Convergence of Bayesian Histogram Filters for Location Estimation

Avik De*, Alejandro Ribeiro*, William Moran† and Daniel E. Koditschek*

***Abstract*— We prove convergence of an approximate Bayesian estimator for the (scalar) location estimation problem by recourse to a histogram approximant. We exploit its tractability to present a simple strategy for managing the tradeoff between accuracy and complexity through the cardinality of the underlying partition. Our theoretical results provide explicit (conservative) sufficient conditions under which convergence is guaranteed. Numerical simulations reveal certain extreme cases in which the conditions may be tight, and suggest that this procedure has performance and computational efficiency favorably comparable to particle filters, while affording the aforementioned analytical benefits. We posit that more sophisticated algorithms can make such piecewise-constant representations similarly feasible for very high-dimensional problems.**

I. Introduction

We investigate the classic problem of Bayesian estimation of the posterior distribution of a scalar location parameter θ from noisy measurements

$$x_n = \theta + z_n, \tag{1}$$

where $\mathsf{E}[z_n] = 0$ and $\mathsf{E}\left[z_n^2\right] < \infty$.

Our motivation for location parameter estimation comes from the field of robotics. Implementing the estimator using a mobile robot generally introduces a specific dependency structure between the z_n (e.g. [1]), so we can neither assume they are independent nor identically distributed.

A. Brief Literature Survey

Bayes consistency in such a setting is still not a solved problem in general [2], however there are specific cases in which much is known. A classic result of Doob [3] shows consistency of Bayes estimates from i.i.d. measurements for almost all parameter values θ drawn from a known distribution. However, from the frequentist point of view where the true parameter value is fixed but unknown, this result is not satisfactory and much more work needs to be done to prove consistency even with relatively restrictive model assumptions [4]. More recent results like the Bernstein-von Mises theorem [5] prove Bayes consistency under less restrictive assumptions, but still with i.i.d. measurements. More importantly, these recent results [6], [7] reveal that the prior distribution plays a pivotal role in the convergence of the estimation algorithm. A recent summary for estimation from i.i.d. measurements can be found in [8], whereas the more complicated case of dependent measurements is investigated in [9], [10]. These results (of a theoretical statistics nature) focus on convergence of the true Bayes posterior, and are difficult to apply to physically implemented Bayes filters because accurate representation of the exact posterior is computationally intractable in high dimensional settings.

Notwithstanding theoretical characterization of the posterior's critical role in convergence of recursive Bayesian estimates, it is well known in robotics and related fields that hard computational constraints necessitate the substitution of approximate representations such as Kalman filters (with enhancements) and particle filters [11] in real application settings. These approaches come with penalties such as model restrictions (for Kalman filters) and/or uncontrollable approximation error (for particle filters). There is a huge recent literature on modifications that address algorithmically some of these issues (summarized well in [12]). For histogram filters in particular, there is some recent work [13] which numerically explores the tradeoff between representation complexity and estimator errors, as well as a higher order piecewise-polynomial (versus piecewise-constant) representation [14]. However, to the best of the authors' knowledge, heretofore there has been no analytical result — even restricted to the domain of piecewise-constant representation — governing the tradeoff between approximation error and convergence.

B. Organization and Contributions of this Paper

In this paper we focus on the same framework as [3], with the added complications of dependent measurements and explicit modeling of the approximation error. However, because of the difficult nature of the general Bayesian consistency problem, in this paper we assume the existence of an estimator which contracts the mean-squared error (MSE), and examine how perturbations due to approximation affect convergence. We show in Section II-A that this assumption is not very restrictive by comparing to a suboptimal LMMSE estimator, and giving exact conditions on the prior under which it holds.

The central contributions of this paper are (a) an explicit strategy for balancing the speed/accuracy tradeoff by controlling the approximation error in Section III, (b) a supermartingale proof of convergence of the estimate of the location parameter from dependent measurements with approximate representations in Section IV-B, (c) a computationally efficient estimation algorithm (some numerical results are presented in Section V together with comparisons to a particle filter implementation), which we posit will be suitable for implementation on even low-power robotic platforms.

*Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA, USA. {avik,aribeiro,kod}@seas.upenn.edu.

†Defence Sciences Institute, University of Melbourne Parkville, VIC, Australia. wmoran@unimelb.edu.au.

This work was supported by AFOSR MURI FA9550-10-1-0567.

We wish to obtain sufficient conditions enabling aggressive approximation (and thus, computational simplification) while still guaranteeing convergence of the estimate. As discussed in Section I-A, the classical statistics literature informs us that the Bayes estimate is sensitive to choice of prior, and that is particularly confounding because approximate representations necessarily modify the prior. Although it is true that in an idealized Bayes estimation problem, the effect of the prior distribution diminishes [5], [15], this can only be shown to hold asymptotically. To the best of our knowledge, the effect of introducing non-zero approximation error at each recursive step of a Bayes filter has not explicitly been examined in the past literature.

II. Model Assumptions

We assume that the random location parameter and measurements lie in the compact space $\mathcal{X} = \mathsf{B}(0, \Gamma) \subset \mathbb{R}$, where $\mathsf{B}(0, \Gamma)$ is the closed ball of radius Γ around 0. Define $\mathcal{P}$ as the space of distributions on $\mathcal{X}$. The parameter to be estimated is sampled from $\theta \sim \pi$, a known prior distribution which is non-informative for location parameters [16]. Let $\mathcal{Y} := \mathcal{X} \times \mathcal{X}$ denote the joint measurement–parameter space.

We define the distribution for z_n in (1) as an unbiased measurement model for x_n conditioned on θ, $\ell_n(x \mid \theta)$. We assume that the measurements are predictive[1] (for Lemma 7), unbiased, and with finite second moment σ_n^2. We label the σ-field generated by the first n measurements $\mathcal{F}_n := \sigma(x_1, \ldots, x_n)$. We assume that σ_n is measurable from $\mathcal{F}_{n-1}$, and that $\sigma_i \leq \sigma_{\max}$ for each i.

Lastly, for the approximation operation of Section III, we make the regularity assumptions on the measurement distribution

1) ℓ_n is C^1 with derivative $\frac{\partial}{\partial x}\ell_n(x, \theta)$, and
2) $\max \left| \frac{\partial}{\partial x}\ell_n(x, \theta) \right| \leq \Upsilon < \infty$.

The conditions above are satisfied by many families of measurement distributions that are commonly encountered in location parameter estimation [17]. Moreover, there is substantial evidence in the literature for received-signal-strength (RSS) wireless sensor models suffering from standard error [1], [18] or variance [19] proportional to the sensing range, thus motivating our decision to not assume independent measurements.

A. MSE Contracting Estimator

As stated in Section I-B, we also assume the existence of an estimator which strictly contracts the MSE (under some conditions on the prior which are derived below). In particular, let $p \in \mathcal{P}$ be an arbitrary prior distribution for θ. Define $\widehat{\theta} = \int_{\mathcal{X}} \theta \mathrm{d}p$, $\mathsf{Pr}(\theta \mid x) \propto \ell(x \mid \theta) p(\theta)$ as the Bayes posterior, and $\widehat{\theta}^+ = \int_{\mathcal{X}} \theta \mathsf{Pr}(\theta \mid x) \mathrm{d}\theta$. Then we require that

$$\mathsf{E}\left[(\widehat{\theta}^+ - \theta)^2\right] \leq \alpha \mathsf{E}\left[(\widehat{\theta} - \theta)^2\right], \tag{2}$$

where $\alpha < 1$. Note that the left- and right-hand sides of (2) are simply $\mathsf{E}\left[\mathsf{Var}[\theta \mid x]\right]$ and $\mathsf{Var}[\theta]$ respectively, and we know from the law of total variance [20] that $\alpha \leq 1$ automatically.

[1] Predictive [15] is a mild condition requiring that the predictive posterior distribution $\Pr(x_{n+1}, \ldots \mid x_1, \ldots, x_n)$ exists.

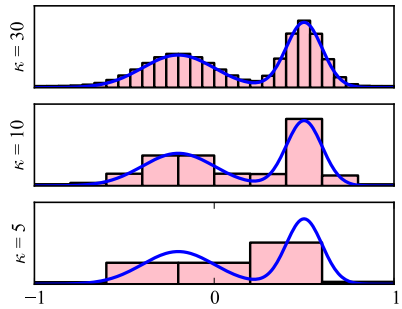

Fig. 1. Illustration of the approximation operator $\mathcal{A}_\kappa$ presented in Section III, showing the tradeoff between cell cardinality, κ, and the approximation error. Here $\mathcal{X} = [-1, 1]$.

In general it is difficult to make claims about posterior variance in Bayesian estimation.[2] We prefer to provide sufficient conditions for (2) by comparing to the suboptimal LMMSE estimator. It is known [22] that even for a nonlinear/nongaussian estimation problem, the LMMSE estimator for (1) attains a bound that only depends on the first and second moments of the joint $\ell(x \mid \theta)p(\theta)$ distribution,

$$\mathsf{lmmse} = \mathsf{E}\left[\theta^2\right] - \mathsf{E}\left[x\theta\right] \left(\mathsf{E}\left[x^2\right]\right)^{-1} \mathsf{E}\left[\theta x\right].$$

Here, $\mathsf{E}\left[x\theta\right] = \mathsf{E}_\theta[\theta \mathsf{E}_{x|\theta}[x]] = \mathsf{E}\left[\theta^2\right]$ using the law iterated expectations. Together with $\mathsf{E}\left[\theta\right] = 0$, we get the condition (on p)

$$\mathsf{Var}[\theta] \geq \sigma_{\max}^2 \left(\frac{1}{\alpha} - 1\right), \tag{3}$$

to ensure $\mathsf{lmmse} \leq \alpha \mathsf{Var}[\theta]$. By definition of MMSE, it must attain a MSE at least as good, thus satisfying (2).

III. Approximate Representations

For approximating our belief state, we wish to have (a) finite parameterizability for computational efficiency, (b) a functional form for the posterior to enable computation of statistics such as moments and entropy, and (c) a bound on the total variation distance between the original and approximated distributions (informed by the proof in Section IV-B).

Historically, Stein's method has seen many applications in finding approximations of probability distributions among specific parameterized families such as Poisson [23], geometric [24] and mixtures of similar families [25]. However, we wish to not commit to a specific family of distributions and inadvertently impose constraints such as unimodality or any other structure through rigidity of representation. Early work on non-parametric approximations from data [26] is also not applicable because of how much data is required to get an estimate.

[2] With normal measurements and general classes of priors, it is possible to compute the posterior moments exactly if there is exact knowledge of the first and second derivatives of the marginal distribution [21], but in our case the prior is arbitrary enough to not permit a similar analysis.

A. Histogram Distributions

Our preferred family of approximations is the set of "histograms" — piecewise constant distributions — because of the ease of representation and computation together with the ability to control the "coarseness" of the approximation by picking the cardinality of the partition. The authors of [27] examine histograms as a means of density estimation from data and provide conditions on bounding the $\mathcal{L}_2$-norm of the error. In this paper we use histograms to directly approximate densities in $\mathcal{P}$ and derive bounds for the approximation error in the form of the $\mathcal{L}_1$-norm of the residual (and thus, total variation distance [28]).

Definition 1 (Uniform Partition)**.** For given $\kappa \in \mathbb{Z}_+$, consider a collection of evenly-spaced boundary points $\xi_i := \varphi(i/\kappa)$ where $\varphi : x \mapsto \Gamma(-1+2x)$ maps $[0,1] \mapsto \mathcal{X}$. Define

$$\mathcal{R}_i = [\xi_{i-1}, \xi_i) \tag{4}$$

for $i \in \{1, \ldots, \kappa\}$, and note that $\cup_i \mathcal{R}_i = \mathcal{X}$ (ignoring the zero-measure end-point) is a κ-cell partition.

Definition 2 (Histogram)**.** Define the space of histogram distributions as $\mathcal{Q} \subset \mathcal{P}$. A κ-cell histogram distribution $q \in \mathcal{Q}$ is defined by a set of levels $(q_i)_{i=1}^{\kappa} \in \mathbb{R}^{\kappa}_{\geq 0}$ together with a κ-cell partition $(\mathcal{R}_i)$ such that

$$q(\theta) = \sum_{i=1}^{\kappa} \mathbb{I}\left(\theta \in \mathcal{R}_i\right) q_i, \tag{5}$$

where $\mathbb{I}\left(\cdot\right)$ is an indicator function.

Let us define $\rho \ll 1$ as a desired minimum mass for each cell.[3] Also, define

$$\delta(\kappa) := \tfrac{2\Gamma}{\kappa} \tag{6}$$

as the size of a cell in a κ-cell partition.

For our purposes of using approximate representations in a Bayes filter (Section IV-A), we only need to restrict our attention to a special subset of all possible distributions.

Definition 3 (Approximable Distributions)**.** Given arbitrary κ, define the set of approximable distributions, $\mathcal{P}_\kappa$, as the set of $p \in \mathcal{P}$ having the properties

1) p is piecewise C^1 with some points of discontinuity that coincide with cell boundaries of the κ-cell partition, and
2) $\max |p'| \leq \frac{\Upsilon}{\rho} < \infty$.

Note that distributions that are C^1 are in $\mathcal{P}_\kappa$ for each κ. In Section IV-A.2 we provide proof that these assumptions are enforced automatically by our usage in Section IV.

Definition 4 (Approximation)**.** Given arbitrary κ and $p \in \mathcal{P}_\kappa$, we define an approximation operator $\mathcal{A}_\kappa : \mathcal{P}_\kappa \to \mathcal{Q}$ that finds a κ-cell histogram approximation $q = \mathcal{A}_\kappa(p)$ by

1) forming a κ-cell uniform partition $\cup_i \mathcal{R}_i$, and
2) setting $q_i = \max\{p(\overline{\mathcal{R}}_i), \frac{\rho}{2\Gamma}\}$, where $\overline{\mathcal{R}}_i$ is the centroid of $\mathcal{R}_i$.

Figure 1 demonstrates this simple algorithm operating on a smooth distribution in $\mathcal{P}$.

Proposition 5. *Let $p \in \mathcal{P}_\kappa$ and $q = \mathcal{A}_\kappa(p)$. Then, the approximation has the properties*

1) $\|p - q\|_1 = \int_{\mathcal{X}} |(p - \mathcal{A}_\kappa(p))(\theta)| \, \mathrm{d}\theta \leq \frac{2\Gamma^2\Upsilon}{\rho\kappa}$,
2) $\int_{\mathcal{X}} |\theta(p - q)(\theta)| \, \mathrm{d}\theta \leq \delta(\kappa) + \rho\Gamma^2$.

Proof. We prove the claims in sequence:

1) Our histogram is reminiscent of a Riemann sum approximation. With a little plane geometry we can easily obtain the error bound

$$\|p - q\|_1 \leq \sum_{i=1}^{\kappa} \frac{\delta(\kappa)^2}{2} \max_{\theta \in \mathcal{R}_i} |p'(\theta)| \leq \frac{2\Gamma^2\Upsilon}{\rho\kappa},$$

where we used the fact that the derivative p' is well-defined and bounded in the interior of $\mathcal{R}_i$, per the conditions assumed from Definition 3.

2) In the rule for setting q_i in Definition 4, note that we can write the approximation as the sum of a linear operator[4] $(Gp)(\theta) = p(\overline{\mathcal{R}}_i)$ for $\theta \in \mathcal{R}_i$, and a nonlinear "error component" of bounded magnitude $|(h(p))(\theta)| \leq \frac{\rho}{2\Gamma}$, i.e. for each $\theta \in \mathcal{X}$,

$$(\mathcal{A}_\kappa(p))(\theta) = (Gp)(\theta) + (h(p))(\theta).$$

Define id as the identity map. The residual error can be written as $\left\langle \theta, (\mathrm{id} - \mathcal{A}_\kappa)(p) \right\rangle_{\mathcal{L}_2}$, an $\mathcal{L}_2$-inner product with respect to measure θ. Using the Hölder inequality, the linear component of the decomposition above is

$$\begin{aligned} \left\langle \theta, (\mathrm{id} - G)(p) \right\rangle_{\mathcal{L}_2} &= \left\langle (\mathrm{id} - G)(\theta), p \right\rangle_{\mathcal{L}_2} \\ &\leq \|(\mathrm{id} - G)(\theta)\|_\infty \|p\|_1 \\ &= \|(\mathrm{id} - G)(\theta)\|_\infty = \delta(\kappa). \end{aligned}$$

Here $(\mathrm{id} - G)(\theta)$ cancels the "constant" component in each cell, and since the θ function has slope 1, the contribution of each cell is the size of the cell, $\delta(\kappa)$. The $\mathcal{L}_\infty$-norm picks out the maximum value of its argument, and so the result is just $\delta(\kappa)$.
The nonlinear component is bounded by

$$\int_{\mathcal{X}} \left|\theta h(p(\theta)) \mathrm{d}\theta\right| \leq \|h(p)\|_\infty \|\theta\|_1 = \rho\Gamma^2.$$

While not explicitly stated, note that $\kappa < \infty$, i.e. we can always find a finite-dimensional arbitrarily close approximation. □

[3] The intuition behind this is "Cromwell's Rule" [29]; it also helps us establish bounds in Proposition 6.

[4] To see the linearity of G, suppose $p_1, p_2 \in \mathcal{P}$. Note that G creates the same partition $\mathcal{R}_i$ irrespective of its operand, and so $G(p_1 + p_2)$ has the same cell structure as $G(p_i)$. The level in cell i of $G(\cdot)$ is set by simply looking evaluating the operand at θ_i, the centroid of $\mathcal{R}_i$. Since $p_1(\theta_i) + p_2(\theta_i) = (p_1 + p_2)(\theta_i)$, the superposition property holds for G. Similarly, the cell structure is invariant to scaling the operand, and $\lambda \cdot p(\theta_i) = (\lambda p)(\theta_i)$. (Technically, scaling a distribution by anything other than unity while staying within $\mathcal{P}$ is not possible; we still ensure that it does not affect the linearity of G as an operator.)

The approximation procedure above is not "optimal" in any way, but an optimal partitioning scheme (such as a generalized Lloyd's procedure [30] with $\mathcal{L}_1$-norm objective) is readily applicable to the theoretical framework presented in this paper.

IV. A Bayes Filter using Histograms

Histogram filters have seen some use in robotics for localization [17], [31] due to their computational efficiency. However, they do not offer an easy way to guarantee convergence due to information loss in the approximation step. We implement a similar filter here, but with the goal in mind of controlling approximation error in order to be able to guarantee almost sure convergence.

Define the following time-indexed (random) quantities:

- Prior distributions $\widehat{p}_0 = \widetilde{p}_0 := \pi$.
- Unknown true posterior $\widehat{p}_n(\theta) := \mathsf{Pr}(\theta \mid x_1, \ldots, x_n)$.
- Measurement-updated belief (calculated using Bayes rule) $\widetilde{p}_n^-(\theta) \propto \ell_n(x_n \mid \theta)\widetilde{p}_{n-1}(\theta)$.
- κ_n-cell approximated belief $\widetilde{p}_n := \mathcal{A}_{\kappa_n}(\widetilde{p}_n^-)$.
- MMSE estimators $\widehat{\theta}_n := \int_{\mathcal{X}} \theta \mathrm{d}\widehat{p}_n$, $\widetilde{\theta}_n := \int_{\mathcal{X}} \theta \mathrm{d}\widetilde{p}_n$, and $\widetilde{\theta}_n^- := \int_{\mathcal{X}} \theta \mathrm{d}\widetilde{p}_n^-$.

For brevity, we denote $\Delta\widehat{\theta}_n := \widehat{\theta}_n - \theta$, $\Delta\widetilde{\theta}_n := \widetilde{\theta}_n - \theta$, and $\Delta\widetilde{\theta}_n^- := \widetilde{\theta}_n^- - \theta$ for each n.

A. Implementation with Histograms

1) Conditions on the Approximation: For the purposes of our convergence proof in Section IV-B, we require for an approximated distribution $\widetilde{p}_n \in \mathcal{Q}$ that the variance has a lower-bound given by (3). This does not impose any additional constraints, because our stopping condition in Theorem 11 is already a threshold for the posterior variance, and we simply need to ensure that this additional condition is met at every iteration.

2) Partition Refinement: For our implementation in this paper, we restrict ourselves to histogram belief distributions which are refinements (in the sense of its partition) of previous beliefs. If κ_0 is the initial cell cardinality, we choose to double the cell count every m steps, i.e.

$$\kappa_i = 2^{\lfloor i/m \rfloor}\kappa_0. \tag{7}$$

This strategy ensures that the cell boundary points of a κ_n-cell partition automatically contain the cell boundary points of a κ_{n-1}-cell partition. This and Proposition 6 below together ensure that both conditions assumed in Definition 3 are automatically enforced.

Proposition 6. *Summed over the interiors of the cell where the derivative is defined,*

$$\frac{\sum_i \max_{\theta\in\mathcal{R}_i} |(\widetilde{p}_n^-)'|}{\kappa_{n-1}} \leq \frac{\Upsilon}{\rho}.$$

Proof. From the definition of $\widetilde{p}_n^-$, note that for each cell i,

$$\begin{aligned}\max_{\theta\in\mathcal{R}_i} |(\widetilde{p}_n^-)'| &= \frac{(\widetilde{p}_{n-1})_i \max_{\theta\in\mathcal{R}_i} \left|\frac{\partial}{\partial\theta}\ell_n(x_n,\theta)\right|}{\sum_j (\widetilde{p}_{n-1})_j \int_{\mathcal{R}_j} \ell_n(x_n \mid \theta)\mathrm{d}\theta} \\ &\leq \frac{(\widetilde{p}_{n-1})_i \Upsilon}{\sum_j (\widetilde{p}_{n-1})_j \int_{\mathcal{R}_j} \ell_n(x_n \mid \theta)\mathrm{d}\theta},\end{aligned}$$

because $(\widetilde{p}_{n-1})' = 0$ in the interior of a cell. Note that for the histogram $\widetilde{p}_{n-1} \in \mathcal{P}_{\kappa_{n-1}}$,

$$\sum_i (\widetilde{p}_{n-1})_i \int_{\theta\in\mathcal{R}_i} \mathrm{d}\theta = 1 \implies \sum_i (\widetilde{p}_{n-1})_i = \frac{1}{\delta(\kappa_{n-1})}.$$

Summing $\max_{\theta\in\mathcal{R}_i} |(\widetilde{p}_n^-)'|$ over all cells,

$$\begin{aligned}\frac{\sum_i \max_{\theta\in\mathcal{R}_i} |(\widetilde{p}_n^-)'|}{\kappa_{n-1}} &\leq \frac{\Upsilon}{2\Gamma \sum_j (\widetilde{p}_{n-1})_j \int_{\mathcal{R}_j} \ell_n(x_n \mid \theta)\mathrm{d}\theta} \\ &\leq \frac{\Upsilon}{\rho \int_{\mathcal{X}} \ell_n(x_n \mid \theta)\mathrm{d}\theta} = \frac{\Upsilon}{\rho},\end{aligned}$$

using the assumption from Section III-A that we have a minimum mass ρ in each cell.[5] The last equality above can be seen by changing coordinates back to z_n in (1), wherein (with our assumption of measurable variance) the integral just becomes $\int_{\mathcal{X}} \mathrm{Pr}(z_n)dz_n = 1$. □

Choosing the refinement (7) also results in a summable approximation error series. As in Proposition 5, the actual cost of approximation at step i is bounded by $\varepsilon_i \leq \frac{2\Gamma^2\Upsilon}{\rho\kappa_i}$, and so

$$\sum_{i=1}^{n} \frac{1}{\kappa_i} \leq 2m \implies \sum_{i=1}^{n} \varepsilon_i \leq \frac{2\Gamma^2\Upsilon m}{\rho\kappa_0} =: \varepsilon_{\lim}. \tag{8}$$

This strategy guarantees that the errors are summable, while giving us some degree of control as to how often we refine the partition.[6]

B. Convergence Proof using Supermartingales

For the proof of convergence, we use the supermartingale convergence theorem [32]. This method does not require independence between successive (in time) stochastic quantities, and lets us focus on incremental approximation errors.

The main result in this section is Theorem 11, which shows that under the conditions we impose, the MSE converges for our algorithm. Lemmas 8, 9 and 10 show how different constituents of the error signal are controlled, and Lemmas 7 establishes bounds that help control these errors.

Lemma 7. *The accumulated approximation error at each time n obeys the bound $\mathsf{E}\left[\|\widehat{p}_n - \widetilde{p}_n\|_1 \mid \mathcal{F}_{n-1}\right] \leq \varepsilon_{\lim}$, as defined in (8).*

Proof. We assert the subclaim

$$\mathsf{E}\left[\|\widehat{p}_n - \widetilde{p}_n^-\|_1 \mid \mathcal{F}_{n-1}\right] \leq \|\widehat{p}_{n-1} - \widetilde{p}_{n-1}\|_1.$$

Here $\widehat{p}_{n-1}$ and $\widetilde{p}_{n-1}$ can be thought of as differing opinions in the sense of [15]. As in Section II, our measurements are predictive, and as mentioned in Section IV-A, we ensure that $\widehat{p}_{n-1}$ is absolutely continuous with respect to $\widetilde{p}_{n-1}$.

[5] We provide a very conservative bound in this proof, but observe that the denominator would be 1 if the prior were uniform and non-informative (i.e. $(\widetilde{p}_{n-1})_j = \frac{1}{2\Gamma}$), and with informative measurements, intuitively > 1.

[6] We typically set $m \approx 40$; we attain convergence in the order of m steps and the cell cardinality never "blows up".

Following the martingale argument in the proof in [15], we observe that the expected total variation distance d_{tv} remains the same after one measurement, i.e.

$$\mathsf{E}\left[d_{\mathrm{tv}}(\widehat{p}_n, \widetilde{p}_n^-) \mid \mathcal{F}_{n-1}\right] = d_{\mathrm{tv}}(\widehat{p}_{n-1}, \widetilde{p}_{n-1}).$$

The same property holds for the $\mathcal{L}_1$-norm because of their simple algebraic relation [28]. This shows the subclaim.

Repeatedly using the triangle inequality,

$$\begin{aligned}\mathsf{E}\left[\|\widehat{p}_n - \widetilde{p}_n\|_1\right] &\leq \mathsf{E}\left[\|\widehat{p}_n - \widetilde{p}_n^-\|_1\right] + \mathsf{E}\left[\|\widetilde{p}_n^- - \widetilde{p}_n\|_1\right] \\ &\leq \mathsf{E}\left[\|\widehat{p}_{n-1} - \widetilde{p}_{n-1}\|_1\right] + \varepsilon_n \\ &\leq \ldots \leq \|\pi - \pi\|_1 + \sum_{i=1}^{n} \varepsilon_i \leq \varepsilon_{\mathrm{lim}},\end{aligned}$$

where ε_i and $\varepsilon_{\mathrm{lim}}$ are as in (8). □

Lemma 8. *The incremental (mean-squared) error added by the approximation at time n is bounded by*

$$\mathsf{E}\left[(\Delta\widetilde{\theta}_n)^2 \mid \mathcal{F}_{n-1}\right] \leq \mathsf{E}\left[(\Delta\widetilde{\theta}_n^-)^2 \mid \mathcal{F}_{n-1}\right] + \gamma_n,$$

where $\gamma_n := \delta(\kappa_n)^2 + \rho\Gamma^2 + (2\Gamma)\varepsilon_{\mathrm{lim}}$, and $\delta(\cdot)$ is defined in (6).

Proof. Note that

$$\begin{aligned}\mathsf{E}\left[\Delta\widetilde{\theta}_n^- \mid \mathcal{F}_{n-1}\right] = \int_{\mathcal{Y}} \Delta\widetilde{\theta}_n^- \ell_n(x \mid \theta)\widetilde{p}_{n-1}(\theta)\mathrm{d}x\mathrm{d}\theta + \\ \int_{\mathcal{Y}} \Delta\widetilde{\theta}_n^- \ell_n(x \mid \theta)(\widehat{p}_{n-1}(\theta) - \widetilde{p}_{n-1}(\theta))\mathrm{d}x\mathrm{d}\theta,\end{aligned}$$

where the first summand is 0 (MMSE is unbiased), and the second summand can be upper-bounded by

$$\left\|\int_{\mathcal{X}} \Delta\widetilde{\theta}_n^- \ell_n(x \mid \theta)\mathrm{d}x\right\|_\infty \|\widehat{p}_{n-1} - \widetilde{p}_{n-1}\|_1 \leq 2\Gamma\varepsilon_{\mathrm{lim}}.$$

Also, using Proposition 5,

$$\left|\widetilde{\theta}_n - \widetilde{\theta}_n^-\right| = \left|\int_{\mathcal{X}} \theta\left(\widetilde{p}_n^- - \mathcal{A}_{\kappa_n}(\widetilde{p}_n^-)\right)\mathrm{d}\theta\right| \leq \delta(\kappa_n) + \rho\Gamma^2.$$

Expanding the square and using the above,

$$\begin{aligned}\mathsf{E}\left[(\Delta\widetilde{\theta}_n)^2 \mid \mathcal{F}_{n-1}\right] \leq \mathsf{E}\left[(\Delta\widetilde{\theta}_n^-)^2 \mid \mathcal{F}_{n-1}\right] + \\ \delta(\kappa_n)^2 + \rho\Gamma^2 + 2\Gamma\varepsilon_{\mathrm{lim}},\end{aligned}$$

as required to prove. □

Lemma 9. *The improvement due to the new measurement at step n is*

$$\mathsf{E}\left[(\Delta\widetilde{\theta}_n^-)^2 \mid \mathcal{F}_{n-1}\right] \leq \alpha\mathsf{E}\left[(\Delta\widetilde{\theta}_{n-1})^2\right] + (2\Gamma)^2\varepsilon_{\mathrm{lim}},$$

where $\varepsilon_{\mathrm{lim}}$ is defined in (8).

Proof. For

$$\mathsf{E}\left[(\Delta\widetilde{\theta}_n^-)^2 \mid \mathcal{F}_{n-1}\right] = \int_{\mathcal{Y}} (\Delta\widetilde{\theta}_n^-)^2 \ell_n(x \mid \theta)\widehat{p}_{n-1}(\theta)\mathrm{d}x\mathrm{d}\theta,$$

we decompose $\widehat{p}_{n-1}(\theta) = \widetilde{p}_{n-1}(\theta) + (\widehat{p}_{n-1}(\theta) - \widetilde{p}_{n-1}(\theta))$. The integral with $\widetilde{p}_{n-1}(\theta)$ satisfies

$$\begin{aligned}&\int_{\mathcal{Y}} (\Delta\widetilde{\theta}_n^-)^2 \ell_n(x \mid \theta)\widetilde{p}_{n-1}(\theta)\mathrm{d}x\mathrm{d}\theta \\ &\leq \alpha\int_{\mathcal{Y}} (\Delta\widetilde{\theta}_{n-1})^2 \ell_n(x \mid \theta)\widetilde{p}_{n-1}(\theta)\mathrm{d}x\mathrm{d}\theta \\ &\leq \alpha\int_{\mathcal{Y}} (\Delta\widetilde{\theta}_{n-1})^2 \ell_n(x \mid \theta)\widehat{p}_{n-1}(\theta)\mathrm{d}x\mathrm{d}\theta = \alpha\mathsf{E}\left[(\Delta\widetilde{\theta}_{n-1})^2\right]\end{aligned}$$

because of the assumption (2), and using the fact that $\widetilde{\theta}_{n-1}$ is the MMSE (and must attain the smallest MSE) with respect to $\widetilde{p}_{n-1}$.

For the integral with $(\widehat{p}_{n-1}(\theta) - \widetilde{p}_{n-1}(\theta))\mathrm{d}\theta$, we just use Hölder's inequality, to upper bound it by

$$\left\|\int_{\mathcal{X}} (\Delta\widetilde{\theta}_n^-)^2 \ell_n(x \mid \theta)\mathrm{d}x\right\|_\infty \|\widehat{p}_{n-1} - \widetilde{p}_{n-1}\|_1 \leq (2\Gamma)^2\varepsilon_{\mathrm{lim}}.$$

Putting these together, we obtain the result. □

Lemma 10. *At each time n, the random sequence of approximated (mean-squared) errors obeys*

$$\mathsf{E}\left[(\Delta\widetilde{\theta}_n)^2 \mid \mathcal{F}_{n-1}\right] \leq \alpha\mathsf{E}\left[(\Delta\widetilde{\theta}_{n-1})^2\right] + \tau_n,$$

where $\tau_n := \delta(\kappa_n)^2 + \rho\Gamma^2 + ((2\Gamma) + (2\Gamma)^2)\varepsilon_{\mathrm{lim}}$, $\delta(\cdot)$ is defined in (6) and $\varepsilon_{\mathrm{lim}}$ is defined in (8).

Proof. The proof follows directly from Lemmas 8 and 9. Observe that the τ_i are bounded, and let $\tau = \max_n \tau_n$. □

With the help of the lemmas above, we can state the following theorem that proves covergence of the approximate Bayesian filter proposed above to a "small" limiting MSE.

Theorem 11. *The Bayesian estimate using the approximate beliefs converges in MSE to a neighborhood of 0,*

$$\mathsf{E}\left[(\Delta\widetilde{\theta}_n)^2\right] \to \mathsf{B}\left(0, \tfrac{\tau}{1-\alpha}\right).$$

Proof. Define $M_n := \mathsf{E}\left[(\Delta\widetilde{\theta}_n)^2\right]$. From Lemma 10, we know that $\mathsf{E}\left[M_n \mid \mathcal{F}_{n-1}\right] \leq \alpha\mathsf{E}\left[M_{n-1}\right] + \tau$. Now define

$$S_n := M_n \cdot \mathbb{I}\left(M_m > \tfrac{\tau}{1-\alpha} \;\; \forall\; m < n\right),$$

and observe that $S_i \geq 0$ and

$$\mathsf{E}\left[S_n \mid \mathcal{F}_{n-1}\right] \leq \alpha S_{n-1}.$$

Using the supermartingale convergence theorem [32], we know that S_n converges to some limiting random variable, S_∞. This also means $\mathsf{E}\left[S_n\right] \to \mathsf{E}\left[S_\infty\right]$, and we can iterate the inequality above to observe that

$$\mathsf{E}\left[S_n\right] \leq \alpha\mathsf{E}\left[S_{n-1}\right] \leq \alpha^2\mathsf{E}\left[S_{n-2}\right] \leq \cdots \leq \alpha^n\mathsf{E}\left[S_0\right].$$

Taking limit $n \to \infty$, we see that $\mathsf{E}\left[S_\infty\right] = 0$. Since the random variable $S_\infty \geq 0$, we conclude by writing out the expectation that

$$\Pr(S_\infty > 0) = 0,$$

i.e. we have almost sure convergence to 0.

Now we can conclude that $M_n \to \mathsf{B}\left(0, \tfrac{\tau}{1-\alpha}\right)$. □

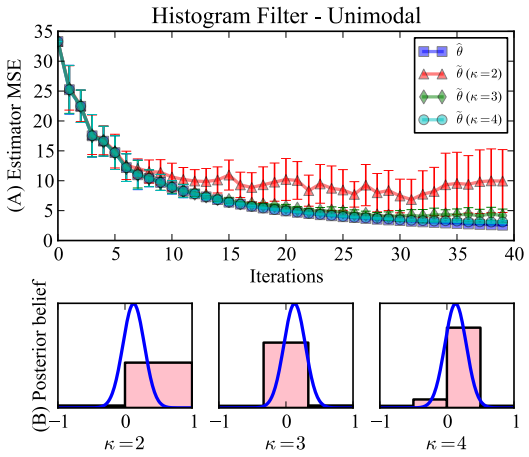

Fig. 2. Performance of the histogram filter for a unimodal measurement model, showing in (A) that the MSE of the estimator approaches the (best attainable) MSE, $\mathsf{E}\left[(\Delta\widehat{\theta})^2\right]$ (blue curve with square markers), for all but the crudest representations (e.g. $\kappa = 2$). (B) shows posterior distributions for one of the trials presented in (A), with the true posterior in solid blue, and the approximated belief distributions presented as the shaded red histogram.

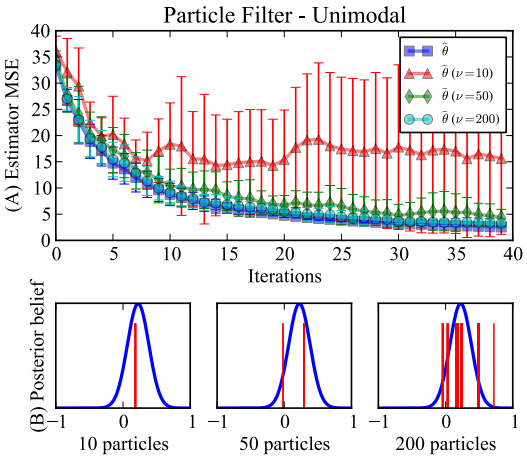

Fig. 4. Performance of a simple particle filter [17] with resampling at every iteration, demonstrating that a comparatively large number of particles is needed to attain similar MSE to the crude histograms of Fig. 2. Plot (B) shows the true posterior distribution in blue, and red vertical bars representing the samples that comprise the particle filter representation of the posterior.

The coefficients controlling the error radius $\tau/(1-\alpha)$ are functions of (a) the measurement (captured in α), which is set by our hypothesis in (2), and (b) the approximation error (captured in τ), which is examined in Lemmas 8, 9 and 10.

V. Numerical Simulation

In this section we present results from numerical simulation of the Bayes filter of Section IV. The theory of the preceding section gives us conservative sufficient conditions, and with numerical simulation in this section we wish to suggest the nature of regimes in which (some refinement of) those conditions become necessary, i.e. when the approximation is "too crude" for the estimator to converge.

For Fig. 2 we use a Normal distribution with standard deviation $\sigma = 1$ for measurements, subplot (A) shows estimator MSE averaged over 10 trials. We use a fixed cell-count (labeled as κ) for the histogram filters presented here; this is the equivalent of setting κ_0 as stated and having $m > 40$ in the sense of (8). Even 3 or 4 cells are seen to perform decently, but the $\kappa = 2$ estimator seems to run into problems attaining an arbitrary posterior mean because of its crude representation. An adaptively refining partition would alleviate this problem.

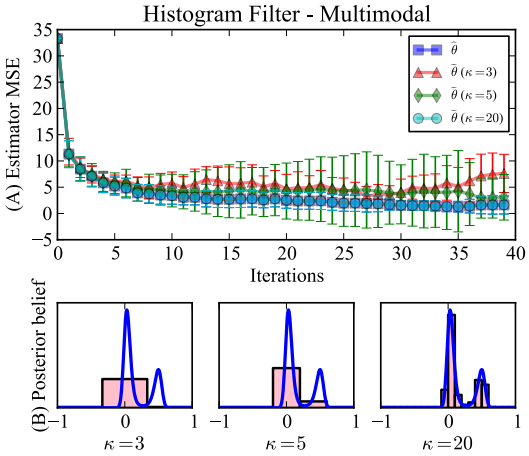

Fig. 3. Performance of the histogram filter for a multimodal measurement model, providing evidence that for more complex measurement models, more complex representations are required for the histogram filter to be accurate (in MSE terms). Note that the number of cells used is larger than those in Fig. 2.

For Fig. 3 we use a mixture model of two Normal distributions for measurements, each with standard deviation $\sigma = 0.2$ to make the bimodality apparent. These more complex measurements require comparately finer partitions from those in Fig. 2, giving some feeling for the more interesting practical settings where poorer measurements or more intricate environments might necessitate some (suitably tightened) version of the sufficient conditions in Section IV.

For some illustrative (clearly, not exhaustive) comparison, we also implement a simple particle filter [17] which resamples at every iteration, and plot the corresponding results in Fig. 4. In order to attain MSE similar to that of the histograms of Fig. 2, we find we need on the order of $\nu \approx 100$ particles, and in the case of our naïve low-dimensional implementations, a ≈ 10-cell histogram filter is relatively more computationally efficient.[7] We do not claim any conclusive computational performance benefits (since we did not rigorously optimize either simulation), but there is certainly some preliminary evidence that computationally tractable crude histogram representations perform comparably to relatively complex particle simulations.

VI. Conclusions and Future Work

We have presented a proof of convergence of a MMSE location parameter estimator using approximate piecewise-

[7] It must be stressed that our present formal results are much more conservative: for example, to reach an error bound of ≈ 0.1 in this example Theorem 11 would call for a cell count of $\approx 10^4$.

constant belief distributions. This means that adding approximation error (which is often unmodeled, yet, necessarily introduced in implementation) need not preclude a proof of convergence of a Bayesian estimator. To the best of our understanding these are the first results in the literature to provide conditions on the approximation quality of a Bayesian estimation scheme sufficient for convergence when the measurements are dependent.

In this initial paper we have presented a very simple approximation operator (Section III), and provided explicit model assumptions (Section II) under which this estimator can be proved to converge to a neighborhood of the true parameter value (Section IV-B). We have implemented this histogram Bayes filter (Section V), and provided numerical evidence that our proofs of convergence, while conservative, introduce conditions that usefully guide the allowable approximation error in regimes where the computational goal of a simplified representation conflicts with the intricacy of the environment being represented. We also compared the performance of the resulting filtering algorithm to a simple particle filter implementation, and empirically found comparable performance with similar or lower computational effort in the low-dimensional settings considered.

The approximation strategy implemented in this paper is motivated by simplicity, but the strict refinement policy causes monotonic growth in the representation complexity. Future work will consider more efficient control of quantization that could include the capability to both increase and decrease cell cardinality using adaptive refinement paired with cell agglomeration in regions of low interest. In fact, this technique seems especially attractive in very high-dimensional problems where large "uninteresting" portions of the estimation space could be combined into a small number of cells, thus drastically reducing computational complexity while still affording analytical leverage similar to what we exploited in this paper.

REFERENCES

[1] N. Patwari, J. N. Ash, S. Kyperountas, A. O. Hero III, R. L. Moses, and N. S. Correal, "Locating the nodes: cooperative localization in wireless sensor networks," *Signal Processing Magazine, IEEE*, vol. 22, no. 4, p. 5469, 2005.

[2] P. Diaconis and D. Freedman, "On inconsistent bayes estimates of location," *The Annals of Statistics*, vol. 14, no. 1, pp. 68–87, Mar. 1986.

[3] J. L. Doob, "Application of the theory of martingales," *Colloques Internationaux du Centre National de la Recherche Scientifique*, pp. 23–27, 1949.

[4] D. A. Freedman, "On the asymptotic behavior of bayes' estimates in the discrete case," *The Annals of Mathematical Statistics*, vol. 34, no. 4, pp. 1386–1403, Dec. 1963.

[5] L. Schwartz, "On bayes procedures," *Probability Theory and Related Fields*, vol. 4, no. 1, p. 1026, 1965.

[6] A. R. Barron, "Discussion: On the consistency of bayes estimates," *The Annals of Statistics*, vol. 14, no. 1, pp. 26–30, Mar. 1986.

[7] S. Ghosal, J. K. Ghosh, and R. V. Ramamoorthi, "Consistent semiparametric bayesian inference about a location parameter," *Journal of statistical planning and inference*, vol. 77, no. 2, p. 181193, 1999.

[8] S. Ghosal, "A review of consistency and convergence of posterior distribution," in *Varanashi Symposium in Bayesian Inference*, Banaras Hindu University, 1997.

[9] C. R. Shalizi, "Dynamics of bayesian updating with dependent data and misspecified models," *Electronic Journal of Statistics*, vol. 3, pp. 1039–1074, 2009.

[10] S. Ghosal and A. Van der Waart, "Convergence rates of posterior distributions for noniid observations," *The Annals of Statistics*, vol. 35, no. 1, pp. 192–223, Feb. 2007.

[11] A. Doucet, N. de Freitas, and N. Gordon, *Sequential Monte Carlo Methods in Practice*, ser. Statistics for Engineering and Information Science Series. Springer, 2001.

[12] D. Fox, "Adapting the sample size in particle filters through KLD-Sampling," *International Journal of Robotics Research*, vol. 22, p. 2003, 2003.

[13] J. Aughenbaugh and B. Cour, "Particle-inspired motion updates for grid-based bayesian trackers," in *2011 Proceedings of the 14th International Conference on Information Fusion (FUSION)*, July 2011, pp. 1–8.

[14] J. Aughenbaugh, J. Kurtz, and B. Cour, "A polynomial-adaptive scheme for bayesian tracking," in *2011 Proceedings of the 14th International Conference on Information Fusion (FUSION)*, July 2011, pp. 1–8.

[15] D. Blackwell and L. Dubins, "Merging of opinions with increasing information," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 882–886, Sept. 1962.

[16] A. R. Syversveen, "Noninformative bayesian priors. interpretation and problems with construction and applications." *Preprint Statistics*, vol. 3, 1998.

[17] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents series)*, ser. Intelligent robotics and autonomous agents. The MIT Press, Aug. 2005, published: Hardcover.

[18] X. Li, "Performance study of RSS-based location estimation techniques for wireless sensor networks," in *IEEE Military Communications Conference, 2005. MILCOM 2005*, Oct. 2005, pp. 1064 –1068 Vol. 2.

[19] B. Charrow, N. Michael, and V. Kumar, "Cooperative multi-robot estimation and control for radio source localization," *Proc. of the Intl. Sym. on Exp. Robot., Quebec City, Quebec*, 2012.

[20] C. G. Bowsher and P. S. Swain, "Identifying sources of variation and the flow of information in biochemical networks," *Proceedings of the National Academy of Sciences*, vol. 109, no. 20, p. E1320E1328, 2012.

[21] L. R. Pericchi and A. F. M. Smith, "Exact and approximate posterior moments for a normal location parameter," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 54, no. 3, pp. 793–804, Jan. 1992.

[22] S. Uhlich, B. Loesch, and B. Yang, "Polynomial LMMSE estimation: A case study," in *IEEE/SP 15th Workshop on Statistical Signal Processing, 2009. SSP '09*, Sept. 2009, pp. 65 –68.

[23] R. Arratia, L. Goldstein, and L. Gordon, "Poisson approximation and the chen-stein method," *Statistical Science*, vol. 5, no. 4, pp. 403–424, Nov. 1990.

[24] E. Pekz, A. Rllin, and N. Ross, "Total variation error bounds for geometric approximation," *arXiv:1005.2774*, May 2010.

[25] F. Daly, "On stein's method, smoothing estimates in total variation distance and mixture distributions," *Journal of Statistical Planning and Inference*, vol. 141, no. 7, pp. 2228–2237, July 2011.

[26] P. Whittle, "On the smoothing of probability density functions," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 20, no. 2, pp. 334–343, Jan. 1958.

[27] D. Freedman and P. Diaconis, "On the histogram as a density estimator: L_2 theory," *Probability theory and related fields*, vol. 57, no. 4, p. 453476, 1981.

[28] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 99th ed. Wiley-Interscience, Aug. 1991.

[29] S. Jackman, *Bayesian Analysis for the Social Sciences*, ser. Wiley Series in Probability and Statistics. Wiley, 2009.

[30] R. Gray and D. Neuhoff, "Quantization," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2325–2383, Oct. 1998.

[31] C. W. Jemmott, R. L. Culver, and J. W. Langelaan, "Comparison of particle filter and histogram filter performance for passive sonar localization," *Proceedings of Meetings on Acoustics*, vol. 8, no. 1, p. 055001, 2009.

[32] A. Ribeiro, "Ergodic stochastic optimization algorithms for wireless communication and networking," *IEEE Transactions on Signal Processing*, vol. 58, no. 12, pp. 6369–6386, Dec. 2010.