valued measurement which can be a DRV or CRV.

The model for the dynamics of the system and measurements may be nonlinear and time-varying:

$$\begin{aligned} x(k) &= q_{k-1}(x(k-1), v(k-1)), \qquad k = 1, 2, \ldots \\ z(k) &= h_k(x(k), w(k)), \end{aligned}$$

where $x(0)$, $\{v(\cdot)\}$, and $\{w(\cdot)\}$ are mutually independent with known PDFs. Any known input to the system is not explicitly included as it can be embedded into q_{k-1} and h_k.

Let $z(1 : k)$ denote the set $\{z(1), \ldots, z(k)\}$. The goal is to *efficiently* compute $p(x(k)|z(1 : k))$, that is, the full conditional probability density function of the state.

2.2 Recursive Estimator

Bayesian tracking is a two step recursive algorithm:

- **Prior update**: the state estimate is predicted forward using the process model.
- **Measurement update**: the prior is combined with observations / measurements.

Proposition 2.1 (Prior Update)**.** *We can predict the PDF of $x(k)$ based on past measurements $z(1 : k - 1)$ using*

$$p(x(k)|z(1 : k - 1)) = \sum_{x(k-1)\in\mathcal{X}} \underbrace{p(x(k)|x(k-1))}_{\text{process model}} \underbrace{p(x(k-1)|z(1 : k - 1))}_{\text{previous iteration}},$$

where $p(x(k)|x(k - 1))$ can be computed from $p(v(k - 1))$ and $q_{k-1}(\cdot, \cdot)$ using changes of variables.

Remark. The change of variables required in the prior update may not be straightforward.

Proposition 2.2 (Measurement Update)**.** *We combine the new observation using Bayes' rule and get*

$$p(x(k)|z(1 : k)) = p(x(k)|z(k), z(1 : k - 1)) = \frac{\overbrace{p(z(k)|x(k), z(1 : k - 1))}^{\text{measurement model}} \overbrace{p(x(k)|z(1 : k - 1))}^{\text{prior}}}{\underbrace{p(z(k)|z(1 : k - 1))}_{\text{normalization}}}$$

where the normalization can be computed using the total probability theorem

$$p(z(k)|z(1 : k - 1)) = \sum_{x(k)\in\mathcal{X}} p(z(k)|x(k))p(x(k)|z(1 : k - 1)).$$

To implement this on a computer, there is the following algorithm.

Algorithm 2.1 (Recursive Estimator)**.** Enumerate the state space $\mathcal{X} = \{0, 1, \ldots, N - 1\}$. Define:

- $\mathbf{a}_{k|k}^i = p_{x(k)|z(1:k)}(i|\bar{z}(1 : k))$ with $i = 0, 1,$ ..., $N - 1$, an array with N elements used to store the posterior PDF at time k.
- $\mathbf{a}_{k|k-1}^i = p_{x(k)|z(1:k-1)}(i|\bar{z}(1 : k - 1))$, $i = 0,$ ..., $N - 1$ to store the prior PDF at time k.

Then, compute an expression for

- $p_{z(k)|x(k-1)}(i|j)$ using the process model $x(k) = q_{k-1}(x(k-1), v(k-1))$ and $p_{v(k-1)}(\bar{v}(k-1))$.
- $p_{z(k)|x(k)}(\bar{z}(k)|i)$ using the observation model $z(k) = h_k(x(k), u$ and $p_{w(k)}(\bar{w}(k))$.

procedure ESTIMATOR($p_{x(0)}(0), \ldots, p_{x(0)}(N - 1)$)
  ▷ *Initialization*
  $\mathbf{a}_{0|0}^i \leftarrow p_{x(0)}(i), \quad \forall i \in \{0, \ldots, N - 1\}$
  loop
    for all $i \in \{0, \ldots, N - 1\}$ **do**
      ▷ *Prior update*
      $\mathbf{a}_{k|k-1}^i \leftarrow \sum_{j=0}^{N-1} p_{x(k)|x(k-1)}(i|j)\ \mathbf{a}_{k-1|k-1}^j$
      ▷ *Measurement update*
      $\mathbf{a}_{k|k}^i \leftarrow \frac{p_{z(k)|x(k)}(\bar{z}(k)|i)\ \mathbf{a}_{k|k-1}^j}{\sum_{j=0}^{N-1} p_{z(k)|x(k)}(\bar{z}(k)|j)\ \mathbf{a}_{k|k-1}^j}$
    end for
  end loop
end procedure

3 Extracting Estimates

The conditional PDF $p_{x|z}(\bar{x}|\bar{z})$ for the quantity x and its observations z captures the full information that one has about x in the Bayesian sense. However, we may be interest in just on an estimate $\hat{x}$ of x.

3.1 Maximum Likelihood (ML)

ML applies when $x \in \mathcal{X}$ is an unknown (constant) parameter without a (known) probabilistic description $p_x(\bar{x})$.

Definition 3.1 (Maximum Likelihood Estimator)**.** Let $z \in \mathcal{Z}$ be the measurement of the observation model $p_{z|x}(\bar{z}|\bar{x})$. For a given observation $\bar{z}$ ML seeks the value for the parameter x that makes the observation $\bar{z}$ most likely:

$$\hat{x}^{\mathrm{ML}} = \arg\max_{\bar{x}\in\mathcal{X}} p_{z|x}(\bar{z}|\bar{x}).$$

In this context $p_{z|x}(\bar{z}|\bar{x})$ is called *likelihood function*.

Example (Least Squares)*.* Suppose there is an observation model, whereby we have m measurements of n states:

$$\begin{bmatrix} z_1 \\ \vdots \\ z_m \end{bmatrix} = \begin{bmatrix} h_{11} & \cdots & h_{1n} \\ \vdots & \ddots & \vdots \\ h_{m1} & \cdots & h_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} w_1 \\ \vdots \\ w_m \end{bmatrix}.$$

Or, more compactly $z = Hx + w$, with $z, w \in \mathbb{R}^m$, $x \in \mathbb{R}^n$, $H \in \mathbb{R}^{m\times n}$, $m > n$ and $w_i \sim \mathcal{N}(0, 1)$. Further, we assume that the w_i are mutually independent and that H has full column rank ($Hx = 0$ implies $x = 0$).

We apply the multivariate change of variables of proposition 1.4, so in this case, $z = g_x(w)$ and the unique solution is $w = h_x(z) = z - Hx$. Here, the Jacobian is nonsingular since

$$\det \frac{\partial g_x}{\partial w} = \det \partial_w (Hx + w) = 1.$$

Hence

$$p_{z|x}(\bar{z}|\bar{x}) = p_w(\bar{z} - H\bar{x}) = \prod_{i=1}^{m} p_w(\bar{z}_i - H_i\bar{x}),$$

where in the first equality the conditioning of x can be dropped because w and x are assumed to be independent, then the second equality is justified by the mutual independence of the w_i's, and we use H_i to denote the i-th row of H (measurement). Now, because p_w is Gaussian

$$\prod_{i=1}^{m} p_w(\bar{z}_i - H_i\bar{x}) \propto \exp\left(-\frac{1}{2}\sum_{i=1}^{m}(\bar{z}_i - H_i\bar{x})^2\right).$$

To apply ML we need to maximize the sum in the exponent. Differentiating it and setting it to zero yields

$$H^{\mathsf{T}}(\bar{z} - H\bar{x}) = 0 \implies \bar{x} = \left(H^{\mathsf{T}}H\right)^{-1}H^{\mathsf{T}}\bar{z},$$

that is, the least squares solution.

3.2 Maximum a Posteriori (MAP)

The maximum a posteriori estimate applies when x is a RV with known PDF.

Definition 3.2 (Maximum a Posteriori Estimator)**.** Let z be the measurement of tbe observation model $p_{z|x}(\bar{z}|\bar{x})$. For a given observervation $\bar{z}$ MAP looks for the most likely choice of $\bar{x}$ given the previous belief about x:

$$\hat{x}^{\mathrm{MAP}} = \arg\max_{\bar{x}\in\mathcal{X}} p_{z|x}(\bar{z}|\bar{x})p_x(\bar{x}).$$

Remark. If p_x is constant (uniform), then $\hat{x}^{\mathrm{MAP}} = \hat{x}^{\mathrm{ML}}$, i.e. the prior does not provide any additional information.

Example. Consider the scalar observation model $z = x + w$ with $w \sim \mathcal{N}(0, 1)$, $x \sim \mathcal{N}(\mu, \sigma^2)$, w and x independent. Then

$$p_x(\bar{x}) \propto \exp\left(-\frac{1}{2}\frac{(\bar{x} - \mu)^2}{\sigma^2}\right), \text{ and}$$

$$p_{z|x}(\bar{z}|\bar{x}) \propto \exp\left(-\frac{1}{2}(\bar{z} - \bar{x})^2\right).$$

To apply MAP compute $p_{x|z}(\bar{x}|\bar{z})$, differentiate w.r.t. $\bar{x}$ and set to zero. The result is

$$\bar{x} = \frac{1}{1 + \sigma^2}\mu + \frac{\sigma^2}{1 + \sigma^2}\bar{z},$$

a weighted sum. Notice that when $\sigma \to \infty$ (x is uniformly distributed), $\bar{x} = \bar{z}$ is the ML.

3.3 Recursive Least Squares (RLS)

3.3.1 Problem Statement

We consider the observation model for a constant x

$$z(k) = H(k)x + w(k), \quad z(k), w(k) \in \mathbb{R}^m, x \in \mathbb{R}^n.$$

The given prior knowledge are the mean $\hat{x}_0 = \mathrm{E}\{x\}$ and variance $P_x = \mathrm{Var}\{x\}$ of x. Further, we assume that the noise $w(k)$ is zero-mean $\mathrm{E}\{w(k)\} = 0$, with known variance $R(k) = \mathrm{Var}\{w(k)\}$ and mutually independent in time k and also with x. Typically it is also the case that we have more equations than observations, i.e. $n > m$, and information is obtained "over time".

The goal is to compute an estimate $\hat{x}(k)$ of x from the observations $\{\bar{z}(1), \bar{z}(2), \ldots, \bar{z}(k)\}$ in the least square sense, that is, minimizing a quadratic error.

3.3.2 Standard Weighted LS

A naive strategy is to wait until one has collected enough samples to perform a weighted LS. Neglecting the prior, if we have data until time k:

$$\begin{bmatrix} z(1) \\ \vdots \\ z(k) \end{bmatrix} = \begin{bmatrix} H(1) \\ \vdots \\ H(k) \end{bmatrix} x + \begin{bmatrix} w(1) \\ \vdots \\ w(k) \end{bmatrix},$$

or $\mathbf{z} = \mathbf{H}x + \mathbf{w}$, then $\mathbf{R} = \mathrm{blockdiag}(R(1), \ldots, R(k))$

$$\begin{aligned} \hat{x}^{\mathrm{WLS}}(k) &= \arg\min_{\hat{x}}\{(\bar{\mathbf{z}} - \mathbf{H}\hat{x})^{\mathsf{T}}\mathbf{R}^{-1}(\bar{\mathbf{z}} - \mathbf{H}\hat{\mathbf{x}})\} \\ &= (\mathbf{H}^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\mathsf{T}}\mathbf{R}^{-1}\bar{\mathbf{z}} \end{aligned}$$

The obvious problem with this approach is that as time progresses (k increases) the LS problem grows in dimension, until it becomes computationally intractable.

3.3.3 Recursive LS

A more effective solution is a recursive algorithm. This is a precursor to the Kalman filter.

Algorithm 3.1 (Recursive LS)**.** Initialize an estimate of the initial state $\hat{x}_0$ with $P_0 = P_x = \mathrm{Var}\{x\}$. The recursion is as follows: Observe $\bar{z}_k$, then update

$$\begin{aligned} K_k &= P_{k-1}H_k^{\mathsf{T}}(H_kP_{k-1}H_k^{\mathsf{T}} + R_k)^{-1}, \\ \hat{x}_k &= \hat{x}_{k-1} + K_k(\bar{z}_k - H_k\hat{x}_{k-1}), \\ P_k &= (I - K_kH_k)P_{k-1}(I - K_kH_k)^{\mathsf{T}} + K_kR_kK_k^{\mathsf{T}}. \end{aligned}$$

An intuition for the structure of this estimator is given by the fact that if the measurement coincides with the estimate then $\bar{z}_k - H_k\hat{x}_{k-1} = 0$ and $\hat{x}_k = \hat{x}_{k-1}$.

4 Kalman Filter

The Kalman filter (KF) is a Bayesian estimator for linear time-varying (LTV) systems with Gaussian process and measurement noise. The KF is particularly exceptional because it has a closed form analytical solution.

4.1 Problem Statement

Consider a LTV system

$$x(k) = A(k-1)x(k-1) + u(k-1) + v(k-1),$$
$$z(k) = H(k)x(k) + w(k),$$

where $x(k)$ is the state, $u(k)$ is a known control input, $v(k) \sim \mathcal{N}(0, Q(k))$ is the process noise, $z(k)$ is the measurement and $w(k) \sim \mathcal{N}(0, R(k))$ is the sensor noise. Further the initial state $x(0) \sim \mathcal{N}(x_0, P_0)$ and $x(0), \{w(\cdot)\}$ and $\{v(\cdot)\}$ are mutually independent.

Remark. If $v(k)$ has nonzero mean, say $v(k) \sim \mathcal{N}(\alpha, Q(k))$ define $\bar{u} = u - \alpha$. Similarly if $w(k) \sim \mathcal{N}(\beta, R(k))$, redefine $\bar{z} = z - \beta$.

4.2 Bayesian Formulation

For the Bayesian interpretation of the KF we reformulate the problem using auxiliary variables "p" for the prediction, and "m" for measurement:

$$x_{\mathrm{m}}(0) = x(0)$$
$$x_{\mathrm{p}}(k) = A(k-1)x_{\mathrm{m}}(k-1) + u(k-1) + v(k-1)$$
$$z_{\mathrm{m}}(k) = H(k)x_{\mathrm{p}}(k) + w(k)$$

where $x_{\mathrm{m}}(k)$ is defined via its PDF

$$p_{x_{\mathrm{m}}(k)}(\xi) = p_{x_{\mathrm{p}}(k)|z_{\mathrm{m}}(k)}(\xi|\bar{z}(k)), \quad \forall \xi.$$

Lemma 4.1. *With the above formulation*

$$p_{x_{\mathrm{p}}(k)}(\xi) = p_{x(k)|z(1:k-1)}(\xi|\bar{z}(1:k-1)),$$
$$p_{x_{\mathrm{m}}(k)}(\xi) = p_{x(k)|z(1:k)}(\xi|\bar{z}(1:k)),$$

for all ξ and $k = 1, 2, \ldots$. That is, $x_{\mathrm{p}}(k)$ is the RV $x(k)$ conditioned on $z(1:k-1)$ and $x_{\mathrm{m}}(k)$ is the RV $x(k)$ conditioned on $z(1:k)$.

Now, introducing the following notation for the mean and variance of the prediction and measurement

$$\hat{x}_{\mathrm{p}}(k) = \mathrm{E}\left\{x_{\mathrm{p}}(k)\right\}, \qquad P_{\mathrm{p}}(k) = \mathrm{Var}\left\{x_{\mathrm{p}}(k)\right\},$$
$$\hat{x}_{\mathrm{m}}(k) = \mathrm{E}\left\{x_{\mathrm{m}}(k)\right\}, \qquad P_{\mathrm{m}}(k) = \mathrm{Var}\left\{x_{\mathrm{m}}(k)\right\},$$

we make use of the following fact:

Lemma 4.2. *For all k, $x_{\mathrm{p}}(k)$ and $x_{\mathrm{m}}(k)$ are GRVs.*

Hence, we can compute expressions for $\hat{x}_{\mathrm{p}}(k)$, $P_{\mathrm{p}}(k)$, $\hat{x}_{\mathrm{m}}(k)$ and $P_{\mathrm{m}}(k)$, i.e. the Kalman filter equations.

Theorem 4.1 (Kalman Filter Equations)**.** *The prior update or prediction step is*

$$\hat{x}_{\mathrm{p}}(k) = A(k-1)\hat{x}_{\mathrm{m}}(k-1) + u(k-1),$$
$$P_{\mathrm{p}}(k) = A(k-1)P_{\mathrm{m}}(k-1)A^{\mathsf{T}}(k-1) + Q(k-1),$$

then, the a posteriori update or measurement step is

$$P_{\mathrm{m}}(k) = \left(P_{\mathrm{p}}^{-1}(k) + H^{\mathsf{T}}(k)R^{-1}(k)H(k)\right)^{-1},$$
$$\hat{x}_{\mathrm{m}}(k) = \hat{x}_{\mathrm{p}}(k) + P_{\mathrm{m}}(k)H^{\mathsf{T}}(k)R^{-1}(k)(\bar{z}(k) - H(k)\hat{x}_{\mathrm{p}}(k)).$$

Therefore, the Kalman filter is the analytical solution to the Bayesian state estimation problem for a linear system with Gaussian distributions.

4.3 Alternate Formulation

A more common formulation for the Kalman filter is as follows.

Theorem 4.2 (KF with Kalman Gain)**.** *Let*

$$K(k) = P_{\mathrm{p}}(k)H^{\mathsf{T}}(k)\left(H(k)P_{\mathrm{p}}(k)H^{\mathsf{T}}(k) + R(k)\right)^{-1}$$

be the Kalman filter gain. Then, the a posteriori update can be computed with

$$\hat{x}_{\mathrm{m}}(k) = \hat{x}_{\mathrm{p}}(k) + K(k)(\bar{z}(k) - H(k)\hat{x}_{\mathrm{p}}(k)),$$
$$P_{\mathrm{m}}(k) = (I - K(k)H(k))P_{\mathrm{p}}(k).$$

Lemma 4.3 (Joseph form of the covariance update)**.** *In the a posteriori update of the KF with Kalman Gain $P_{\mathrm{m}}(k)$ can be computed with*

$$P_{\mathrm{m}}(k) = (I - K(k)H(k))P_{\mathrm{p}}(k)(I - K(k)H(k))^{\mathsf{T}} + K(k)R(k)K(k)^{\mathsf{T}}.$$

This form is more computationally expensive, but less sensitive to numerical errors.

This is the same as the recursive least square algorithm 3.1, therefore, the KF can also be applied to non-Gaussian RVs. Then, the KF can be interpreted as a linear unbiased estimator that minimizes the mean square error (MMSE). However, it the KF will no longer be optimal in the Bayesian sense.

Remark. If $A(k)$, $H(k)$, $Q(k)$, $R(k)$ and P_0 are known for all k, $P_{\mathrm{p}}(k)$, $P_{\mathrm{m}}(k)$ and $K(k)$ can all be precomputed offline.

Remark. The KF assumes positive definiteness for $P_0 \succ 0$, $Q(k) \succ 0$, $R(k) \succ 0$, however the KF also makes sense when they are positive semidefinite, as long the matrix inversions are well defined. That is, when some states are know exactly.

4.4 Detectability and Stabilizability

Definition 4.1 (Detectability)**.** Consider the deterministic system

$$x(k+1) = Ax(k), \quad z(k) = Hx(k),$$

where $k = 0, 1, \ldots$ and $x(0) = x_0 \in \mathbb{R}^n$. The system is said to be detectable if

$$\lim_{k\to\infty} z(k) = 0 \implies \lim_{k\to\infty} x(k) = 0,$$

for any x_0.

The above implies that the system is detectable iff $Hw \neq 0$ for any eigenvector w corresponding to an eigenvalue $|\lambda| \geq 1$ of the matrix A. In other words, we need to be able to see unstable modes. This can be expressed with the conditions:

$$\operatorname{rank}\begin{bmatrix} A - \lambda I \\ H \end{bmatrix} = n, \quad \forall \lambda \in \mathbb{C}, \quad |\lambda| \geq 1,$$

i.e. is full rank (PBH Test); Or, the eigenvalues of $A - LH$ or $(I - LH)A$ can be placed within the unit circle by a suitable choice of $L \in \mathbb{R}^{n\times m}$

Definition 4.2 (Stabilizability)**.** Consider the deterministic system

$$x(k+1) = Ax(k) + Bu(k), \quad z(k) = Hx(k),$$

where $k = 0, 1, \ldots$ and $x(0) = x_0$, $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$. The system is said to be stabilizable if

$$\exists u(0:k-1) \text{ such that } \lim_{k\to\infty} x(k) = 0$$

for any x_0.

Equivalently, if $\begin{bmatrix} A - \lambda I & B \end{bmatrix}$ is full rank for all $\lambda \in \mathbb{C}$ with $|\lambda| \geq 1$ (PBH Test); Or if the eigenvalues of $A - BK$ or $(I - BK)A$ can be placed within the unit circle by choosing $K \in \mathbb{R}^{m\times n}$. Stabilizability is the dual of detectability, i.e. (A, B) is stabilizable iff $(A^{\mathsf{T}}, B^{\mathsf{T}})$ is detectable.

4.5 The Steady-State KF

If the variance $P_{\mathrm{p}}(k)$ converges, so does $P_{\mathrm{m}}(k)$ and the KF gain: $\lim_{k\to\infty} K(k) = K_\infty$. The time invariance of K_∞ makes the implementation easier.

Definition 4.3 (Steady-State KF Gain, DARE)**.** Assuming $P_{\mathrm{p}}(k)$ converges to P_∞, then

$$P_{\mathrm{p}}(k+1) = AP_{\mathrm{p}}(k)A^{\mathsf{T}} + Q - AP_{\mathrm{p}}(k)H^{\mathsf{T}}(HP_{\mathrm{p}}(k)H^{\mathsf{T}} + R)^{-1}HP_{\mathrm{p}}(k)A^{\mathsf{T}}$$

(obtained by combining the KF equations) becomes the *discrete algebraic Riccati equation* (DARE)

$$P_\infty = AP_\infty A^{\mathsf{T}} + Q - AP_\infty H^{\mathsf{T}}(HP_\infty H^{\mathsf{T}} + R)^{-1}HP_\infty A^{\mathsf{T}},$$

and $K_\infty = P_\infty H^{\mathsf{T}}(HP_\infty H^{\mathsf{T}} + R)^{-1}$.

Theorem 4.3. *Assume $R \succ 0$, $Q \succeq 0$, and let G be any matrix such that $Q = GG^{\mathsf{T}}$. Then the following statements are equivalent*

- *(A, H) is detectable and (A, G) is stabilizable.*
- *The DARE has a unique solution $P_\infty \succeq 0$, $(I - K_\infty H)A$ is stable and*

$$\lim_{k\to\infty} P_{\mathrm{p}}(k) = P_\infty \text{ for any } P_{\mathrm{p}}(1) \succeq 0$$

and hence any $P_{\mathrm{m}}(0) = P_0 \succeq 0$.

Remark. (A, H) detectable means that all unstable modes can be observed, and (A, G) stabilizable means that noise excites unstable modes (if $Q \succ 0$, then this always holds).

5 Extended Kalman Filter (EKF)

5.1 Problem Statement

Consider the nonlinear discrete-time system

$$x(k) = q_{k-1}(x(k-1), v(k-1)), \quad z(k) = h_k(x(k), w(k)),$$

where

$$\mathrm{E}\left\{x(0)\right\} = x_0, \qquad \mathrm{Var}\left\{x(0)\right\} = P_0,$$
$$\mathrm{E}\left\{v(k-1)\right\} = 0, \qquad \mathrm{Var}\left\{v(k-1)\right\} = Q(k-1),$$
$$\mathrm{E}\left\{w(k)\right\} = 0, \qquad \mathrm{Var}\left\{w(k)\right\} = R(k).$$

Moreover, $x(0)$, $\{v(\cdot)\}$, $\{w(\cdot)\}$ are mutually independent, q_{k-1} is continuously differentiable wrt $x(k-1)$ and $v(k-1)$, and h_k is continuously differentiable wrt $x(k)$ and $w(k)$. That is, a system that is mildly nonlinear. Any known input to the system is implicitly absorbed in q_{k-1}.

5.2 The EKF Equations

The EKF works by linearizing the nonlinear system at the current state estimate and then apply the KF equations.

Theorem 5.1 (EKF process update equations)**.** *Linearizing $q_{k-1}(x(k-1), v(k-1))$ about $\hat{x}_{\mathrm{m}}(k-1)$ and $\mathrm{E}\left\{v(k-1)\right\} = 0$ yields*

$$x(k) \approx A(k-1)x(k-1) + L(k-1)v(k-1) + \xi(k-1),$$

where

$$A(k-1) = \partial_x q_{k-1}(\hat{x}_{\mathrm{m}}(k-1), 0),$$
$$L(k-1) = \partial_v q_{k-1}(\hat{x}_{\mathrm{m}}(k-1), 0),$$
$$\xi(k-1) = q_{k-1}(\hat{x}_{\mathrm{m}}(k-1), 0) - A(k-1)\hat{x}_{\mathrm{m}}(k-1).$$

The noise $L(k-1)v(k-1)$ is zero-mean with variance $L(k-1)Q(k-1)L^{\mathsf{T}}(k-1)$. The update equations are

$$\hat{x}_{\mathrm{p}}(k) = q_{k-1}(\hat{x}_{\mathrm{m}}(k-1), 0)$$
$$P_{\mathrm{p}}(k) = A(k-1)P_{\mathrm{m}}(k-1)A^{\mathsf{T}}(k-1) + L(k-1)Q(k-1)L^{\mathsf{T}}(k-1).$$

Intuition: Predict the mean state estimate forward using the nonlinear process model and update the variance according to the linearized equations.

Theorem 5.2 (EKF measurement update equations)**.** *Linearizing $h_k(x(k), w(k))$ about $\hat{x}_{\mathrm{p}}(k)$ and $\mathrm{E}\left\{w(k)\right\} = 0$ yields*

$$z(k) \approx H(k)x(k) + M(k)w(k) + \zeta(k),$$

where

$$H(k) = \partial_x h_k(\hat{x}_{\mathrm{p}}(k), 0),$$
$$M(k) = \partial_w h_k(\hat{x}_{\mathrm{p}}(k), 0),$$
$$\zeta(k) = h_k(\hat{x}_{\mathrm{p}}(k), 0) - H(k)\hat{x}_{\mathrm{p}}(k).$$

The term $\zeta(k)$ known and can be removed by defining the auxiliary measurement $z(k) - \zeta(k)$. The noise $M(k)w(k)$ is zero-mean with variance $M(k)R(k)M^{\mathsf{T}}(k)$, and the update equations are

$$K(k) = P_{\mathrm{p}}(k)H^{\mathsf{T}}(k)(H(k)P_{\mathrm{p}}(k)H^{\mathsf{T}}(k) + M(k)R(k)M^{\mathsf{T}}(k))^{-1},$$
$$\hat{x}_{\mathrm{m}}(k) = \hat{x}_{\mathrm{p}}(k) + K(k)(\bar{z}(k) - h_k(\hat{x}_{\mathrm{p}}(k), 0)),$$
$$P_{\mathrm{m}}(k) = (I - K(k)H(k))P_{\mathrm{p}}(k).$$

Intuition: correct for the mismatch between the actual measurement $\bar{z}(k)$ and its nonlinear prediction $h_k(\hat{x}_{\mathrm{p}}(k), 0)$, and correct the variance according to the linearized equations.

Remark. In this case the Kalman gain cannot be computed offline even if the noise distributions are known for all k, hence the EKF is more computationally expensive.

The EKF variables $\hat{x}_{\mathrm{p}}(k)$, $\hat{x}_{\mathrm{m}}(k)$, $P_{\mathrm{p}}(k)$ and P_{m} are only approximations! The EKF would be exact if q_{k-1} and $\mathrm{E}\left\{\cdot\right\}$ commuted:

$$\mathrm{E}\left\{q_{k-1}(x, v)\right\} = q_{k-1}(\mathrm{E}\left\{x\right\}, \mathrm{E}\left\{v\right\})$$

(the same is also for h_k), which is not the case for general nonlinear q_{k-1}, but true for linear q_{k-1}.

Therefore, the EKF does not have general convergence guarantees, but it works well for mildly nonlinear systems with unimodal noise distributions.

5.3 Hybrid EKF

In practice the process dynamics are usually continuous in time, and measurement taken a discrete time steps. That is

$$\dot{x}(t) = q(x(t), v(t), t), \quad z[k] = h_k(x[k], w[k]),$$

with $\mathrm{E}\left\{w[k]\right\} = 0$, $\mathrm{Var}\left\{w[k]\right\} = R$ (assumed constant for simplicity). We use the notation $x[k] = x(kT)$ with T being a constant sampling time.

Remark. One could discretize the dynamics and work with the EKF above. However, if the process is "fast" (needs very small T), working with a lot of samples may end up being more expensive than the hybrid EKF which works with the continuous-time dynamics.

Definition 5.1 (White Noise)**.** A discrete time signal $v_{\mathrm{d}}[k]$ is said to be white noise if $\mathrm{E}\left\{v_{\mathrm{d}}[k]\right\} = 0$ and $\mathrm{E}\left\{v_{\mathrm{d}}[k]v_{\mathrm{d}}{}^{\mathsf{T}}[k+n]\right\} = Q\delta_{\mathrm{d}}[n]$, where n is an integer and $\delta_{\mathrm{d}}[n]$ is the Kroneker delta, i.e $\delta_{\mathrm{d}}[0] = 1$ and $\delta_{\mathrm{d}}[n] = 0$ when $n \neq 0$.

Similarly a continuous time signal $v(t)$ is white noise if $\mathrm{E}\left\{v(t)\right\} = 0$ and $\mathrm{E}\left\{v(t)v(t)^{\mathsf{T}}\right\} = Q_c\delta(\tau)$ where $\delta(\tau)$ is the Dirac delta, which may be defined as

$$\delta(\tau) = \lim_{\epsilon\to 0} \begin{cases} 1/(2\epsilon) & -\epsilon < \tau < \epsilon \\ 0 & \text{otherwise} \end{cases}.$$

Remark. True continuous white noise cannot exist because it would have infinite power (has constant power spectral density), but it is nonetheless a useful approximation.

Theorem 5.3. *The Dirac pulse has the property that*

$$\int_a^b \xi(\tau)\delta(\tau)\, d\tau = \xi(0),$$

for all $a < 0$ and $b > 0$ and any real valued function $\xi(t)$ that is continuous at 0.

Theorem 5.4 (Hybrid EKF process update)**.** *Solve in the interval $(k-1)T \leq t \leq kT$ the ODE*

$$\dot{\hat{x}}(t) = q(\hat{x}(t), 0, t), \quad \hat{x}((k-1)T) = \hat{x}_{\mathrm{m}}[k-1]$$

and set $\hat{x}_{\mathrm{p}}[k] = \hat{x}(kT)$. Then solve in the same interval the matrix ODE

$$\dot{P}(t) = A(t)P(t) + P(t)A^{\mathsf{T}}(t) + L(t)Q_cL^{\mathsf{T}}(t),$$

where $A(t) = \partial_x q(\hat{x}(t), 0, t)$ and $L(t) = \partial_v q(\hat{x}, 0, t)$ with $P((k-1)T) = P_{\mathrm{m}}[k-1]$. Then set $P_{\mathrm{p}}[k] = P(kT)$.

The measurement equations are the same as the discrete-time EKF.

Theorem 5.5 (Hybrid EKF measurement update)**.** *Let*

$$H[k] = \partial_x h_k(\hat{x}_{\mathrm{p}}[k], 0), \quad M[k] = \partial_w h_k(\hat{x}_{\mathrm{p}}[k], 0).$$

Then the update equations are

$$K[k] = P_{\mathrm{p}}H^{\mathsf{T}}[k](H[k]P_{\mathrm{p}}[k]H^{\mathsf{T}}[k] + M[k]RM^{\mathsf{T}}[k])^{-1},$$
$$\hat{x}_{\mathrm{m}}[k] = \hat{x}_{\mathrm{p}}[k] + K[k](\bar{z} - h_k(\hat{x}_{\mathrm{p}}[k], 0)),$$
$$P_{\mathrm{m}}[k] = (I - K[k]H[k])P_{\mathrm{p}}[k].$$

Remark. Solving the matrix ODE is usually done with numerical ODE solvers such as Runge-Kutta (MATLAB e.g. `ode45`), and the accuracy largely depends on the order of the solver. Numerical accuracy is often at the cost of increased computation.

6 Particle Filter (PF)

The basic idea of the particle filter is to approximate the Bayesian state estimator for nonlinear systems and general (non Gaussian) noise distributions, by representing the state PDF by a large number of samples called *particles*. The overview of the particle filter is as follows:

- the particles are propagate through the process model,
- the particles are then weighted according to the measurement likelihood,
- a resampling generates a new set of particles.

6.1 Problem Statement

Consider the nonlinear discrete-time system

$$x(k) = q_{k-1}(x(k-1), v(k-1)), \quad z(k) = h_k(x(k), w(k)),$$

where $x(0)$, $\{v(\cdot)\}$ and $\{w(\cdot)\}$ are mutually independent and can be CRVs or DRVs with known PDF (no assumption on the shape of the PDF). Any known input is implicitly absorbed in $q_{k-1}(\cdot)$.

6.2 Monte Carlo (MC) Sampling

MC sampling is a basic technique o fusing a large number of samples called particles to approximate the PDF of a RV.

MC approximant of a DRV Let $y \in \mathcal{Y} = \{1, 2, \ldots, \bar{Y}\}$ be a DRV with PDF p_y. Then let $\{y^1, y^2, \ldots, y^N\}$ i.i.d with PDF p_y be DRVs model N random samples of y, and define

$$s_i^n = \delta(i - y^n) = \begin{cases} 1 & \text{if } y^n = i \\ 0 & \text{otherwise} \end{cases},$$

where $i = 1, \ldots, \bar{Y}$ and $n = 1, \ldots N$ (there are $N \times \bar{Y}$ s_i^n's). By the law of the unconscious statistician

$$\mathop{\mathrm{E}}_{y^n}\left\{s_i^n\right\} = \sum_{\bar{y}^n=1}^{\bar{Y}} \delta(i - \bar{y}^n)p_y(\bar{y}^n) = p_y(i).$$

Conversely, let the DRV s_i denote the average of s_i^n over n: $s_i = \frac{1}{N}\sum_{n=1}^N s_i^n$. By the law of large numbers (LLN) $\lim_{N\to\infty} s_i = \mathrm{E}_{y^n}\left\{s_i^n\right\} = p_y(i)$. By denoting the values that the random variables s_i^n take by $\bar{s}_i^n$, the can approximately write

$$p_y(i) \approx \frac{1}{N}\sum_{n=1}^N \bar{s}_i^n.$$

MC approximant of a CRV Let y be a CRV with PDF p_y. First we bin the CRV by letting Δy be a fixed bin size, and let y^n, $n = 1, 2, \ldots, N$ be i.i.d with PDF p_y. Like in the DRV define

$$s_a^n = \int_a^{a+\Delta y} \delta(\xi - y^n)\, d\xi = \begin{cases} 1 & \text{if } y^n \in [a, a + \Delta y) \\ 0 & \text{otherwise} \end{cases}.$$

with a indexing the bins. Again, by the law of the unconscious statistician

$$\mathop{\mathrm{E}}_{y^n}\left\{s_a^n\right\} = \int_{\mathbb{R}} \left[\int_a^{a+\Delta y} \delta(\xi - \bar{y}^n)\, d\xi\right] p_y(\bar{y}^n)\, d\bar{y}^n = \int_a^{a+\Delta y} p_y(\bar{y}^n)\, d\bar{y}^n = \Pr\left\{a \leq y < a + \Delta y\right\}.$$

Hence $\lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^N s_a^n = \mathrm{E}_{y^n}\left\{s_a^n\right\} = \Pr\left\{y \in [a, a + \Delta y)\right\}$ by the LLN, since

$$\frac{1}{N}\sum_{n=1}^N s_a^n = \frac{1}{N}\sum_{n=1}^N \int_a^{a+\Delta y} \delta(\xi - y^n)\, d\xi = \int_a^{a+\Delta y} \frac{1}{N}\sum_{n=1}^N \delta(\xi - y^n)\, d\xi \to \int_a^{a+\Delta y} p_y(\bar{y})\, d\bar{y}.$$

Thus for a smooth and bounded p_y and small Δy we approximate

$$p_y(\xi) \approx \frac{1}{N}\sum_{n=1}^N \delta(\xi - \bar{y}^n), \quad \forall \xi,$$

where we understand it in the sense that if you integrate both you get similar numbers.

Change of variables for MC approximant Consider $x = g(y)$, with $x \in \mathcal{X} = g(\mathcal{Y})$. Let $x^n = g(y^n)$, $j \in \mathcal{X}$, and $r_j^n = \delta(j - x^n)$ (similar to s_i^n), then

$$p_x(j) \approx \frac{1}{N}\sum_{n=1}^N \delta(j - g(\bar{y}^n)), \quad j \in g(\mathcal{Y}),$$

i.e. we can approximate p_x by using samples from p_y. This also holds for joint RVs

$$p_x(\xi) \approx \frac{1}{N}\sum_{n=1}^N \delta(\xi - \bar{x}^n), \quad \forall \xi,$$

where ξ and $\bar{x}^n$ may be vectors. Moreover, $\mathcal{X}$ and $\mathcal{Y}$ may also be infinite.

6.3 The Particle Filter

We take the general Bayesian estimator of §2 and similar to §4.2 we define the auxiliary variables $x_{\mathrm{p}}(k), x_{\mathrm{m}}(k), z_{\mathrm{m}}(k)$, however, unlike in the KF we make no assumption on the shape of the PDF nor assume a linear q_{k-1}. Hence

$$x_{\mathrm{p}}(k) = q_{k-1}(x_{\mathrm{m}}(k-1), v(k-1))$$
$$z_{\mathrm{m}}(k) = h_k(x_{\mathrm{p}}(k), w(k))$$
$$p_{x_{\mathrm{m}}}(k) = p_{x_{\mathrm{p}}|z_{\mathrm{m}}(k)}(\xi|\bar{z}(k)), \quad \forall \xi.$$

with $x_{\mathrm{m}}(0) = x(0)$, and with lemma 4.1:

$$p_{x_{\mathrm{p}}(k)}(\xi) = p_{x(k)|z(1:k-1)}(\xi|\bar{z}(1:k-1)),$$
$$p_{x_{\mathrm{m}}(k)}(\xi) = p_{x(k)|z(1:k)}(\xi|\bar{z}(1:k)).$$

Prior Update Given the PDF $p_{x_{\mathrm{m}}(k-1)}$ we construct $p_{x_{\mathrm{p}}(k)}$ by approximating both with MC sampling. Let

$$p_{x_{\mathrm{m}}(k-1)}(\xi) \approx \frac{1}{N}\sum_{n=1}^N \delta(\xi - \bar{x}_{\mathrm{m}}^n(k-1)), \quad \forall \xi$$

where $\{\bar{x}_{\mathrm{m}}^n(k-1)\}$ are N particles to approximate $x_{\mathrm{m}}(k-1)$. Then

$$p_{x_{\mathrm{p}}(k)}(\xi) \approx \frac{1}{N}\sum_{n=1}^N \delta(\xi - \bar{x}_{\mathrm{p}}^n(k)), \quad \forall \xi$$
$$\text{where} \quad \bar{x}_{\mathrm{p}}^n(k) = q_{k-1}(\bar{x}_{\mathrm{m}}^n(k-1), \bar{v}^n(k-1)),$$

and $\{\bar{v}^n(k-1)\}$ are MC samples of $p_{v(k-1)}$. In words: we "simply" propagate the particles through the dynamics.

Measurement Update (and Resampling) Given the PDF $p_{x_{\mathrm{p}}(k)}$ of $x_{\mathrm{p}}(k) \in \mathcal{X}$ and a measurement $\bar{z}(k)$ we approximately construct $p_{x_{\mathrm{m}}(k)}$ using MC sampling. By Bayes' rule (proposition 2.2 in §2.2):

$$p_{x_{\mathrm{m}}(k)}(\xi) = \frac{p_{z_{\mathrm{m}}(k)|x_{\mathrm{p}}(k)}(\bar{z}(k)|\xi)\; p_{x_{\mathrm{p}}(k)}(\xi)}{\sum_{\zeta\in\mathcal{X}} p_{z_{\mathrm{m}}(k)|x_{\mathrm{p}}(k)}(\bar{z}(k)|\zeta)\; p_{x_{\mathrm{p}}(k)}(\zeta)}, \quad \forall \xi.$$

Substituting the MC approximation[1] for $p_{x_{\mathrm{p}}(k)}$

$$p_{x_{\mathrm{m}}(k)}(\xi) \approx \sum_{n=1}^N \beta_n \delta(\xi - \bar{x}_{\mathrm{p}}^n(k))$$

where $\sum_{n=1}^N \beta_n = 1$, and

$$\beta_n = \alpha\; p_{z_{\mathrm{m}}(k)|x_{\mathrm{p}}(k)}(\bar{z}|\bar{x}_{\mathrm{p}}^n(k)),$$
$$\alpha = \left(\sum_{n=1}^N p_{z_{\mathrm{m}}(k)|x_{\mathrm{p}}(k)}(\bar{z}(k)|\bar{x}_{\mathrm{p}}^n(k))\right)^{-1}.$$

In words: at points of high prior there are many particles, in the posterior they are scaled by the measurement likelihood.

To complete the measurement update we need to resample the particles. This is algorithm 1.1 of §1.3. Repeat N times:

- Select a random number $r \sim \mathcal{U}(0, 1)$,
- Pick particle $\bar{n}$ such that

$$\sum_{n=1}^{\bar{n}-1} \beta_n < r \quad \text{and} \quad \sum_{n=1}^{\bar{n}} \beta_n \geq r.$$

The result are N *new* particles $\bar{x}_{\mathrm{m}}^n(k)$ from a subset of the old particles that have all equal weight.

6.4 Sample Impoverishment

A possible problem of the PF is that all particles may converge to the same one and become a bad representation of the PDF. This is because we have a finite number of samples N, and is called *sample impoverishment*. The simplest solution to prevent this is *roughening*.

After the resampling perturb the particles with

$$\bar{x}_{\mathrm{m}}^n(k) \leftarrow \bar{x}_{\mathrm{m}}^n(k) + \Delta x^n(k),$$

where $\Delta x^n(k)$ is drawn from a zero-mean, finite-variance distribution. To choose the variance of said distribution a simple way is to let

$$\sigma_i = KE_iN^{-1/d},$$

where $K \ll 1$ is a tuning parameter, d is the dimension of the state space, $E_i = \max_{n_1,n_2} |x_{\mathrm{m},i}^{n_1} - x_{\mathrm{m},i}^{n_2}|$ is the maximum inter-sample variability and $N^{-1/d}$ is related to the spacing between nodes of a uniform grid.

[1] and making use of the fact that $f(\xi)\delta(\xi) = f(0)\delta(\xi)$