Predictively-Oriented Kalman Filtering

Zheyang Shen
Newcastle University
zheyangshen@gmail.com

Gerardo Duran-Martin
University of Oxford
G.Duran@me.com

Chris. J. Oates
Newcastle University
Chris.Oates@ncl.ac.uk

Abstract

This paper presents a post-Bayesian approach to online filtering in nonlinear state-space models, capable of avoiding over-confident inferences in settings where either the dynamical model, the measurement model, or *both*, could be misspecified. This is addressed using *predictively oriented* (PrO) posteriors, an emerging paradigm in which learning (i.e., posterior concentration) occurs if and only if the overall model is well-specified, without strict adherence to Bayes' theorem. As the characterisation of PrO posteriors is challenging, our main technical contribution is a fast approximate linear-Gaussian update procedure, based on convex optimisation and analogous to an (iterated) extended Kalman filter. The methodology, which we call EKF-PrO, has no tunable hyper-parameters and has a computational cost comparable to that of existing filtering methods. Performance is empirically assessed on a range of linear and non-linear applications, in which the state-space model is systematically misspecified.

1 Introduction

Our setting is a state-space model

$$x_k = f(x_{k-1}, \epsilon_k) \tag{1}$$
$$y_k = h(x_k, \xi_k) \tag{2}$$

where the noise terms $(\epsilon_i)_{i\in\mathbb{N}}$ and $(\xi_i)_{i\in\mathbb{N}}$ are independent. The functions f and h may be nonlinear and only the y_k are observed. The *filtering* task refers to the inference of the state variable x_k given the observations $y_{1:k} := (y_i)_{i=1}^k$. Note that this set-up also allows for f and h to depend on one or more unknown (static) parameters θ, which can be cast as additional components of the state variable x_k and held constant under the dynamic model (1). State-space models of the form (1) and (2) arise in time series analysis [Durbin and Koopman, 2012], tracking in computational vision [Chen, 2011], and sensor fusion in robotics [Thrun, 2002].

Since the latent state x_k may not be strongly constrained from the observations $y_{1:k}$, a Bayesian approach to quantify uncertainty in x_k is natural in the filtering context. The *Kalman filter* (KF) was originally derived as an optimal (minimum mean square error) linear estimator of the moments of x_k given $y_{1:k}$ [Kalman, 1960]; for linear f, h, and Gaussian ϵ_k, ξ_k with known mean and covariance, these moments characterise the exact posterior (or *filtering*) distributions $p(x_k|y_{1:k})$, and it is this distributional interpretation of the KF that we adopt [Särkkä and Svensson, 2023]. For nonlinear f, h, algorithms such as the *extended* Kalman filter (EKF) exploit local linearisation to approximate the filtering distribution while retaining the efficient online nature of the KF. A textbook treatment of algorithms for Bayesian inference in the state-space model in (1) and (2) can be found in Särkkä and Svensson [2023].

The focus of this paper is the filtering problem in settings where *the true data-generating process is not an instance of the state-space model* in (1) and (2). In such settings Bayesian inference, and the various numerical methods that approximate Bayesian inference (e.g. EKF), can produce inferences

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

and predictions that are *over-confident* because they are not capable of recognising and adapting to model misspecification. Accordingly, several solutions have been proposed to mitigate this effect:

Misspecification of the Measurement Error Model One common cause of misspecification for the measurement error model (2) is the presence of *outliers* in the dataset [Liu, 2020]. Accordingly several potential remedies have been proposed, including hierarchical measurement models; see e.g., [Ting et al., 2007, Piché et al., 2012, Agamennoni et al., 2012, Nurminen et al., 2015, Huang et al., 2016, Wang et al., 2018], or the use of robust (e.g. Huber) loss functions in place of the usual negative log-likelihood [e.g. Boncelet and Dickinson, 1983, Karlgaard, 2015, Das, 2023]. The latter can be considered as *generalised* Bayesian methods [Bissiri et al., 2016, Knoblauch et al., 2022], and were further developed from this perspective in Boustati et al. [2020], Duran-Martin et al. [2024, 2025]. In particular, by departing from the traditional Bayesian update it is possible to down-weight the influence of outliers as they are encountered. That is, if an observed datum y_k is far into the tail of the data predictive distribution based on $y_{1:k-1}$, then the effect of y_k on inferences and predictions can be manually discounted. This idea enabled Duran-Martin et al. [2024] to rigorously establish outlier-robustness properties of their generalised Bayes method, called the *weighted observation likelihood filter* (WoLF). However, while generalised Bayesian methods are capable of ignoring occasional outliers, they do not provide a solution to *systematic* model misspecification; if all data were deemed to be outliers, then all data would be ignored.

Misspecification of the Dynamical Model An alternative form of misspecification arises from an incomplete knowledge of the latent dynamics (1). This occurs, for example, when tracking high-speed maneuvering objects, when a system undergoes unmodelled phase transitions (see e.g., bottom row of Figure 1), or when the transition function is incorrectly parametrised. In such cases, the resulting observations may exhibit abrupt jumps, or the estimation error can increase over time—a phenomenon commonly referred to as *divergence* [Schlee et al., 1967, Fitzgerald, 2003]. To address this form of misspecification, various methods that admit closed-form (including fixed-point) solutions have been proposed. This includes approaches that assume heavy-tailed noise in the observation and latent space [Roth et al., 2013, Huang et al., 2017], variational Bayesian methods for estimating unknown system dynamics [Roth et al., 2017a, Huang et al., 2020, Vilmarest and Wintenberger, 2024], as well as heuristics that perform *covariance inflation* for adaptation Mehra [2003], Kelly [2002], Chang et al. [2023]. However, these approaches typically rely on explicit parametrisations of the latent noise process $(\epsilon_i)_{i\in\mathbb{N}}$, or specification of hyper-parameters, requiring *a priori* knowledge of how the dynamical model might be misspecified. Bayesian deep learning offers a point of contrast, enabling dynamics to be *learned* [e.g. Dahan et al., 2025], but at the expense of introducing hyper-parameters and substantially increased computational overhead.

In summary, the state-of-the-art for closed-form Bayesian filtering under model misspecification mainly focuses on either defining a more sophisticated hierarchical state-space model and developing efficient computational methods to facilitate Bayesian inference, or down-weighting the contribution from particular observations y_k within a generalised Bayesian framework. Neither approach is ideal; we are typically unable to write down a well-specified state-space model, and typical (generalised) Bayesian posteriors do not provide calibrated uncertainty when the statistical model is misspecified, e.g. as demonstrated in the recent work of Shen et al. [2025]. In the filtering context, a collapse of uncertainty causes overconfident inferences and predictions when the state-space model is misspecified, as illustrated in the first two columns of Figure 1.

1.1 Our Contributions

To deal with generic model misspecification in (1) and (2), we contend that an alternative to (generalised) Bayesian inference is required; rather than attempting to fit a more accurate model in a Bayesian framework in the hope that the resulting posterior predictions will be well-calibrated, we propose instead to target the *predictive* performance of the original state-space model. Our inspiration comes from deep learning, where the pitfalls of Bayesian inference are explicitly acknowledged and predictively oriented techniques such as posterior tempering are routinely used [Kuleshov et al., 2018, Wenzel et al., 2020]. However, *post hoc* techniques such as posterior tempering introduce additional hyper-parameters, a problem which we seek to avoid.

To achieve this we pursue a promising and emerging research direction in post-Bayesian methodology, which was termed *predictively oriented* (PrO) posteriors in McLatchie et al. [2025]. PrO posteriors

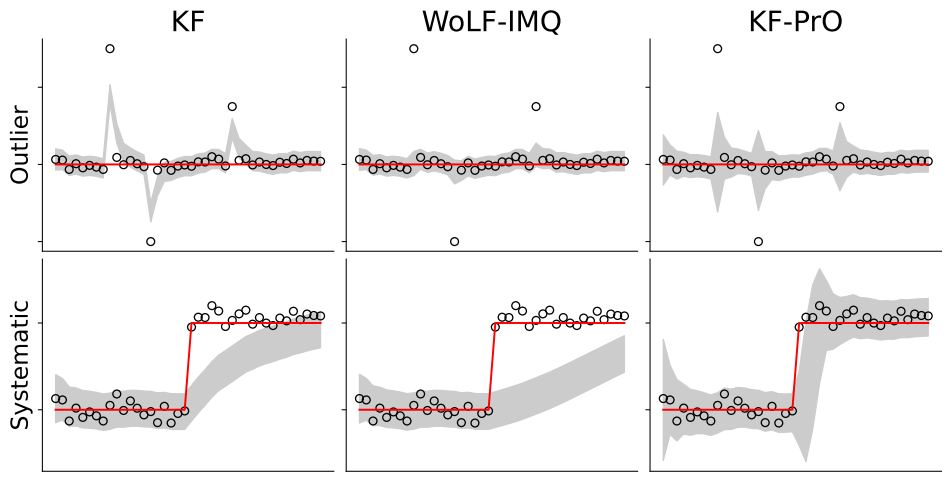

Figure 1: Illustration on the Gaussian location model. Data $(y_k)_{k=1}^n$ are shown as circles and the true latent state is in red; in the top row the data are corrupted by outliers, while in the bottom row the model is systematically misspecified. Posterior predictives are displayed for KF (left), a generalised Bayes method called WoLF-IMQ due to Duran-Martin et al. [2024] (centre) and our proposed KF-PrO (right). The grey shadings identify the 95% predictive confidence interval.

seek a distribution over latent variables for which the implied data predictive distribution most closely matches the observed dataset. This represents a fundamental departure from Bayesian and generalised Bayesian methods. For example, in the case of a static model of observations parametrised by θ, the PrO posterior will concentrate around a single "best" parameter $\theta_\star$ if and only if the model is well-specified. In contrast, both Bayesian and generalised Bayesian posteriors concentrate around $\theta_\star$ irrespective of whether the model is well-specified [Miller, 2021]; this leads to a problematic *collapse of uncertainty* when the model is misspecified.

The main technical challenge in developing a PrO posterior for state-space models is that PrO posteriors are only implicitly defined, in contrast (generalised) Bayesian posteriors for which one can express the posterior density as the product of the prior density and a (generalised) likelihood. Although PrO posteriors can be approximated to arbitrary precision using techniques such as gradient flows and mean-field Langevin dynamics, these techniques are computationally intensive and therefore unsuitable in the state-space context [Chazal et al., 2025]. Our main technical contribution in Section 2 is therefore a fast approximate update for PrO posteriors in linear (or linearised) state-space models, analogous to how the EKF enables fast approximate inference in the Bayesian framework.

Through detailed empirical assessment in Section 3, we demonstrate that EKF-PrO out-performs existing robust and adaptive filtering methods when the state-space model is systematically misspecified, at the price of a modest loss of statistical efficiency relative to the KF when the model is well-specified. A discussion of our findings is contained in Section 4.

2 Methodology

Section 2.1 casts probabilistic filtering as an optimisation task, relating common filtering algorithms to optimisation objectives that are implicitly being minimised. Section 2.2 builds on this perspective to introduce a novel PrO filtering algorithm in terms of its associated optimisation objective, initially for the case where f and h are linear, and Section 2.3 compares the PrO filtering distribution with standard KF. The general case of the PrO Kalman filter, where f and h can be nonlinear, is addressed by the EKF-PrO algorithm, presented in Section 2.4.

2.1 Filtering as an Optimisation Task

This section casts probabilistic filtering as an optimisation task. The latent state x_k is assumed to take values in $\mathbb{R}^d$ and we let $\mathcal{P}(\mathbb{R}^d)$ denote the set of probability distributions on $\mathbb{R}^d$. The

conditional distribution of x_k (resp. y_k) given x_{k-1} implied by (1) and (2) is denoted $p(x_k|x_{k-1})$ (resp. $p(y_k|x_k)$), and a prediction for x_k given the observations $y_{1:l}$ is denoted $q_{k|l} \in \mathcal{P}(\mathbb{R}^d)$, where to improve presentation we interchange freely between densities and distributions, and use the standard filtering notation in Särkkä and Svensson [2023].

Through the lens of optimisation, existing Bayesian and generalised Bayesian filtering methods can be formulated as minimising objective functionals defined on $\mathcal{P}(\mathbb{R}^d)$. Following standard filtering terminology, we present this in two steps: the *prediction* step uses the filtering distribution from the previous step $q_{k-1|k-1}$ to calculate the latent predictive distribution $q_{k|k-1}$, then the *update* step minimises an objective function in which $q_{k|k-1}$ appears as a regulariser via the Kullback–Leibler divergence (KLD):

Prediction Step Given the filtering distribution from the previous time step, $q_{k-1|k-1}(x_{k-1})$, the latent predictive distribution is obtained using the Chapman–Kolmogorov equation

$$q_{k|k-1}(x_k) := \int p(x_k|x_{k-1}) q_{k-1|k-1}(x_{k-1}) \mathrm{d}x_{k-1}. \tag{3}$$

Update Step Given the latent predictive distribution in (3), we can then formulate the filtering distribution as the solution of an infinite-dimensional optimisation problem on the set of probability measures

$$q_{k|k} := \underset{q \in \mathcal{P}(\mathbb{R}^d)}{\arg\min} \quad \ell(y_k; q) + \mathrm{KL}(q \| q_{k|k-1}) \tag{4}$$

where the loss function ℓ is to be specified.

This optimisation-centric formulation unifies Bayesian and generalised Bayesian filtering, which correspond to different choices of loss function ℓ in (4):

Example 1 (Bayes; KF)**.** *Standard Bayesian inference is recovered using the (negative) average log-likelihood loss*

$$\ell^{(\mathrm{Bayes})}(y_k; q) := -\int \log(p(y_k|x_k)) q(x_k) \mathrm{d}x_k, \tag{5}$$

a result which follows from the Donsker–Varadhan characterisation of the KLD [see e.g. Knoblauch et al., 2022]. For linear f, h and Gaussian ϵ_k, ξ_k, the above updates exactly implement the KF*. More broadly, this perspective encompasses methods which attempt Bayesian inference for a variety of more flexible state-space models, mentioned in Section 1.*

Example 2 (Generalised Bayes; WoLF)**.** *As an exemplar of generalised Bayesian filtering methods, the WoLF algorithm of Duran-Martin et al. [2024] employs*

$$\ell^{(\mathrm{WoLF})}(y_k; q) := -\int w(y_k, \hat{y}_k) \log(p(y_k|x_k)) q(x_k) \mathrm{d}x_k$$

where w is a non-negative weight function*, comparing a prediction $\hat{y}_k$ based on $y_{1:k-1}$ to the actual datum y_k observed. The weight function is user-specified; for example as an inverse multi-quadric weight function gives rise to the* WoLF-IMQ *method, while a variant based on Mahalanobis distance was termed* WoLF-TMD*. In either case, if the observed y_k is far from the predicted $\hat{y}_k$, then y_k is deemed to be unreliable and its influence on the filtering distribution is reduced.*

The principal drawback of (generalised) Bayesian methods in this context is that they lack awareness that the state-space model could be misspecified. The issue is seen most easily in the simplest setting where $x_k = \theta$ is a static parameter and we model $y_k = \theta + \xi_k$ as a noisy observation of θ; given sufficient data, both Bayesian and generalised Bayesian posteriors collapse onto a single "best" parameter $\theta_\star$ [Miller, 2021], while such arbitrarily confident predictions are inappropriate when the state-space model on which they are based is misspecified. To address this problem we adopt an alternative, predictively-oriented perspective next.

2.2 Predictively-Oriented Kalman Filtering

Our novel PrO filtering method employs a loss function which explicitly promotes *predictive* performance of the filtering distribution. Taking inspiration from the earlier work of Masegosa [2020],

Jankowiak et al. [2020], Sheth and Khardon [2020], Morningstar et al. [2022], Lai et al. [2026], Shen et al. [2025], McLatchie et al. [2025], Liu et al. [2025] in the static inference context, we define the PrO filtering objective as follows:

Definition 1 (PrO filtering objective)**.** *The* PrO filter *employs the loss*

$$\ell^{(\mathrm{PrO})}(y_k; q) := -\log\left(\int p(y_k|x_k)q(x_k)\mathrm{d}x_k\right). \tag{6}$$

Unlike attempts to perform Bayesian inference with a more sophisticated state-space model, or generalised Bayesian solutions in which additional hyper-parameters must be user-specified, the PrO filter performs a probabilistic *lifting* of the original measurement model

$$p(y_k|x_k) \quad \rightsquigarrow \quad p_q(y_k) := \int p(y_k|x_k)q(x_k)\mathrm{d}x_k$$

viewing q (rather than x_k) as the 'parameter' of the model, in this case an (infinite) mixture model in a similar manner to nonparametric maximum likelihood [Laird, 1978]. As such, our construction circumvents the requirement to manually design a more sophisticated model, and does not introduce any hyper-parameter which need to be specified (cf. Table 1).

Compared to the Bayesian loss (5), the order of the logarithm and the integral are interchanged; this key difference causes the PrO filter to behave in a more desirable way compared to (generalised) Bayesian methods when the state-space model is misspecified. Indeed, considering y_k to be a sample from the true data-generating distribution p_{true}, we can see that

$$\mathbb{E}_{y_k\sim p_{\text{true}}(\cdot|x_k)}[\ell^{(\mathrm{PrO})}(y_k; q)] = \underbrace{\mathrm{KL}\left(p_{\text{true}}(\cdot|x_k)\,\|p_q\right)}_{\text{predictive fit}} + \text{constant}, \tag{7}$$

which demonstrates that PrO filtering explicitly minimises the KLD between the predictive p_q and the true data distribution $p_{\text{true}}(\cdot|x_k)$, automatically calibrating predictions even when the model is misspecified [McLatchie et al., 2025].

Why Does PrO Filtering Adapt to Misspecification? The difference between the Bayesian loss (5) and the PrO loss (6) is whether the filtering distribution is asked to explain the datum y_k *pointwise* or *as a whole*. The Bayesian loss scores q by the average log-likelihood of individual states, so by Jensen's inequality any dispersion in q is penalised irrespective of whether it improves the fit; the filtering distribution therefore contracts at the usual rate even when no state value can explain the data. The PrO loss instead scores the single predictive density $p_q(y_k)$ that q implies, so that dispersion is penalised only insofar as it degrades the predictive fit. If the model is well-specified, the best predictive is obtained by concentrating on the true state and the familiar behaviour is recovered; if it is misspecified, no point mass can reproduce the data and a more diffuse q yields a better predictive, so uncertainty is retained. Both forms of misspecification are detected through the same channel, a poor one-step-ahead predictive fit: under-stated measurement noise and a displaced propagated mean (due to misspecified dynamics) each manifest as a large innovation, to which the PrO filter responds by inflating its covariance (cf. Proposition 4). The PrO filter thus reacts to the symptom rather than to a hypothesised cause, needing neither to diagnose which component is misspecified nor to be told how it might fail, which allows it to remain hyper-parameter free. This behaviour is *adaptive* rather than *robust* in the usual sense; instead of discounting anomalous data, the PrO filter increases its uncertainty when misspecification is detected.

Remark 1 (General scoring rules)**.** *Alternative scoring rules could be used in* (6) *and would correspond to alternative divergences in* (7) *[Gneiting and Raftery, 2007]. However, the log scoring rule ensures that both the loss and the regulariser in* (4) *are in the same units (i.e. nats), avoiding the need to introduce a 'learning rate' to balance the size of the two terms [Liu et al., 2025, Remark 2].*

The key technical question that we address in this work is how to efficiently (and perhaps approximately) perform the update in (4) when the PrO filtering objective is used. Techniques such as gradient flows and mean-field Langevin dynamics can in principle be applied to approximate $q_{k|k}$ [Shen et al., 2025, McLatchie et al., 2025, Chazal et al., 2025], but they are computationally intensive and therefore impractical when rapid calculation is required, as is often the case in a filtering context. Instead, we develop a fast update analogous to the `EKF`, based on recursive Gaussian variational approximations [Lambert et al., 2022, Jones et al., 2024], which we call the `EKF-PrO`.

Computing the PrO Filter For presentational purposes we begin by considering the simplest case where f, h are linear and ϵ_k, ξ_k are Gaussian, i.e. the setting of the KF:

$$x_k|x_{k-1} \sim \mathcal{N}(A_{k-1}x_{k-1}, \Lambda_{k-1})$$
$$y_k|x_k \sim \mathcal{N}(H_k x_k, R_k)$$

for positive definite Λ_{k-1} and R_k. The optimal filtering distribution q minimising the PrO objective (4) is not necessarily Gaussian, even with a Gaussian prior $q_{k|k-1}$. To relieve the computational burden, we define the KF-PrO as the optimal filtering distribution within the Gaussian variational family

$$q_{k|k} := \operatorname*{arg\,min}_{q \in \mathcal{Q}(\mathbb{R}^d)} -\log\left(\int p(y_k|x_k)q(x_k)\mathrm{d}x_k\right) + \mathrm{KL}(q\|q_{k|k-1}), \tag{8}$$

where $\mathcal{Q}(\mathbb{R}^d)$ is the set of all Gaussian distributions on $\mathbb{R}^d$.

Given a filtering distribution $q_{k-1|k-1}$ from the previous time step of the form $\mathcal{N}(m_{k-1}, P_{k-1})$, the **prediction step** coincides with the standard KF, as the Chapman–Kolmogorov equation yields a closed-form solution

$$q_{k|k-1} = \mathcal{N}(\tilde{m}_k, \tilde{P}_k), \qquad \tilde{m}_k = A_{k-1}m_{k-1}, \qquad \tilde{P}_k = A_{k-1}P_{k-1}A_{k-1}^\top + \Lambda_{k-1}. \tag{9}$$

However, the **update step** does not have a closed form in general. Nevertheless, the optimisation objective in (8) can be explicitly computed:

Proposition 1 (Explicit form of PrO filtering objective)**.** *Parametrising $q \in \mathcal{Q}(\mathbb{R}^d)$ as $q = \mathcal{N}(m, P)$, we can write down the objective function in* (8) *as*

$$\mathcal{J}_k(m, P) \overset{+C}{:=} \frac{1}{2}\left[\mathrm{tr}\left(\tilde{P}_k^{-1}P\right) + \left\|\tilde{P}_k^{-\frac{1}{2}}(m - \tilde{m}_k)\right\|^2 + \log\frac{|S_k(P)|}{|P|} + \left\|S_k^{-\frac{1}{2}}(P)v_k(m)\right\|^2\right] \tag{10}$$

where $S_k(P) := H_kPH_k^\top + R_k$ and $v_k(m) := y_k - H_km$.

The proof of Proposition 1 is contained in Section A.1. To minimise (10), we first note that $\mathcal{J}_k$ is convex in m when P is fixed, and that the minimiser $m_k^\star(P)$ can be explicitly computed:

Proposition 2 (Explicit minimisation over m for fixed P)**.** *Let $\mathcal{J}_k$ be the objective function in* (10) *and let the covariance matrix P be fixed. Then $m \mapsto \mathcal{J}_k(m, P)$ is uniquely minimised at*

$$m_k^\star(P) = \tilde{m}_k + K_k(P)(y_k - H_k\tilde{m}_k), \tag{11}$$

where $K_k(P) := [\tilde{P}_k^{-1} + H_k^\top S_k^{-1}(P)H_k]^{-1}H_k^\top S_k(P)^{-1}$.

The proof of Proposition 2 is contained in Section A.2. Similarly to the KF, (11) takes the form of a prior mean ($\tilde{m}_k$) plus a 'correction' term which depends on how much the observed datum (y_k) deviates from the predictive mean ($H_k\tilde{m}_k$). The matrix K_k in (11) is an analogue of the *Kalman gain* matrix for the KF-PrO; we elaborate on this relationship in Section 2.3.

Substituting $m_k^\star(P)$ into (10) yields a reduced objective $\Phi_k(P) := \mathcal{J}_k(m_k^\star(P), P)$ over covariance matrices, which turns out to be convex:

Proposition 3 (Convexity of the reduced PrO filtering objective)**.** *The reduced objective can be written as*

$$\Phi_k(P) \overset{+C}{=} \frac{1}{2}\Big[\mathrm{tr}(\tilde{P}_k^{-1}P) + \log\left|P^{-1} + H_k^\top R_k^{-1}H_k\right| + v_k(\tilde{m}_k)^\top\left(S_k(P) + H_k\tilde{P}_kH_k^\top\right)^{-1}v_k(\tilde{m}_k)\Big],$$

where each of the three terms is convex in P on the cone of symmetric positive definite matrices. In particular, Φ_k is convex and attains its minimum in this cone.

The proof of Proposition 3 is contained in Section A.3. Convexity ensures that any stationary point of Φ_k is a global minimiser, so that standard convex optimisation methods can be applied to obtain the KF-PrO covariance $P_k = \arg\min_P \Phi_k(P)$, and hence also the PrO filtering mean $m_k = m_k^\star(P_k)$; full implementational details are contained in Section B.

2.3 Comparison with the Standard Kalman Filter

This section compares the one-step behaviour of KF with our proposed KF-PrO, with the main findings summarised in the following result:

Proposition 4 (Conservativity of PrO filtering relative to KF)**.** *Let* $\mathtt{m}_k$ *and* $\mathtt{P}_k$ *denote the mean and covariance associated with the* KF*, and assume both the* KF *and* KF-PrO *start from a common prior* $q_{k-1|k-1}$*, with predictive mean* $\tilde{m}_k$ *defined in* (9)*. Then*

1. $m_k = \tilde{m}_k + \tilde{P}_k H_k^\top (H_k \tilde{P}_k H_k^\top + S_k(P_k))^{-1}(y_k - H_k \tilde{m}_k)$*, i.e.* m_k *is the* KF *mean computed with the measurement covariance* R_k *replaced by* $S_k(P_k) = R_k + H_k P_k H_k^\top \succeq R_k$*. In particular, if* y_k *is scalar, then* $\|m_k - \tilde{m}_k\| \leq \|\mathtt{m}_k - \tilde{m}_k\|$*.*
2. $P_k \succeq \mathtt{P}_k$*, where* $\succeq$ *denotes the Loewner order.*

The proof of Proposition 4 is contained in Section A.4. The first part shows that the KF-PrO mean update acts as if the measurement noise were inflated by the current state uncertainty (for scalar y_k, this ensures m_k is no more sensitive to y_k than the KF mean), while the second shows that the KF-PrO covariance is always at least as large as that of the KF. Moreover, the first-order optimality condition for P_k (see Section A.4) contains a term proportional to the outer product of the innovation $v_k(\tilde{m}_k)$, so that a large innovation results in additional inflation; i.e. KF-PrO performs *data-dependent covariance inflation* relative to the KF.[1]

Figure 1 illustrates the difference in behaviour of the KF and KF-PrO in the setting of a simple state space model: $A_k = 1, R_k = \sigma^2 = 1, \Lambda_k = \epsilon^2 = 10^{-4}, H_k = 1$. The dataset in the first row is corrupted by outliers, and we observe both the conservative mean update and the data-driven covariance inflation for KF-PrO. The second row represents a misspecified dynamics model, where the true state abruptly changes after $k = 20$. Like the KF, the KF-PrO mean m_{21} is not strongly affected by the first outlier y_{21} (cf. Part 1 of Proposition 4), however unlike the KF the KF-PrO covariance P_{21} is increased (cf. Part 2 of Proposition 4). This large P_{21} enables the KF-PrO to rapidly update the filter mean m_{22} at the next step using (11), resulting in better predictions compared with KF and WoLF-IMQ for all subsequent values of k considered. Behaviour over multiple time steps, including stability, is discussed in Section C.

2.4 A PrO Extended Kalman Filter

The EKF extends the KF using local linearisation to perform approximate inference in settings where f and h are nonlinear [Maybeck, 1982]. Analogously, in this section we extend the KF-PrO to the EKF-PrO. Firstly, the **prediction step** follows the same recipe as the EKF [Särkkä and Svensson, 2023, Algorithm 7.5], linearising f around the mean from the previous time step (i.e. m_{k-1}):

$$\tilde{m}_k = f(m_{k-1}, 0),$$
$$\tilde{P}_k = \mathbf{F}_x(m_{k-1}, 0)\, P_{k-1}\, \mathbf{F}_x(m_{k-1}, 0)^\top + \mathbf{F}_\epsilon(m_{k-1}, 0)\, \Lambda_{k-1}\, \mathbf{F}_\epsilon(m_{k-1}, 0)^\top,$$

where without loss of generality we assume that $\epsilon_k \sim \mathcal{N}(0, \Lambda_{k-1})$ and $\xi_k \sim \mathcal{N}(0, R_k)$. Here $\mathbf{F}_x$ and $\mathbf{F}_\epsilon$ denote the Jacobian matrices of the dynamic model f in (1). For the **update step**, we could proceed similarly to the EKF and linearise h around the predictive mean from the previous time step (i.e. $\tilde{m}_k$). However, since the PrO update must in any case be computed numerically, we can at little additional cost linearise around a more suitable origin, namely the current estimate $m_k^\star$ of the filtering mean, analogous to the *iterated* EKF (see Gelb, 1974, and Algorithm 7.9 in Särkkä and Svensson, 2023). Thus we take

$$S_k(P) = \mathbf{H}_x(m_k^\star, 0)\, P\, \mathbf{H}_x(m_k^\star, 0)^\top + \mathbf{H}_\xi(m_k^\star, 0)\, R_k\, \mathbf{H}_\xi(m_k^\star, 0)^\top, \tag{12}$$
$$v_k(m) = y_k - h(m_k^\star, 0) - \mathbf{H}_x(m_k^\star, 0)(m - m_k^\star), \tag{13}$$

where $\mathbf{H}_x, \mathbf{H}_\xi$ refer to the Jacobian matrices of the measurement model h in (2). For a fixed linearisation point $m_k^\star$, the update step is exactly the linear-Gaussian problem of Section 2.2, so that Φ_k is convex in P (Proposition 3) and can be minimised using L-BFGS [Liu and Nocedal, 1989]; an outer loop then updates the linearisation point $m_k^\star \leftarrow m_k^\star(P_k)$ until convergence, as summarised in Algorithm 1. When h is linear a single outer iteration suffices, and Algorithm 1 reduces to computing the KF-PrO.

[1] The posterior covariance $\mathtt{P}_k$ of KF is y_k-independent, a consequence of assuming the state-space model is well-specified.

Algorithm 1 One time step of EKF-PrO

Input: m_{k-1}, P_{k-1}, y_k **Output:** m_k, P_k

compute $\tilde{m}_k$, $\tilde{P}_k$ by linearising f around m_{k-1}; set $m_k^{(0)} \leftarrow \tilde{m}_k$, $P_k^{(0)} \leftarrow \tilde{P}_k$, $t \leftarrow 0$
while `not converged` **do**
  $t \leftarrow t+1$; compute $S_k(\cdot)$ and $v_k(\cdot)$ in (12)–(13) with $m_k^\star = m_k^{(t-1)}$ ▷ linearise h
  $P_k^{(t)} \leftarrow$ L-BFGS(Φ_k), initialised at $P_k^{(t-1)}$ ▷ convex problem; Proposition 3
  $m_k^{(t)} \leftarrow m_k^\star(P_k^{(t)})$ ▷ explicit; Proposition 2
end while
return $m_k \leftarrow m_k^{(t)}$, $P_k \leftarrow P_k^{(t)}$

Method	Cost	#HP	Ref
EKF	$O(d^3)$	0	Kalman [1960]
I-EKF	$O(I\,d^3)$	0	Jazwinski [1970]
EKF-B	$O(I\,d^3)$	2	Wang et al. [2018]
EKF-IW	$O(I\,d^3)$	1	Agamennoni et al. [2012]
EKF-AF	$O(d^3)$	2	Xia et al. [1994]
WoLF-IMQ	$O(d^3)$	1	Duran-Martin et al. [2024]
WoLF-TMD	$O(d^3)$	1	Duran-Martin et al. [2024]
EKF-PrO	$O(I\,d^3)$	0	(Ours)

Table 1: Computational complexity of the update step for each baseline method. Here d is the dimension of the latent state, I is the number of iterations that are performed (for EKF-PrO, the total number of L-BFGS iterations across all linearisations), and #HP is the number of hyper-parameters to be specified.

3 Empirical Results

This section presents a range of results to demonstrate the properties of the EKF-PrO, relative to baselines described in Section 3.1. Tracking of a 2D object is considered in Section 3.2, and prediction for a chaotic dynamical system is considered in Section 3.3.

Code to reproduce our experiments can be found at `https://github.com/zheyang-shen/pro-kalman-filter`.

3.1 Baseline Methods

For baselines, we consider the standard EKF (EKF), the iterated EKF (I-EKF; Gelb, 1974) and also methods that are representative of recent state-of-the-art approaches to robust filtering: the Bernoulli filter of Wang et al. [2018] (EKF-B; an example of a *detect-and-reject* strategy, identifying outliers and then ignoring them); the inverse-Wishart filter of Agamennoni et al. [2012] (EKF-IW; an example of a *compensation-based* strategy, which anticipates heavier-than-Gaussian tails in the dataset), the adaptive fading filter of Xia et al. [1994] (EKF-AF; an example of *innovation-based covariance inflation*, which inflates the predictive covariance by a fading factor estimated from a window of recent innovations, and is thus closest in spirit to EKF-PrO), and the weighted observation likelihood filter Duran-Martin et al. [2024] (a generalised Bayesian approach, here in two variants, WoLF-IMQ and WoLF-TMD). Note that we do not compare against more sophisticated hierarchical methods nor methods based on particle filtering, because typically these do not scale well to high-dimensional state spaces; the computational complexity of EKF-PrO is designed to be comparable to the fast baseline methods, as shown in Table 1.

3.2 Tracking an Object in 2D

For our first experiment we consider the classical problem of tracking an object moving with constant velocity in 2D (see e.g., Example 8.2.1.1 in Murphy, 2023 or Example 4.5 in Särkkä and Svensson, 2023). A typical linear state-space model for this task is $f(x, \epsilon) = Ax + \Lambda^{1/2}\epsilon$ and

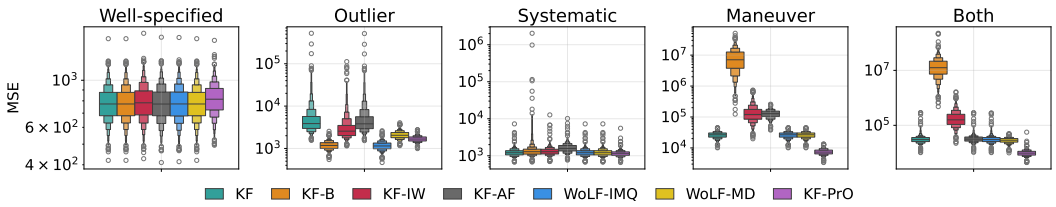

Figure 2: Tracking an object in 2D. Here the mean square error for the filtering mean of the object location is displayed for each of the methods in Table 1, with either no misspecification (**well-specified**), or **outlier**, **systematic**, **maneuver**-type, or **both** outlier and maneuver-type misspecification present in the dataset; cf. Section 3.2. Note that in this linear setting both EKF and I-EKFare identical to the classical KF. A total of 1000 replicates were performed.

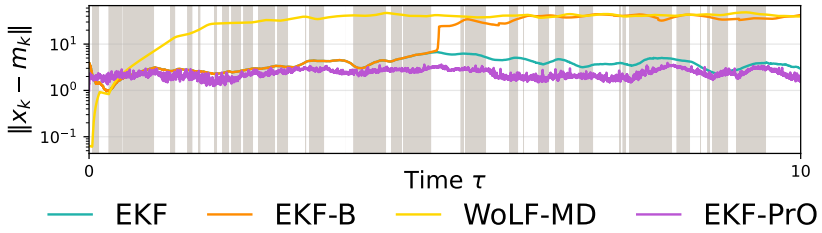

Figure 3: The misspecified Lorenz96 model. The error $\|x_k - m_k\|$ of the filtering mean is plotted against continuous time $\tau = k\Delta$; shaded regions indicate when the data-generating forcing differs from that of the model ($s_k = 2$), and white indicates no misspecification.

$h(x, \xi) = Hx + R^{1/2}\xi$, where

$$A = \begin{pmatrix} I_2 & 0.1I_2 \\ \mathbf{0} & I_2 \end{pmatrix}, \quad H = (I_2 \quad \mathbf{0}), \quad \Lambda = 0.1I_4, \quad R = 10I_2$$

with the state variable $x_k \in \mathbb{R}^4$ representing the location and velocity of the object. Five types of data are considered. In the **well-specified** setting the data are generated from the state-space model itself, so that the KF is optimal. The remaining four settings are not accurately described by the model; **systematic** misspecification draws $\epsilon_k \sim t(3)$, so that the state noise is sampled from a student-t distribution with 3 degrees of freedom, **outlier** misspecification draws $\xi_k \sim t(2.01)$, so that the measurement noise is heavier-tailed, **maneuver** introduces random piecewise constant shifts to the velocity component, and **both** combines the **maneuver** dynamics with the **outlier** measurement noise, so that both the dynamical and measurement models are misspecified. We simulated $T = 1,000$ time steps; full experimental protocol is contained in Section D.1.

In the **well-specified** setting the KF achieves the lowest positional mean square error (MSE), as expected, with KF-PrO incurring a modest loss of efficiency. When only the measurement model is misspecified (**outlier**), the data-driven approach to the mean update enables WoLF-IMQ to achieve the lowest positional MSE; however, when misspecification occurs in the dynamical model (**systematic**, **maneuver**, and **both**), KF-PrO is to be preferred. The covariance-inflating KF-AF improves slightly on the KF under **outlier**, but substantially underperforms the KF under **maneuver**; inflating the covariance allows KF-AF to bend the predicted position towards each new observation without correcting the latent velocity, while its window of past innovations introduces a lag and confuses the bias due to the maneuver with variance. Thus the performance of KF-PrO is not explained by covariance inflation alone; both KF-AF and KF-IW also enlarge the filtering covariance, yet perform substantially worse than KF-PrO under **maneuver** and **both**.

3.3 Lorenz96 Process

For our final experiment, we consider the classical nonlinear Lorenz96 model, commonly used to simulate the atmosphere [Lorenz, 1996]. The latent state $x_k \in \mathbb{R}^d$, $d = 60$, evolves according to an

Euler discretisation, with step size Δ, of the ODE $\frac{\mathrm{d}x_i}{\mathrm{d}\tau} = (x_{i+1} - x_{i-2})x_{i-1} - x_i + u_i$, $i = 1, \ldots, d$ (indices taken cyclically), driven by a forcing vector $u \in \mathbb{R}^d$, and is observed with additive Gaussian noise. The state-space model uses the canonical forcing $u = 8 \cdot \mathbf{1}$, whereas the data are generated with a forcing that switches between $u^{(1)} = 8 \cdot \mathbf{1}$ and a perturbed $u^{(2)}$ according to a hidden Markov regime $s_k \in \{1, 2\}$, so that the model is misspecified whenever $s_k = 2$. Our protocol is similar to Roth et al. [2017b]; full details are contained in Section D.2.

Interestingly, Figure 3 illustrates that outlier-robust filters alone cannot alleviate systematic misspecification, and can sometimes worsen it: `EKF-B` and `WoLF-TMD` underperform `EKF` because an "outlier" observation can be rejected if it is too far from the predicted value from the prior, however, when the outlier is caused by state-space misspecification, this leads to consecutive rejections and slower updates in the state-space. Therefore, `EKF-PrO` is particularly beneficial in cases where the precise source of model misspecification is undetermined.

4 Discussion

In applications it is rare for a state-space model to be well-specified. Existing solutions either discount observations to hedge against outliers, reducing the information available to the filter, or add degrees of freedom to the model, introducing hyper-parameters which need to be elicited or estimated. Our contribution established `EKF-PrO`, a novel hyper-parameter-free post-Bayesian filtering method whose predictions are explicitly calibrated against the data-generating process, cf. (7), suitable for situations where *both* the dynamical model and measurement model may be misspecified. Our work contributes to an emerging literature on PrO posteriors [Lai et al., 2026, Shen et al., 2025], extending their applicability to state-space models.

Limitations The main limitation of `EKF-PrO` is that it is sub-optimal when the state-space model is well-specified, in the sense of being statistically inefficient. This is inevitable when stepping outside the Bayesian framework, since Bayesian updating is optimal in the well-specified context [e.g. Zellner, 1988]; indeed, $P_k \succeq \mathtt{P}_k$ irrespective of whether the model is misspecified (Proposition 4), and the **well-specified** setting in Section 3.2 quantifies the loss of efficiency. More generally, since misspecification is detected only through the innovations, covariance inflation may be *excessive* when large innovations arise by chance, and *ineffective* when misspecification is not visible in the innovations (e.g. affecting only weakly-observed state components); see Section C. Theoretical properties of `EKF-PrO` (e.g. statistical efficiency, stability) are not yet well-understood, and will require extending existing analysis of PrO posteriors [McLatchie et al., 2025]. Finally, while all experiments ran within a few minutes on an M2 Macbook Pro, `EKF-PrO` was 5–10$\times$ slower than the extremely rapid `KF`.

Extensions Our efficient update relied on linearisation, Gaussian variational approximation, and convex optimisation; relaxing the linearity and Gaussian assumptions is possible, albeit potentially at the cost of hyper-parameters and computation. The **prediction step** could use moment matching [Maybeck, 1982], cubature [Ito and Xiong, 2002], or unscented transforms [Julier, 1996] in place of linearisation. The Gaussian **update step** can express misspecification only through an inflated covariance, rather than e.g., multimodality; Lau et al. [2026] have studied the constraints imposed by a Gaussian variational approximation in a Gaussian process setting. Particle methods [Evensen, 2003, Chazal et al., 2025] could relax this restriction, but raise practical challenges (e.g. number of particles, discretisation bias, iterations per time step) that merit a separate investigation.

Acknowledgments and Disclosure of Funding

CJO and ZS were supported by EPSRC (EP/W019590/1). CJO was supported by a Philip Leverhulme Prize (PLP-2023-004). The authors would like to thank François-Xavier Briol for feedback on an earlier draft of this manuscript, and the anonymous reviewers and area chair for their constructive feedback, in particular for pointing out the convexity of the objective in Proposition 3.

References

G. Agamennoni, J. I. Nieto, and E. M. Nebot. Approximate inference in state-space models with heavy-tailed noise. *IEEE Transactions on Signal Processing*, 60(10):5024–5037, 2012.

P. G. Bissiri, C. C. Holmes, and S. G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):1103–1130, 2016.

C. G. Boncelet and B. W. Dickinson. An approach to robust Kalman filtering. In *IEEE Conference on Decision and Control*, pages 304–305, 1983.

A. Boustati, O. D. Akyildiz, T. Damoulas, and A. Johansen. Generalised Bayesian filtering via sequential Monte Carlo. *Advances in Neural Information Processing Systems*, 34:418–429, 2020.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

P. G. Chang, G. Durán-Martín, A. Shestopaloff, M. Jones, and K. P. Murphy. Low-rank extended Kalman filtering for online learning of neural networks from streaming data. In *Conference on Lifelong Learning Agents*, pages 1025–1071. PMLR, 2023.

C. Chazal, H. Kanagawa, Z. Shen, A. Korba, and C. J. Oates. A computable measure of suboptimality for entropy-regularised variational objectives. *arXiv preprint arXiv:2509.10393*, 2025.

S.-Y. Chen. Kalman filter for robot vision: A survey. *IEEE Transactions on Industrial Electronics*, 59(11):4409–4420, 2011.

Y. Dahan, G. Revach, J. Dunik, and N. Shlezinger. Bayesian KalmanNet: Quantifying uncertainty in deep learning augmented Kalman filter. *IEEE Transactions on Signal Processing*, 73:2558–2573, 2025.

S. Das. *Robust State Estimation Methods for Robotics Applications*. PhD thesis, West Virginia University, 2023.

G. Duran-Martin, M. Altamirano, A. Y. Shestopaloff, L. Sánchez-Betancourt, J. Knoblauch, M. Jones, F.-X. Briol, and K. Murphy. Outlier-robust Kalman filtering through generalised Bayes. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

G. Duran-Martin, L. Sánchez-Betancourt, A. Shestopaloff, and K. Murphy. A unifying framework for generalised Bayesian online learning in non-stationary environments. *Transactions on Machine Learning Research*, 2025.

J. Durbin and S. J. Koopman. *Time Series Analysis by State Space Methods*. Oxford University Press, 2012.

G. Evensen. The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dynamics*, 53(4):343–367, 2003.

R. Fitzgerald. Divergence of the Kalman filter. *IEEE Transactions on Automatic Control*, 16(6): 736–747, 2003.

A. Gelb. *Applied Optimal Estimation*. MIT Press, 1974.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Y. Huang, Y. Zhang, N. Li, and J. Chambers. A robust Gaussian approximate filter for nonlinear systems with heavy tailed measurement noises. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4209–4213, 2016.

Y. Huang, Y. Zhang, N. Li, Z. Wu, and J. A. Chambers. A novel robust student's t-based Kalman filter. *IEEE Transactions on Aerospace and Electronic Systems*, 53(3):1545–1554, 2017.

Y. Huang, F. Zhu, G. Jia, and Y. Zhang. A slide window variational adaptive Kalman filter. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 67(12):3552–3556, 2020.

K. Ito and K. Xiong. Gaussian filters for nonlinear filtering problems. *IEEE Transactions on Automatic Control*, 45(5):910–927, 2002.

M. Jankowiak, G. Pleiss, and J. Gardner. Parametric Gaussian process regressors. In *International Conference on Machine Learning*, pages 4702–4712. PMLR, 2020.

A. Jazwinski. *Stochastic Processes and Filtering Theory*. Academic Press, 1970.

M. Jones, P. Chang, and K. Murphy. Bayesian online natural gradient (BONG). *Advances in Neural Information Processing Systems*, 37:131104–131153, 2024.

S. J. Julier. A general method for approximating non-linear transformations of probability distributions. 1996. URL `http://www.robots.ox.ac.uk/~siju/work/work.html`.

R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(Series D):35–45, 1960.

C. D. Karlgaard. Nonlinear regression Huber–Kalman filtering and fixed-interval smoothing. *Journal of Guidance, Control, and Dynamics*, 38(2):322–330, 2015.

R. Kelly. Reducing geometric dilution of precision using ridge regression. *IEEE Transactions on Aerospace and Electronic Systems*, 26(1):154–168, 2002.

Y.-H. Kim and S.-J. Kim. On the convexity of $\log\det(i + kx^{-1})$. *arXiv preprint arXiv:cs/0611043*, 2006.

J. Knoblauch, J. Jewson, and T. Damoulas. An optimization-centric view on Bayes' rule: Reviewing and generalizing variational inference. *Journal of Machine Learning Research*, 23(132):1–109, 2022.

V. Kuleshov, N. Fenner, and S. Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International Conference on Machine Learning*, 2018.

J. Lai, A. R. Linero, and Y. Yao. Predictive variational inference: Learn the predictively optimal posterior distribution. In *International Conference on Machine Learning*. PMLR, 2026.

N. Laird. Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73(364):805–811, 1978.

M. Lambert, S. Bonnabel, and F. Bach. The recursive variational Gaussian approximation (R-VGA). *Statistics and Computing*, 32(1):10, 2022.

C. Lau, J. Knoblauch, and L. Sharrock. Predictively oriented gaussian process posteriors, 2026. URL `https://arxiv.org/abs/2610.03201`.

D. C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1):503–528, 1989.

G. Liu. Data quality problems troubling business and financial researchers: A literature review and synthetic analysis. *Journal of Business & Finance Librarianship*, 25(3-4):315–371, 2020.

Q. Liu, M. A. Fisher, Z. Shen, K. Tant, X. Zhao, A. Curtis, and C. J. Oates. Detecting model misspecification in Bayesian inverse problems via variational gradient descent. *arXiv preprint arXiv:2512.01667*, 2025.

E. N. Lorenz. *Predictability of Weather and Climate*, chapter Predictability: A problem partly solved. Cambridge University Press, 1996.

A. Masegosa. Learning under model misspecification: Applications to variational and ensemble methods. *Advances in Neural Information Processing Systems*, 34, 2020.

P. S. Maybeck. *Stochastic Models, Estimation, and Control*. Academic Press, 1982.

Y. McLatchie, B.-E. Cherief-Abdellatif, D. T. Frazier, and J. Knoblauch. Predictively oriented posteriors. *arXiv preprint arXiv:2510.01915*, 2025.

R. Mehra. Approaches to adaptive filtering. *IEEE Transactions on Automatic Control*, 17(5):693–698, 2003.

J. W. Miller. Asymptotic normality, concentration, and coverage of generalized posteriors. *Journal of Machine Learning Research*, 22(168):1–53, 2021.

W. R. Morningstar, A. Alemi, and J. V. Dillon. PACm-Bayes: Narrowing the empirical risk gap in the misspecified Bayesian regime. In *International Conference on Artificial Intelligence and Statistics*, 2022.

K. P. Murphy. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023.

F. Nogueira. Bayesian Optimization: Open source constrained global optimization tool for Python, 2014. URL `https://github.com/bayesian-optimization/BayesianOptimization`.

H. Nurminen, T. Ardeshiri, R. Piché, and F. Gustafsson. Robust inference for state-space models with skewed measurement noise. *IEEE Signal Processing Letters*, 22(11):1898–1902, 2015.

R. Piché, S. Särkkä, and J. Hartikainen. Recursive outlier-robust filtering and smoothing for nonlinear systems using the multivariate student-t distribution. In *IEEE International Workshop on Machine Learning for Signal Processing*, 2012.

M. Roth, E. Özkan, and F. Gustafsson. A student's t filter for heavy tailed process and measurement noise. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5770–5774. IEEE, 2013.

M. Roth, T. Ardeshiri, E. Özkan, and F. Gustafsson. Robust Bayesian filtering and smoothing using Student's-t distribution. *arXiv preprint arXiv:1703.02428*, 2017a.

M. Roth, G. Hendeby, C. Fritsche, and F. Gustafsson. The ensemble Kalman filter: A signal processing perspective. *Journal on Advances in Signal Processing*, 2017:1–16, 2017b.

F. Schlee, C. Standish, and N. Toda. Divergence in the Kalman filter. *AIAA Journal*, 5(6):1114–1120, 1967.

Z. Shen, J. Knoblauch, S. Power, and C. J. Oates. Prediction-centric uncertainty quantification via MMD. In *International Conference on Artificial Intelligence and Statistics*, 2025.

R. Sheth and R. Khardon. Pseudo-Bayesian learning via direct loss minimization with applications to sparse Gaussian process models. In *Symposium on Advances in Approximate Bayesian Inference*, 2020.

S. Särkkä and L. Svensson. *Bayesian Filtering and Smoothing (2nd edition)*. Cambridge University Press, 2023.

S. Thrun. Probabilistic robotics. *Communications of the ACM*, 45(3):52–57, 2002.

J.-A. Ting, E. Theodorou, and S. Schaal. Learning an outlier-robust Kalman filter. In *European Conference on Machine Learning*, 2007.

J. d. Vilmarest and O. Wintenberger. Viking: Variational Bayesian variance tracking. *Statistical Inference for Stochastic Processes*, 27(3):839–860, 2024.

H. Wang, H. Li, J. Fang, and H. Wang. Robust Gaussian Kalman filter with outlier detection. *IEEE Signal Processing Letters*, 25(8):1236–1240, 2018.

F. Wenzel, K. Roth, B. Veeling, J. Swiatkowski, L. Tran, S. Mandt, J. Snoek, T. Salimans, R. Jenatton, and S. Nowozin. How good is the Bayes posterior in deep neural networks really? In *International Conference on Machine Learning*, 2020.

Q. Xia, M. Rao, Y. Ying, and X. Shen. Adaptive fading Kalman filter with an application. *Automatica*, 30(8):1333–1338, 1994.

A. Zellner. Optimal information processing and Bayes's theorem. *The American Statistician*, 42(4): 278–280, 1988.

Supplementary Material

This document contains supplementary material for the paper *Predictively-Oriented Kalman Filtering*. Section A contains proofs for all theoretical results in the main text. Section B explains how the **update step** of the EKF-PrO was numerically implemented. Section C discusses the behaviour of the EKF-PrO over multiple time steps, including its stability and the circumstances in which covariance inflation may be excessive or ineffective. Section D contains full details required to reproduce the empirical results in the main text.

A Proofs

This appendix contains proofs for all theoretical results in the main text. The proof of Proposition 1 is contained in Section A.1, the proof of Proposition 2 is contained in Section A.2, the proof of Proposition 3 is contained in Section A.3, and the proof of Proposition 4 is contained in Section A.4.

Remark 2 (Existence of matrix inverses)**.** *Our standing assumption in Section 2.2 that Λ_{k-1} and R_k are (symmetric) positive definite ensures that the matrices $\tilde{P}_k$ and $S_k(P)$ exist and are also (symmetric) positive definite and can hence be inverted.*

A.1 Proof of Proposition 1

Denote the density function of a Gaussian with mean μ and covariance Σ as $\mathcal{N}(\cdot;\mu,\Sigma)$. Parametrising $q \in \mathcal{Q}(\mathbb{R}^d)$ as $q = \mathcal{N}(m,P)$, we can write down the objective function in (8) as the convolution of two Gaussian densities:

$$\begin{aligned}
\mathcal{J}_k(m,P) &= -\log \int \mathcal{N}(y_k; H_k x, R_k)\mathcal{N}(x;m,P)\mathrm{d}x + \mathrm{KL}(\mathcal{N}(m,P)\|\mathcal{N}(\tilde{m}_k,\tilde{P}_k)) \\
&= -\log \mathcal{N}(0; \underbrace{y_k - H_k m}_{=v_k(m)}, \underbrace{H_k P H_k^\top + R_k}_{=S_k(P)}) + \mathrm{KL}(\mathcal{N}(m,P)\|\mathcal{N}(\tilde{m}_k,\tilde{P}_k)) \\
&= \frac{1}{2}\left[\log|S_k(P)| + \left\|S_k^{-\frac{1}{2}}(P)v_k(m)\right\|^2\right] \\
&\qquad + \frac{1}{2}\left[\mathrm{tr}\left(\tilde{P}_k^{-1}P\right) + \left\|\tilde{P}_k^{-\frac{1}{2}}(m-\tilde{m}_k)\right\|^2 + \log\frac{|\tilde{P}_k|}{|P|}\right] \\
&\overset{+C}{=} \frac{1}{2}\left[\mathrm{tr}\left(\tilde{P}_k^{-1}P\right) + \left\|\tilde{P}_k^{-\frac{1}{2}}(m-\tilde{m}_k)\right\|^2 - \log|P| + \log|S_k(P)| + \left\|S_k^{-\frac{1}{2}}(P)v_k(m)\right\|^2\right],
\end{aligned}$$

recovering the final expression in Proposition 1.

A.2 Proof of Proposition 2

For fixed P, the map $m \mapsto \mathcal{J}_k(m,P)$ is a quadratic function; cf. (10). Therefore, again with fixed P, we can find the critical point using the first-order optimality condition:

$$\nabla_m \mathcal{J}_k(m,P) = 0 \implies \tilde{P}_k^{-1}(m-\tilde{m}_k) - H_k^\top S_k^{-1}(P)(y_k - H_k m) = 0$$

Rearranging the above formula,

$$\begin{aligned}
\left(\tilde{P}_k^{-1} + H_k^\top S_k^{-1}(P)H_k\right)m &= \tilde{P}_k^{-1}\tilde{m}_k + H_k^\top S_k^{-1}(P)y_k \\
&= \tilde{P}_k^{-1}\tilde{m}_k + H_k^\top S_k^{-1}(P)(y_k - H_k\tilde{m}_k + H_k\tilde{m}_k) \\
&= \left(\tilde{P}_k^{-1} + H_k^\top S_k^{-1}(P)H_k\right)\tilde{m}_k + H_k^\top S_k^{-1}(P)(y_k - H_k\tilde{m}_k).
\end{aligned}$$

Note that this linear system is uniquely invertible from positive-definiteness of $\tilde{P}_k^{-1}$; cf. Remark 2. Solving this linear system for m, we thus obtain the prediction-correction form of the unique critical point

$$m_k^\star(P) = \underbrace{\tilde{m}_k}_{\text{predict}} + \underbrace{\left(\tilde{P}_k^{-1} + H_k^\top S_k^{-1}(P)H_k\right)^{-1} H_k^\top S_k^{-1}(P)(y_k - H_k\tilde{m}_k)}_{\text{correct}},$$

as claimed.

A.3 Proof of Proposition 3

First we note that

$$v_k(m) = y_k - H_k m = v_k(\tilde{m}_k) - H_k(m - \tilde{m}_k),$$

so that from the definition of $\mathcal{J}_k(m, P)$ in Proposition 1,

$$\mathcal{J}_k(m,P) \overset{+C}{=} \frac{1}{2}\Big[\mathrm{tr}(\tilde{P}_k^{-1}P) + (m-\tilde{m}_k)^\top \tilde{P}_k^{-1}(m-\tilde{m}_k) + \log\frac{|S_k(P)|}{|P|} \\ + (v_k(\tilde{m}_k) - H_k(m-\tilde{m}_k))^\top S_k^{-1}(P)(v_k(\tilde{m}_k) - H_k(m-\tilde{m}_k))\Big]. \tag{14}$$

Let $x := m - \tilde{m}_k$. The terms in (14) depending on x are

$$x^\top \tilde{P}_k^{-1} x + (v_k(\tilde{m}_k) - H_k x)^\top S_k^{-1}(P)(v_k(\tilde{m}_k) - H_k x).$$

Expanding,

$$x^\top \tilde{P}_k^{-1} x + v_k(\tilde{m}_k)^\top S_k^{-1}(P) v_k(\tilde{m}_k) - 2x^\top H_k^\top S_k^{-1}(P) v_k(\tilde{m}_k) + x^\top H_k^\top S_k^{-1}(P) H_k x \\ = x^\top B x - 2x^\top c + v_k(\tilde{m}_k)^\top S_k^{-1}(P) v_k(\tilde{m}_k), \tag{15}$$

where $B := \tilde{P}_k^{-1} + H_k^\top S_k^{-1}(P) H_k$ and $c := H_k^\top S_k^{-1}(P) v_k(\tilde{m}_k)$. Since $B \succ 0$, the minimiser of (15) is $x^\star = B^{-1}c$, and

$$\min_x \left(x^\top B x - 2x^\top c\right) = -c^\top B^{-1} c. \tag{16}$$

Recalling that $\Phi_k(P) = \min_m \mathcal{J}_k(m,P) = \min_x \mathcal{J}_k(x+\tilde{m}_k, P)$, the calculation in (16) establishes that

$$\Phi_k(P) \overset{+C}{=} \frac{1}{2}\Big[\mathrm{tr}(\tilde{P}_k^{-1}P) + \log\frac{|S_k(P)|}{|P|} + v_k(\tilde{m}_k)^\top S_k^{-1}(P) v_k(\tilde{m}_k) - c^\top B^{-1} c\Big]. \tag{17}$$

From the definition of B and c,

$$c^\top B^{-1} c = v_k(\tilde{m}_k)^\top S_k^{-1}(P) H_k \left(\tilde{P}_k^{-1} + H_k^\top S_k^{-1}(P) H_k\right)^{-1} H_k^\top S_k^{-1}(P) v_k(\tilde{m}_k). \tag{18}$$

Using the Woodbury matrix inversion identity,

$$S_k^{-1}(P) - S_k^{-1}(P) H_k \left(\tilde{P}_k^{-1} + H_k^\top S_k^{-1}(P) H_k\right)^{-1} H_k^\top S_k^{-1}(P) = \left(S_k(P) + H_k \tilde{P}_k H_k^\top\right)^{-1}.$$

Combining this identity with (18),

$$v_k(\tilde{m}_k)^\top S_k^{-1}(P) v_k(\tilde{m}_k) - c^\top B^{-1} c = v_k(\tilde{m}_k)^\top \left(S_k(P) + H_k \tilde{P}_k H_k^\top\right)^{-1} v_k(\tilde{m}_k). \tag{19}$$

Substituting (19) back into (17) gives

$$\Phi_k(P) \overset{+C}{=} \frac{1}{2}\Big[\mathrm{tr}(\tilde{P}_k^{-1}P) + \log\frac{|S_k(P)|}{|P|} + v_k(\tilde{m}_k)^\top M_k(P)^{-1} v_k(\tilde{m}_k)\Big], \qquad M_k(P) := S_k(P) + H_k \tilde{P}_k H_k^\top. \tag{20}$$

By the matrix determinant lemma, $|S_k(P)| = |R_k + H_k P H_k^\top| = |R_k|\,|P|\,|P^{-1} + C_k|$ where $C_k := H_k^\top R_k^{-1} H_k$, so that

$$\log\frac{|S_k(P)|}{|P|} = \log|R_k| + \log|P^{-1} + C_k|,$$

and, since $\log|R_k|$ does not depend on P, we obtain the expression for Φ_k claimed in Proposition 3.

It remains to establish convexity of each term, and existence of a minimiser. The first term $\mathrm{tr}(\tilde{P}_k^{-1}P)$ is linear in P. For the third term, the matrix-fractional function $X \mapsto v^\top X^{-1} v$ is convex on the cone of symmetric positive definite matrices [Boyd and Vandenberghe, 2004, Section 3.1.7], and $P \mapsto M_k(P)$ is affine, so their composition is convex. For the second term, let $\Sigma_k(P) := (P^{-1} + C_k)^{-1}$, so that $\log|P^{-1} + C_k| = -\log|\Sigma_k(P)|$. For any gain matrix $G \in \mathbb{R}^{d\times o}$, define the affine map

$$F_G(P) := (I - GH_k) P (I - GH_k)^\top + G R_k G^\top.$$

Completing the square gives $F_G(P) - \Sigma_k(P) = (G - G_P)S_k(P)(G - G_P)^\top \succeq 0$, where $G_P := PH_k^\top S_k(P)^{-1}$; i.e. $F_G(P) \succeq \Sigma_k(P)$ with equality at $G = G_P$ (this is the familiar Joseph form of the Kalman covariance update). Thus, for symmetric positive definite P_1, P_2, $\lambda \in [0,1]$, $P_\lambda := \lambda P_1 + (1-\lambda)P_2$, and $G := G_{P_\lambda}$,

$$\Sigma_k(P_\lambda) = F_G(P_\lambda) = \lambda F_G(P_1) + (1-\lambda)F_G(P_2) \succeq \lambda\Sigma_k(P_1) + (1-\lambda)\Sigma_k(P_2),$$

so that Σ_k is concave with respect to the Loewner order. Since $X \mapsto -\log|X|$ is both convex and decreasing with respect to the Loewner order on the cone of symmetric positive definite matrices,

$$-\log|\Sigma_k(P_\lambda)| \le -\log|\lambda\Sigma_k(P_1) + (1-\lambda)\Sigma_k(P_2)| \le -\lambda\log|\Sigma_k(P_1)| - (1-\lambda)\log|\Sigma_k(P_2)|,$$

which establishes convexity of the second term [see also Kim and Kim, 2006].

For existence of a minimiser, note that $P^{-1} + C_k \succeq P^{-1}$ implies $\log|P^{-1} + C_k| \ge -\log|P|$, and the third term is non-negative, so that $\Phi_k(P) \ge \frac{1}{2}[\mathrm{tr}(\tilde{P}_k^{-1}P) - \log|P|] + C$. The right hand side diverges both as P approaches the boundary of the cone of symmetric positive definite matrices and as $\|P\| \to \infty$, so the sublevel sets of the continuous function Φ_k are compact subsets of this cone and a minimiser exists.

A.4 Proof of Proposition 4

In the statement and proof of Proposition 4, we use $\preceq$ to denote the Loewner ordering on the cone of (real) symmetric positive semi-definite matrices; i.e. $A \preceq B$ if and only if $B - A$ is a positive semi-definite matrix.

The standard KF update equations are

$$\mathtt{P}_k^{-1} = \tilde{P}_k^{-1} + H_k^\top R_k^{-1} H_k, \tag{21}$$

$$\mathtt{m}_k = \tilde{m}_k + \mathtt{P}_k H_k^\top R_k^{-1}(y_k - H_k\tilde{m}_k) = \tilde{m}_k + \mathtt{K}_k v_k(\tilde{m}_k), \tag{22}$$

where $\mathtt{K}_k := \mathtt{P}_k H_k^\top R_k^{-1}$ is the classical Kalman gain matrix.

To establish the first claim, we use the push-through identity $(A^{-1} + B^\top C^{-1}B)^{-1}B^\top C^{-1} = AB^\top(BAB^\top + C)^{-1}$, valid for symmetric positive definite A and C. Applying this identity with $A = \tilde{P}_k$, $B = H_k$, and either $C = S_k(P_k)$ or $C = R_k$, we obtain

$$K_k(P_k) = \tilde{P}_k H_k^\top \big(H_k\tilde{P}_k H_k^\top + S_k(P_k)\big)^{-1}, \qquad \mathtt{K}_k = \tilde{P}_k H_k^\top \big(H_k\tilde{P}_k H_k^\top + R_k\big)^{-1},$$

so that $m_k = m_k^\star(P_k) = \tilde{m}_k + K_k(P_k)v_k(\tilde{m}_k)$ is the KF mean computed with R_k replaced by $S_k(P_k) = R_k + H_k P_k H_k^\top \succeq R_k$, as claimed. If y_k is scalar, then $b := H_k\tilde{P}_k H_k^\top \ge 0$, $r := R_k > 0$ and $s := S_k(P_k) \ge r$ are scalars, and

$$\|m_k - \tilde{m}_k\| = \frac{|v_k(\tilde{m}_k)|}{b+s}\|\tilde{P}_k H_k^\top\| \le \frac{|v_k(\tilde{m}_k)|}{b+r}\|\tilde{P}_k H_k^\top\| = \|\mathtt{m}_k - \tilde{m}_k\|.$$

(For vector-valued y_k, the inequality $\|m_k - \tilde{m}_k\| \le \|\mathtt{m}_k - \tilde{m}_k\|$ need not hold in general, since $X \mapsto X^{-2}$ is not monotone with respect to the Loewner order.)

To establish the second claim, recall from (20) that

$$\Phi_k(P) \overset{+C}{=} \frac{1}{2}\Big[\mathrm{tr}(\tilde{P}_k^{-1}P) - \log|P| + \log|S_k(P)| + v_k(\tilde{m}_k)^\top M_k(P)^{-1}v_k(\tilde{m}_k)\Big].$$

Computing the gradient, using the standard identities $\nabla_P \mathrm{tr}(AP) = A^\top$, $\nabla_P \log|P| = P^{-\top}$, $\nabla_P \log|f(P)| = (\partial f/\partial P)^\top f^{-\top}$, and $\nabla_P[v^\top f(P)^{-1}v] = -(\partial f/\partial P)^\top f^{-\top}vv^\top f^{-\top}$, we obtain that

$$\nabla_P\Phi_k(P) = \frac{1}{2}\Big[\tilde{P}_k^{-1} - P^{-1} + H_k^\top S_k^{-1}(P)H_k - H_k^\top M_k^{-1}(P)v_k(\tilde{m}_k)v_k(\tilde{m}_k)^\top M_k^{-1}(P)H_k\Big].$$

From Proposition 3, Φ_k attains its minimum at a point P_k in the (open) cone of symmetric positive definite matrices, at which the first-order optimality condition $\nabla_P\Phi_k(P_k) = 0$ must hold:

$$P_k^{-1} = \tilde{P}_k^{-1} + H_k^\top S_k^{-1}(P_k)H_k - H_k^\top M_k^{-1}(P_k)v_k(\tilde{m}_k)v_k(\tilde{m}_k)^\top M_k^{-1}(P_k)H_k. \tag{23}$$

The final term in (23) is a rank-one positive semi-definite matrix, and $S_k(P_k) = R_k + H_k P_k H_k^\top \succeq R_k$ implies $S_k^{-1}(P_k) \preceq R_k^{-1}$. Thus

$$P_k^{-1} \preceq \tilde{P}_k^{-1} + H_k^\top R_k^{-1} H_k = \mathtt{P}_k{}^{-1},$$

and, since matrix inversion reverses the Loewner order, $P_k \succeq \mathtt{P}_k$, as claimed. Further, (23) shows that the extent of covariance inflation depends on the data y_k through the outer product of the innovation $v_k(\tilde{m}_k)$, with larger innovations leading to additional inflation. Note also that $P_k \succeq \mathtt{P}_k$ holds even when $v_k(\tilde{m}_k) = 0$, which reflects the loss of statistical efficiency of the PrO filter relative to the KF when the model is well-specified.

B Numerical Optimisation for the Update Step

This appendix describes our implementation of Algorithm 1 for the **update step** of EKF-PrO.

Outline of the Algorithm For a fixed linearisation point $m_k^\star$ of the measurement model h, the update step reduces to the linear-Gaussian problem of Section 2.2, in which the reduced objective Φ_k is convex in P and attains its minimum (Proposition 3). We therefore compute $P_k = \arg\min_P \Phi_k(P)$ directly using `L-BFGS` [Liu and Nocedal, 1989], and then compute the mean $m_k = m_k^\star(P_k)$ explicitly using Proposition 2. For nonlinear h, an outer loop updates the linearisation point to the current filtering mean and repeats this procedure, analogous to the iterated EKF; for linear h a single outer iteration suffices. Each evaluation of Φ_k and its gradient requires Cholesky factorisations of $d \times d$ and $o \times o$ matrices, at a cost of $O(d^3 + o^3 + d^2 o)$, so that the total cost of the update is $O(I(d^3 + o^3 + d^2 o))$ where I is the total number of `L-BFGS` iterations (cf. Table 1).

Efficient Implementation via Cholesky Decomposition Direct optimisation over P requires care since P must remain a symmetric positive definite matrix. Rather than implementing constrained optimisation over the manifold of such matrices, we instead parametrise P using the lower-triangular Cholesky factor L of the whitened matrix, i.e. $P = \tilde{P}_k^{1/2} L L^\top \tilde{P}_k^{1/2}$. Our implementation applies `L-BFGS` to optimise with respect to L, whose domain is unconstrained. Although $L \mapsto \Phi_k(\tilde{P}_k^{1/2} L L^\top \tilde{P}_k^{1/2})$ is not itself convex, at any L for which the objective is finite the matrix L is invertible, and the gradient with respect to the free (lower-triangular) entries of L is the lower-triangular part of $2WL$, where $W := \tilde{P}_k^{1/2} \nabla_P \Phi_k(P) \tilde{P}_k^{1/2}$ is symmetric. If this vanishes, then WL is strictly upper triangular, so that the symmetric matrix $L^\top W L$ is also strictly upper triangular and hence zero, implying $W = 0$. It follows that every stationary point in L corresponds to a stationary point, and hence (by convexity) a global minimiser, of Φ_k.

Stopping Criterion For simplicity, we perform a fixed number `n_outer` of outer (linearisation) iterations for EKF-PrO. The stopping criterion for the inner `L-BFGS` loop is defined by 3 factors: (i) the hard ceiling of how many iterations to perform is determined by an `n_inner` parameter; (ii) we threshold the norm of the gradient $\|\nabla_L \Phi_k\|$ and upper bound this with `gtol`$=10^{-4}$; (iii) we check the "stalling" effect by checking if the relative improvement $\frac{\Phi_k(\mathtt{P}_{\text{curr}}) - \Phi_k(\mathtt{P}_{\text{new}})}{\max\{1, |\Phi_k(\mathtt{P}_{\text{curr}})|\}} \leq$ `rtol`, and stop if the above condition is achieved for `patience` number of iterations in a row. We set `rtol`$= 10^{-7}$, `patience`$= 3$. `n_outer` and `n_inner` can either be set manually or selected using Bayesian optimisation (Cf. Section D).

C Behaviour over Time and Stability

A formal analysis of the stability of EKF-PrO is beyond the scope of this work, but in this appendix we provide a qualitative discussion, in comparison to the standard KF. For simplicity we consider the linear-Gaussian setting of Section 2.2.

Error Dynamics The KF-PrO mean recursion $m_k = \tilde{m}_k + K_k(P_k)(y_k - H_k \tilde{m}_k)$, $\tilde{m}_k = A_{k-1} m_{k-1}$, has the same functional form as the KF, so that the error dynamics are governed by $(I - K_k(P_k) H_k) A_{k-1}$, and one expects the standard uniform observability and controllability conditions [Jazwinski, 1970] to remain the relevant ones for boundedness of the error. The main difference is with respect to the behaviour of the gain.

Behaviour of the Gain Within a time step, inflation of P_k reduces the gain; e.g. in the scalar case with $H_k = 1$, Proposition 4 gives

$$K_k(P_k) \;=\; \frac{\tilde{P}_k}{\tilde{P}_k + R_k + P_k},$$

which is decreasing in P_k. Across multiple time steps, the inflated P_k propagates into $\tilde{P}_{k+1} = A_k P_k A_k^\top + \Lambda_k$, and the gain is increasing in $\tilde{P}_{k+1}$. The filter therefore damps its response to the first anomalous datum and then becomes progressively more responsive if the mismatch persists, as illustrated in the bottom row of Figure 1.

Comparison of Covariances Across Time Proposition 4 compares the KF and KF-PrO covariances after a single update from a common prior. Since both the prediction step (9) and the KF covariance update $\tilde{P}_k \mapsto (\tilde{P}_k^{-1} + H_k^\top R_k^{-1} H_k)^{-1}$ are monotone with respect to the Loewner order, an induction argument shows that, if both filters are initialised with a common covariance, the KF-PrO covariance dominates the KF covariance at all subsequent times. That is, the KF-PrO is never more confident than the KF.

When is Covariance Inflation Excessive or Ineffective? Since misspecification is detected solely through the one-step-ahead predictive fit (cf. Section 2.2), covariance inflation can be *excessive* in settings where large innovations arise by chance; indeed, from Proposition 4 the KF-PrO covariance is inflated relative to the KF even in the absence of misspecification, which leads to the loss of statistical efficiency seen in the **well-specified** setting of Section 3.2. Under severe and persistent misspecification the covariance P_k can become large, so that the filtering distribution is only weakly informative; this is the intended behaviour, in that it honestly reflects the inability of the model to explain the data, but it implies that the filtering mean then closely follows the most recent observations. Conversely, covariance inflation can be *ineffective* when misspecification is not visible in the innovations; for example, when it affects only components of the state that are weakly observed through H_k, or when it takes the form of a slowly accumulating drift whose effect at each time step is small relative to the predictive dispersion $S_k(P_k)$. Moreover, since the filtering distribution is restricted to be Gaussian, forms of uncertainty that would be better represented by (for example) a multimodal distribution can only be expressed through an inflated covariance.

D Experimental Protocol

Hardware Requirement Experiments were run locally on an Apple M2 Pro machine with 16 GB unified memory using JAX/Metal.

JAX implementation. Our experimental code is implemented in JAX. The code framework builds on the third-party package `rebayes_mini` (`https://github.com/gerdm/rebayes-mini/tree/main`), which provides a lightweight implementation framework for recursive Bayesian filtering methods. We use this package for shared filtering abstractions and baseline implementations, and implement the method-specific updates, experimental configurations, and evaluation scripts used in this paper on top of this framework.

Hyper-parameter selection The hyper-parameters for different robust filters are chosen using the Bayesian Optimisation (BO) package [Nogueira, 2014]. We note that while EKF-PrO is hyper-parameter free, we can use BO to select the optimal number of iterative steps, similar to how the number of iterations in EKF-B and EKF-IW can be tuned using BO. We argue that while these hyper-parameters are tunable, they are a consequence of the optimisation procedure and do not impact the final filtering distribution. For all methods, we perform hyper-parameter tuning by minimising mean-squared error using the first trial of the data-generating distribution.

D.1 Tracking an Object in 2D

Data-generating settings In the **well-specified** setting, data are generated from the state-space model in Section 3.2. In the **outlier** and **systematic** settings, the measurement noise and the state noise, respectively, are replaced by the student-t distributions described in Section 3.2. The **maneuver**

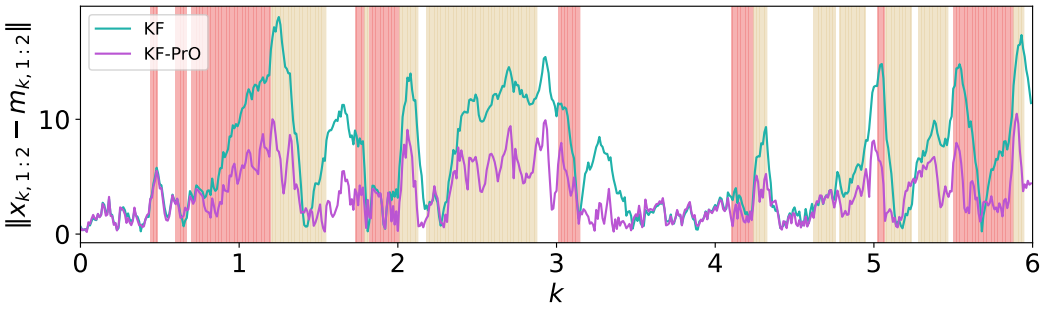

Figure 4: Example of the positional error $\|x_{k,1:2} - m_{k,1:2}\|$, as a function of time k, for KF and KF-PrO in the **maneuver** type of misspecification. The background shading denotes the existence of systematic misspecification, with white background indicating no misspecification. We can see from this plot that KF-PrO starts out similar to KF, but the propagation of overconfident predictions, and the consequent overconfidence priors for the next state leads to prolonged periods of time where KF contains much higher levels of error.

setting is described below, and the **both** setting uses the **maneuver** dynamics together with the student-t measurement noise of the **outlier** setting. The degree of freedom for the measurement noise in both **outlier** and **both** is set to 2.01; and the student-t distribution driving the dynamics model in **systematic** has degree of freedom 3.

Adaptive fading filter For KF-AF we follow Xia et al. [1994], in which the predictive covariance is multiplied by a maximum fading factor $\lambda_{\text{max}} \geq 1$ computed by comparing the empirical covariance of the innovations, over a β-exponentially decaying window of recent time steps, with the covariance predicted by the model. We choose the innovation covariance smoothing coefficient β and the maximum fading factor λ_{max} with Bayesian optimisation.

For each method and each replicate, the positional MSE is calculated as

$$\text{MSE} = \sum_{k=1}^{T} \sum_{i \in \{1,2\}} (x_{k,i} - m_{k,i})^2 = \sum_{k=1}^{T} \|x_{k,1:2} - m_{k,1:2}\|^2,$$

where $x_{k,i}$ is the ith coordinate of the true object location at time k, $m_{k,i}$ is the filtering mean for the ith coordinate of the object location at time k, and $x_{k,1:2} := (x_{k,1}, x_{k,2})$, $m_{k,1:2} := (m_{k,1}, m_{k,2})$.

Generating a maneuvering object We consider a switching linear dynamical system, identical to Example 13.4.2 from [Murphy, 2023], for tracking a maneuvering object. The state is $x_k = (x_{k,1}, x_{k,2}, \dot{x}_{k,1}, \dot{x}_{k,2})^\top$, where $(x_{k,1}, x_{k,2})$ denotes the two-dimensional position and $(\dot{x}_{k,1}, \dot{x}_{k,2})$ denotes the corresponding velocity. Conditional on a discrete regime $s_k = j \in \{1, 2, 3\}$, the state evolves according to

$$p(x_k \mid x_{k-1}, s_k = j) = \mathcal{N}(x_k \mid Ax_{k-1} + b_j, \Lambda),$$

where A and Λ are as in Section 3.2 (i.e. a time step of 0.1 and $\Lambda = 0.1I_4$), and the observations $p(y_k \mid x_k) = \mathcal{N}(y_k \mid Hx_k, R)$ are also generated as in Section 3.2. The regime-dependent bias vectors are

$$b_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \qquad b_2 = \begin{pmatrix} -1.225 \\ -0.35 \\ 1.225 \\ 0.35 \end{pmatrix}, \qquad b_3 = \begin{pmatrix} 1.225 \\ 0.35 \\ -1.225 \\ -0.35 \end{pmatrix}.$$

Thus the discrete regime controls the direction of the maneuver through the additive bias b_j. The regime s_k follows a persistent three-state Markov chain with self-transition probability 0.99, i.e. with transition matrix

$$\Pi = \begin{pmatrix} 0.99 & 0.005 & 0.005 \\ 0.005 & 0.99 & 0.005 \\ 0.005 & 0.005 & 0.99 \end{pmatrix}.$$

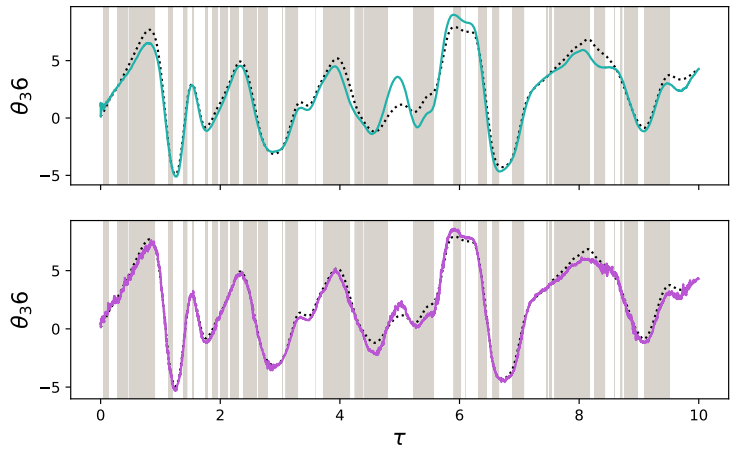

Figure 5: Tracing the mean of EKF-PrO and EKF against the ground truth state (black dotted). Even though EKF is able to capture the trajectory most of the time, it does not match EKF-PrO in accuracy, especially around $\tau = 17.5$.

The comparison between KF and KF-PrO is shown in Figure 4, where we can see the propagation of overconfident predictions leads to prolonged periods of large errors obtained by KF, similar to the results shown in Figure 3.

D.2 Lorenz96 Process

The Lorenz96 process is a $d = 60$-dimensional ordinary differential equation driven by a forcing vector $u \in \mathbb{R}^d$, for which the canonical choice is $u = 8 \cdot \mathbf{1}$, where $\mathbf{1}$ denotes a vector of all 1s. Using an Euler discretisation with step size $\Delta = 0.001$, the data-generating process is

$$x_{k+1,i} = x_{k,i} + \Delta\big[(x_{k,i+1} - x_{k,i-2})x_{k,i-1} - x_{k,i} + u_i^{(s_k)}\big], \qquad i = 1, \ldots, d,$$

with indices taken cyclically, where the regime $s_k \in \{1, 2\}$ selects between the forcing vectors $u^{(1)} = 8 \cdot \mathbf{1}$ and $u^{(2)} \sim \mathcal{N}(8 \cdot \mathbf{1}, I)$, and evolves as a Markov chain with transition matrix

$$\Pi = \begin{pmatrix} 0.995 & 0.005 \\ 0.005 & 0.995 \end{pmatrix}.$$

The state-space model instead assumes that the forcing is always canonical:

$$x_{k+1,i} \sim \mathcal{N}\left(x_{k,i} + \Delta\big[(x_{k,i+1} - x_{k,i-2})x_{k,i-1} - x_{k,i} + u_i^{(1)}\big], \Delta^2\right),$$

together with the measurement model $y_k \sim \mathcal{N}(x_k, 5I)$, which is shared by the data-generating process. We generate a total of $N = 20{,}000$ time steps. Figure 5 illustrates the effect of a misspecified nonlinear state-space model on EKF.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The methodology is introduced as claimed, and is empirically assessed as claimed.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The main limitation is the current lack of theoretical guarantees, which we flag as a direction for future work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All propositions are rigorously stated, and proofs are provided in the supplement.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The full experimental protocol is provided in the supplement.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Python code is provided to reproduce the results in the manuscript.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The full details are provided in the code and the supplement.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Monte Carlo replicates of experiments were performed and the spread of values obtained is indicated.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Our computational hardware and timings are reported in the supplement.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Every effort was made to align with the Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: No direct societal impact is envisaged, though filtering methods are widely used in diverse applications which could have societal impact.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: No risk envisaged.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The experiments are based on synthetic datasets. The JAX implementation of `EKF-PrO` is based on existing code framework `rebayes_mini` version 0.2.0, and credit it in the supplementary material with a link to the upstream repository. The upstream repository did not specify an explicit license in the version we used, so we do not redistribute or modify its source code; instead, the supplementary instructions install it as an external dependency from the upstream repository.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets`

has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: NA

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: NA

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: NA

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: NA

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.