# EFFICIENT IMPORTANCE SAMPLING TECHNIQUES FOR LARGE DIMENSIONAL AND MULTIMODAL POSTERIOR COMPUTATIONS

*Samarjit Das and Namrata Vaswani*

Department of Electrical and Computer Engineering
Iowa State University, Ames, IA 50011

## ABSTRACT

In recent work, we proposed some new ideas for efficient sequential importance sampling in the context of particle filtering. These were specifically designed for problems with multimodal posteriors (particularly those with multimodal likelihoods) and with very large dimensions. In this work, we demonstrate the use of similar ideas to improve the performance of importance sampling (IS) in static problems. The key idea of our proposed method is to split the state space in such a way that the posterior conditioned on a small part of the state space is "unimodal". We can then importance sample from the prior for the small "multimodal" part of the state space while adapting existing efficient IS techniques for the much larger dimensional "unimodal" part. We give a modified version of a result from our recent work to obtain sufficient conditions to ensure posterior unimodality. Also, for a subspace of the "unimodal" state space having small enough prior variance, one can replace IS by just estimating the conditional posterior mode. We call this the mode tracking (MT) approximation of IS. We show, via experiments on a large dimensional temperature field estimation problem, that when the number of samples, N, is small, the MT approximation outperforms any standard IS technique.

***Index Terms***— Importance sampling, multimodal observation likelihood, multimodal posterior computation

## 1. INTRODUCTION

In recent work [1], we proposed two new ideas for efficient sequential importance sampling in the context of particle filtering. These were specifically designed for problems with multimodal posteriors (particularly those with multimodal likelihoods) and with very large dimensions. In this work, we demonstrate the use of similar ideas to improve the performance of importance sampling (IS) to compute the posterior in static problems. Well-known IS techniques for posterior computation include sampling from the prior or sampling from a Gaussian approximation to the posterior about its mode (Laplace's approximation) [2, 3, 4]. Some other techniques for multimodal importance sampling include [5]. Importance sampling from the prior was also used in the first particle filter (PF) paper [6] and importance sampling from a Gaussian approximation about the mode was described as one possible technique for approximating the optimal sequential importance density for PF in [7].

Importance sampling from the prior is often very inefficient, especially when the likelihood is reliable [7, 8]. On the other hand, the Gaussian approximation of the posterior is valid only when it is unimodal (or is unimodal most of the time, i.e. for most values of the observation). In this work, we study the situation where the posterior is multimodal. This happens very often due to multimodality or heavy-tailed-ness of the observation likelihood. A common example is large dimensional temperature field estimation when each sensor has some probability of failure. In computer vision, multimodal likelihoods occur due to background clutter or occlusions [9]. But in most of these problems, even though the posterior is multimodal, it is often unimodal conditioned on a part of the state space. If we can find the "multimodal" part of the state vector conditioned on which the posterior will be unimodal, we can sample from the prior for this part and sample from a Gaussian approximation to the conditional posterior for the rest ("unimodal" states).

In [1], we derived sufficient conditions to test for unimodality of the posterior conditioned on the previous state and a part of the current state. In this work, we adapt that result to obtain sufficient conditions for unimodality of the posterior conditioned on a part of the state space.

For a subspace of the "unimodal" state space having small enough prior variance, one can replace IS by just using the conditional posterior mode as the sample. We call this the mode estimation or mode tracking (MT) approximation of IS [1]. We show, via experiments on a large dimensional temperature field estimation problem, that when the number of samples, N, is small, the MT approximation outperforms any standard IS technique.

### 1.1. Problem Definition

We denote the probability density function (pdf) of a random vector $\mathbf{X}$, $f_{\mathbf{X}}(X)$, using the notation $p(X)$ and we denote

This work was partially funded by NSF grant ECCS 0725849

**Algorithm 1** **EIS. Computing** $p^N_{X|Y}(X) = \sum_{i=1}^N w_t^{(i)} \delta(X - X^i),\ X^i = [X_s^i, X_r^i]$

```
1. Importance Sample from the prior for $X_s$: $\forall i$, sample $X_s^i \sim p(X_s)$.
2. Efficient Importance Sample $X_r$: $\forall i$, sample $X_r^i \sim \mathcal{N}(X_r^i; m^{(i)}, \Sigma_{IS}^i)$. Here $m^i(X_s^i, Y) = \arg\min_{X_r} L^i(X_r)$ and $\Sigma_{IS}^i \triangleq (\nabla^2 L^i(m^i))^{-1}$ and $L^i$ is defined in (8).
3. Weight: $\forall i$, compute $w^i = \frac{\tilde{w}^i}{\sum_{j=1}^N \tilde{w}^j}$ where $\tilde{w}^i = \frac{p(Y|X^i)p(X_r^i|X_s^i)}{\mathcal{N}(X_r^i;\, m^i,\, \Sigma_{IS}^i)}$ where $X^{(i)} = [X_s^i, X_r^i]$.
```

the conditional pdf, $f_{\mathbf{X}|\mathbf{Y}}(X|Y)$, by $p(X|Y)$. The goal is to estimate a large dimensional unobserved state vector $X$ from observation vector $Y$ which is a noise-corrupted and nonlinear function of $X$. The optimal minimum mean squared error estimate (MMSE) is given by the posterior expectation $\mathbb{E}[X|Y] = \int Xp(X|Y)dX$. When the integral cannot be computed analytically, we use (Bayesian) importance sampling to approximate it. Our goal is to design efficient importance densities for problems where the observation likelihood, $p(Y|X)$, treated as a function of $X$, is multimodal and when the dimension of $X$ is large.

The paper is organized as follows. We describe our efficient importance sampling (EIS) technique in Sec. 2. Sufficient conditions for testing for conditional posterior unimodality for a static problem are derived in Sec 3. The Mode tracking (MT) idea is discussed in Sec 4. Simulation results comparing EIS and EIS-MT with existing work for a large dimensional temperature field estimation problem are given in Sec. 5. Conclusions are given in Sec. 6.

## 2. EFFICIENT IMPORTANCE SAMPLING

Denote the posterior by $p^*(X)$, i.e.

$$p^*(X) \quad \triangleq \quad p(X|Y) \propto p(Y|X)p(X) \tag{1}$$

In most cases, the posterior cannot be computed analytically and hence importance sampling is needed (else simple Monte Carlo would suffice). *If $p^*$ is unimodal (at least approximately)*, one can approximate it by a Gaussian about its mode and sample from it (Laplace's approximation)[2, 7]. But, when the observation likelihood $p(Y|X)$ is multimodal, or heavy-tailed, or otherwise not strongly log-concave, $p^*$ will be unimodal only if the prior $p(X)$ is unimodal and narrow enough and the state sample is near enough to an observation likelihood mode. In many situations, this may not hold in all dimensions. But in most such situations, the prior is broad and/or multimodal in only a few directions, which we call the *"multimodal" directions*. It can be shown that if the prior is unimodal and narrow enough in the rest of the directions, $p^*$ will be unimodal conditioned on the *"multimodal states"* (this is proved in Theorem 1). When this holds, we propose to split the state vector as $X = [X_s; X_r]$ in such a way that $X_s$ contains the minimum number of dimensions for which $p^*$ is unimodal conditioned on it, i.e.

$$p^{**,i}(X_r) \triangleq p^*(X|X_s^i) = p(X_r|X_s^i, Y) \tag{2}$$

is unimodal. We sample $X_s$ from its prior (to sample the possibly multiple modes of $p^*$), and use Laplace's approximation to approximate $p^{**,i}$ and sample $X_r$ from it, i.e. sample $X_r^i$ from $\mathcal{N}(m^i, \Sigma_{IS}^i)$ where

$$\begin{aligned} m^i \quad &= \quad m^i(X_s^i, Y) \triangleq \min_{X_r} L^i(X_r), \\ \Sigma_{IS}^i \quad &\triangleq \quad [(\nabla^2 L^i)(m^i)]^{-1}, \text{ where} \\ L^i(X_r) \quad &\triangleq \quad -\log[p^{**,i}(X_r)] + \text{const} \end{aligned} \tag{3}$$

$\nabla^2 L^i$ denotes the Hessian of $L^i$. The weighting step also changes to satisfy the principle of importance sampling. The complete algorithm is given in Algorithm 1. We call it Efficient Importance Sampling (EIS). It is to be noted that if $X_s$ is chosen so that $p^{**,i}$ is unimodal for most particles and at most times (i.e. is unimodal with high probability), the proposed algorithm works well.

## 3. TESTING FOR POSTERIOR UNIMODALITY

We derive sufficient conditions for unimodality of the conditional posterior, $p^{**,i}$, defined in (2). By setting $X_s =$ empty, the same conditions can be used for checking for posterior unimodality. Let $\dim(X_s) \triangleq K$, $\dim(X_r) \triangleq M_r$, $\dim(X) \triangleq M = K + M_r$. Now,

$$p^{**,i}(X_r) \quad = \quad \zeta p(Y|X_s^i, X_r)p(X_r|X_s^i) \tag{4}$$

where $\zeta$ is a proportionality constant.

**Definition 1** *We first define a few terms and symbols.*

1. *The notation $A > 0$ ($A \geq 0$) where $A$ is a square matrix means that $A$ is* positive definite (positive semidefinite). *Also, $A > B$ ($A \geq B$) means $A - B > 0$ ($A - B \geq 0$).*

2. *The term* "minimizer" *refers to the unconstrained local minimizer of a function, i.e. a point $x_0$ s.t. $f(x_0) \leq f(x)\ \forall\ x$ in its neighborhood. Similarly for "maximizer".*

3. *A twice differentiable function, $f(x)$, is* strongly convex *in a region $\mathcal{R}$, if there exists an $m > 0$ s.t. at all points, $x \in \mathcal{R}$, the Hessian $\nabla^2 f(x) \geq mI$. If $f$ is strongly convex in $\mathcal{R}$, it has at most one minimizer in $\mathcal{R}$ and it lies in the interior of $\mathcal{R}$. If $f$ is strongly-convex on $\mathbb{R}^M$, then it has exactly one (finite) minimizer.*

4. *A function is* strongly log-concave *if its negative log is strongly convex. An example is a Gaussian pdf.*

5. *Since a pdf is an integrable function, it will always have at least one (finite) maximizer. Thus a pdf having at most one maximizer is equivalent to it being* unimodal.

6. *The symbol $\mathbb{E}[.]$ denotes expected value.*

7. *We denote the $-\log$ of the observation likelihood using the symbol $E_Y$, i.e.*

$$E_Y(X) \triangleq -\log p(Y|X) + \mathit{const} \tag{5}$$

8. *We denote the $-\log$ of the prior of $X_r$ as*

$$D^i(X_r) \triangleq -\log p(X_r|X_s^i) + \mathit{const} \tag{6}$$

9. *When the prior of $X_r$ is strongly log-concave (assumed in Theorem 1), we denote its unique mode by*

$$f_r^i \triangleq f_r(X_s^i) = \arg\max_{X_r} p(X_r|X_s^i) \tag{7}$$

10. *$[z]_p$ or $z_p$ denotes the $p^{th}$ coordinate of a vector, $z$.*

11. *$\max_p$ is often used in place of $\max_{p=1,2,\ldots M_r}$.*

Combining (4), (5) and (6), $L^i(X_r)$ can be written as

$$L^i(X_r) = E_Y(X_s^i, X_r) + D^i(X_r) \tag{8}$$

Now, $p^{**,i}(X_r)$ will be unimodal if and only if we can show that $L^i$ has at most one minimizer. We derive a set of sufficient conditions on $E_Y$, $D^i$ and $f_r^i$ to ensure this. The main idea is as follows. We assume strong log-concavity (e.g. Gaussianity) of the prior of $X_r$. Thus $D^i(X_r)$ will be strongly convex with a unique minimizer at $f_r^i$. But $E_Y(X)$ (and so $E_Y$ as a function of $X_r$) can have multiple minimizers since observation likelihood can be multimodal. Assume that $E_Y(X_s^i, X_r)$ is locally convex in the neighborhood of $f_r^i$ (this will hold if $f_r^i$ is close enough to any of its minimizers). Denote this region by $\mathcal{R}_{LC}$. Thus, inside $\mathcal{R}_{LC}$, $L^i$ will be strongly convex and hence it will have at most one minimizer. We show that if $\max_p |[\nabla D]_p|$ is large enough outside $\mathcal{R}_{LC}$ (the spread of the prior of $X_r$ is small enough), $L^i$ will have no stationary points (and hence no minimizers) outside $\mathcal{R}_{LC}$ or on its boundary.

This idea leads to Theorem 1 below. Its first condition ensures strong convexity of $D^i$ everywhere. The second one ensures that $\mathcal{R}_{LC}$ exists. The third one ensures that $\exists$ an $\epsilon_0 > 0$, s.t. at all points in $\mathcal{R}_{LC}^c$ (complement of $\mathcal{R}_{LC}$), $\max_p |[\nabla L^i]_p| > \epsilon_0$ (i.e. $L^i$ has no stationary points in $\mathcal{R}_{LC}^c$).

**Theorem 1** *$p^{**,i}(X_r)$ is unimodal with the unique mode lying inside $\mathcal{R}_{LC}$ if the following hold:*

1. *The prior of $X_r$, $p(X_r|X_s^i)$, is strongly log-concave. Its unique mode is denoted by $f_r^i$.*

2. *The $-\log$ of the observation likelihood given $X_s^i$, $E_Y(X_s^i, X_r)$ is twice continuously differentiable almost everywhere and is locally convex in the neighborhood of $f_r^i$. Let $\mathcal{R}_{LC} \subseteq \mathbb{R}^{M_r}$ denote the largest convex region in the neighborhood of $f_r^i$ where $\nabla^2_{X_r} E_Y(X_s^i, X_r) \geq 0$ ($E_Y$ as a function of $X_r$ is locally convex).*

3. *There exists an $\epsilon_0 > 0$ such that*

$$\inf_{X_r \in \cap_{p=1}^{M_r} (\mathcal{A}_p \cup \mathcal{Z}_p)} \max_{p=1,\ldots M_r} [\gamma_p(X_r)] > 1 \tag{9}$$

*where*

$$\gamma_p(X_r) \triangleq \begin{cases} \frac{|[\nabla D^i]_p|}{\epsilon_0 + |[\nabla E_Y]_p|}, & if \;\; X_{t,r} \in \mathcal{A}_p \\ \\ \frac{|[\nabla D^i]_p|}{\epsilon_0 - |[\nabla E_Y]_p|}, & if \;\; X_r \in \mathcal{Z}_p \end{cases} \tag{10}$$

$$\begin{aligned} \mathcal{A}_p &\triangleq \{X_r \in \mathcal{R}_{LC}^c : [\nabla D^i]_p . [\nabla E_Y]_p < 0\} \\ \mathcal{Z}_p &\triangleq \{X_r \in \mathcal{R}_{LC}^c : \\ & \quad [\nabla E_Y]_p . [\nabla D^i]_p \geq 0 \; \& \; |[\nabla E_Y]_p| < \epsilon_0\} \quad (11) \\ \nabla E_Y &\triangleq \nabla_{X_r} E_Y(X_s^i, X_r) \\ \nabla D^i &\triangleq \nabla_{X_r} D^i(X_r) \quad (12) \end{aligned}$$

*Proof:* In the proof, $\nabla$ is used to denote $\nabla_{X_r}$. Also, we remove the superscripts from $L^i$ and $D^i$. $p^{**,i}(X_r)$ will be unimodal iff $L$ defined in (8) has at most one minimizer. We obtain sufficient conditions for this. Condition 1) ensures that $D$ is strongly convex everywhere with a unique minimizer at $f_r^i$. Condition 2) ensures that $\mathcal{R}_{LC}$ exists. By definition of $\mathcal{R}_{LC}$, $E_{Y_t}$ is convex inside it. Thus the first two conditions ensure that $L$ is strongly convex inside $\mathcal{R}_{LC}$. So it has at most one minimizer inside $\mathcal{R}_{LC}$.

We now show that if condition 3) also holds, $L$ will have no stationary points (and hence no minimizers) in $\mathcal{R}_{LC}^c$ or on its boundary. A sufficient condition for this is: $\exists \; \epsilon_0 > 0$ s.t.

$$\max_p |[\nabla L]_p| > \epsilon_0, \; \forall X_r \in \mathcal{R}_{LC}^c \tag{13}$$

We show that condition 3) is sufficient to ensure (13). Note that $\nabla L = \nabla E_Y + \nabla D$. In the regions where for at least one $p$, $[\nabla E_Y]_p . [\nabla D]_p \geq 0$ (have same sign) and $|[\nabla E_Y]_p| > \epsilon_0$, condition (13) will always hold. Thus we only need to worry about regions where, for all $p$, either $[\nabla E_Y]_p . [\nabla D]_p < 0$ or $[\nabla E_Y]_p . [\nabla D]_p \geq 0$ but $|[\nabla E_Y]_p| < \epsilon_0$. This is the region

$$\cap_{p=1}^{M_r} (\mathcal{A}_p \cup \mathcal{Z}_p) \triangleq \mathcal{G}, \;\; \mathcal{A}_p, \mathcal{Z}_p \text{ defined in (11)} \tag{14}$$

Now, $D$ only has one stationary point which is $f_r^i$ and it lies inside $\mathcal{R}_{LC}$ (by definition of $\mathcal{R}_{LC}$), and none in $\mathcal{R}_{LC}^c$. Thus $\nabla D \neq 0$ in $\mathcal{R}_{LC}^c$ and, in particular, inside $\mathcal{G} \subset \mathcal{R}_{LC}^c$. Thus if we can find a condition which ensures that, for all points in $\mathcal{G}$, for at least one $p$, $[\nabla L]_p$ "follows the sign of $[\nabla D]_p$" (i.e. $[\nabla L]_p > \epsilon_0$ where $[\nabla D]_p > 0$ and $[\nabla L]_p < -\epsilon_0$ where $[\nabla D]_p < 0$), we will be done.

We first find the required condition for a given $p$ and a point $X_r \in \mathcal{G}$. For any $p$, if $X_r \in \mathcal{G}$, then it either belongs to $\mathcal{A}_p$ or belongs to $\mathcal{Z}_p$. If $X_r \in \mathcal{A}_p$, $|[\nabla L]_p| > \epsilon_0$ if

$$\frac{|[\nabla D]_p|}{\epsilon_0 + |[\nabla E_Y]_p|} > 1 \tag{15}$$

This is obtained by combining the conditions for the case $[\nabla D]_p > 0$ and the case $[\nabla D]_p < 0$. Proceeding in a similar fashion, if $X_r \in \mathcal{Z}_p$, $|[\nabla L]_p| > \epsilon_0$ if

$$\frac{|[\nabla D]_p|}{\epsilon_0 - |[\nabla E_Y]_p|} > 1 \tag{16}$$

Inequalities (15) and (16) can be combined and rewritten as $\gamma_p(X_r) - 1 > 0$ where $\gamma_p$ is defined in (10). For (13) to hold, we need $|[\nabla L]_p| > \epsilon_0$ for at least one $p$, for all $X_r \in \mathcal{G}$. This will happen if $\inf_{X_r \in \mathcal{G}} \max_p \gamma_p(X_r) > 1$. But this is condition 3. Thus condition 3) implies that $L$ has no minimizers in $\mathcal{R}_{LC}^c$. Thus if conditions 1), 2) and 3) of the theorem hold, $L$ has at most one minimizer which lies inside $\mathcal{R}_{LC}$. Thus $p^{**,i}(X_{t,r})$ has a unique mode which lies inside $\mathcal{R}_{LC}$, i.e. it is unimodal. ■

The most common example of a strongly log-concave pdf is a Gaussian. When the prior of $X_r$ is Gaussian with mean (= mode) $f_r^i$, the above result can be further simplified to get an upper bound on the eigenvalues of its covariance matrix. First consider the case when the covariance is diagonal, denoted $\Delta_r$. In this case, $D^i(X_r) = \sum_p \frac{([X_r - f_r^i]_p)^2}{2\Delta_{r,p}}$ and so $[\nabla D^i]_p = \frac{[X_r - f_r^i]_p}{\Delta_{r,p}}$. By substituting this in condition 3), it is easy to see that we get the following simplified condition:

$$\inf_{X_r \in \cap_{p=1}^{M_r} (\mathcal{A}_p \cup \mathcal{Z}_p)} \max_p [\gamma_p^{num}(X_r) - \Delta_{r,p}] > 0 \tag{17}$$

$$\gamma_p^{num}(X_r) \triangleq \begin{cases} \frac{|[X_r - f_r^i]_p|}{\epsilon_0 + |[\nabla E_Y]_p|}, & if \;\; X_r \in \mathcal{A}_p \\ \\ \frac{|[X_r - f_r^i]_p}{\epsilon_0 - |[\nabla E_Y]_p|}, & if \;\; X_r \in \mathcal{Z}_p \end{cases} \tag{18}$$

$$\begin{aligned} \mathcal{A}_p &\triangleq \{X_r \in \mathcal{R}_{LC}^c : [X_r - f_r^i]_p . [\nabla E_Y]_p < 0\} \\ \mathcal{Z}_p &\triangleq \{X_r \in \mathcal{R}_{LC}^c : \\ &\quad [\nabla E_Y]_p . [X_r - f_r^i]_p \geq 0 \;\&\; |[\nabla E_Y]_p| < \epsilon_0\} \end{aligned} \tag{19}$$

Also, since $\max_p [g_1(p) - g_2(p)] \geq \max_p g_1(p) - \max_p g_2(p)$ for any two functions, $g_1$, $g_2$, a sufficient condition for (17) is

$$\max_p \Delta_{r,p} < \inf_{X_r \in \cap_{p=1}^{M_r} (\mathcal{A}_p \cup \mathcal{Z}_p)} \max_p [\gamma_p^{num}(X_r)] \triangleq \Delta^* \tag{20}$$

Thus, we have the following corollary.

**Corollary 1** *When the prior of $X_r$ is Gaussian with mean $f_r^i$ and* diagonal covariance, $\Delta_r$, *$p^{**,i}(X_r)$ is unimodal if (a) condition 2) of Theorem 1 holds and (b) there exists an $\epsilon_0 > 0$ s.t. (17) holds with $\gamma_p^{num}$ defined in (18) and $\mathcal{A}_p, \mathcal{Z}_p$ defined in (19). A sufficient condition for (17) is (20).*

Now consider the case when the prior of $X_r$ is Gaussian with non-diagonal covariance, $\Sigma_r = U \Delta_r U^T$. Define $\tilde{X}_r = U^T X_r$. Since $\tilde{X}_r$ is a one-to-one and linear function of $X_r$, it is easy to see that $p^{**,i}(X_r)$ is unimodal iff $p^{**,i}(\tilde{X}_r) \triangleq p(\tilde{X}_r | X_s^i, Y)$ is unimodal. The prior of $\tilde{X}_r$ is $\mathcal{N}(U^T f_r^i, \Delta_r)$. Also, its observation likelihood is $p(Y | X_s^i, U\tilde{X}_r)$. Define $\tilde{E}_Y(\tilde{X}_r) \triangleq E_Y(U\tilde{X}_r)$.

**Corollary 2** *When the prior of $X_r$ is Gaussian with mean $f_r^i$ and* non-diagonal covariance, $\Sigma_r = U \Delta_r U^T$, *$p^{**,i}(X_r)$ is unimodal if the conditions of Corollary 1 hold with $E_Y$ replaced by $\tilde{E}_Y$; $f_r^i$ replaced by $U^T f_r^i$ and $X_r$ replaced by $\tilde{X}_r$ everywhere.*

To summarize the above discussion, $p^{**,i}$ is unimodal if

1. The prior of $X_r$ is strongly log-concave (e.g. Gaussian),

2. The mode of the prior of $X_r$ is "close enough" to a mode of [observation likelihood given $X_s^i$], so that condition 2) of Theorem 1 holds. Denote this mode by $X_r^*$.

3. The maximum spread of the prior of $X_r$ is "small enough" to ensure that condition 3) of Theorem 1 holds. In the Gaussian prior case, this translates to the maximum eigenvalue of its covariance being smaller than $\Delta^*$, defined in (20). $\Delta^*$ itself is directly proportional to the distance of $X_r^*$ to the next nearest mode of [observation likelihood given $X_s^i$] and inversely proportional to its strength.

*The last two conditions above automatically hold if [observation likelihood given $X_s^i$] is strongly log-concave* ($\mathcal{R}_{LC}^c$ is empty and so $\Delta^* = \infty$).

## 4. MODE TRACKING (MT) APPROXIMATION OF IS

For any importance sampling (including EIS and the importance sampling techniques used in PF-Gordon [6] or in PF-Doucet [7]), the effective sample size [8, 7] reduces with increasing dimension, i.e. the $N$ required for a given estimation accuracy increases with dimension. This makes the state estimation problem impractically expensive when the dimensionality of the state vector is large. We discuss one possible solution to this problem here.

**Algorithm 2 EIS-MT. Computing** $p^N_{X|Y}(X) = \sum_{i=1}^N w_t^{(i)} \delta(X - X^i)$**,** $X^i = [X^i_s, X^i_r]$**,** $X^i_r = [X^i_{r,s}, X^i_{r,r}]$

```
1. Importance Sample $X_s$: $\forall i$, sample $X^i_s \sim p(X^i_s)$.
2. Efficient Importance Sample $X_{r,s}$: $\forall i$,
    (a) Compute $m^i(X^i_s, Y) = \arg\min_{X_r} L^i(X_r)$ and $\Sigma^i_{IS} \triangleq (\nabla^2 L^i(m^i))^{-1}$ where $L^i$ is defined in (8). Let $m^i = \begin{bmatrix} m^i_s \\ m^i_r \end{bmatrix}$
        and $\Sigma^i_{IS} = \begin{bmatrix} \Sigma_{IS,s} & \Sigma_{IS,s,r} \\ \Sigma_{IS,r} & \Sigma_{IS,r,s} \end{bmatrix}$.
    (b) Sample $X^i_{r,s} \sim \mathcal{N}(m^i_s,\, \Sigma^i_{IS,s})$.
3. Mode Track $X_{r,r}$: $\forall i$,
    (a) Compute $m^{*i}_r$ using (21).
    (b) Set $X^i_{r,r} = m^{*i}_r$
4. Weight: $\forall i$, compute $w^i = \frac{\tilde{w}^i}{\sum_{j=1}^N \tilde{w}^j}$ where $\tilde{w}^i = \frac{p(Y|X^i)p(X^i_r|X^i_s)}{\mathcal{N}(X^i_r;\, m^i,\, \Sigma^i_{IS})}$ where $X^i_r = [X^i_{r,s}, X^i_{r,r}]$.
```

Consider a large dimensional state vector $X$. To apply EIS, we split the state $X$ into $[X_s, X_r]$, such that $p^*$ is unimodal w.h.p. conditioned on $X_s$. As explained earlier, this is ensured if the eigenvalues of $\Sigma_r$ are small enough to satisfy (20). Now, $X_r$ can further be split into $[X_{r,s}, X_{r,r}]$ so that the maximum eigenvalue of the covariance of the prior of $X_{r,r}$ is small enough to ensure that there is little error in approximating the conditional posterior of $X_{r,r}$ by a Dirac delta function at its mode. We call this the Mode Tracking (MT) approximation of importance sampling (IS), or IS-MT. We *refer to $\tilde{X}_s \triangleq [X_s, X_{r,s}]$ as the "effective" state and to $\tilde{X}_r \triangleq X_{r,r}$ as the "residual" state.* We explain IS-MT in detail below.

In EIS, we IS $X^i_s$ from its prior, and we EIS $X^i_r$ from $\mathcal{N}(m^i, \Sigma^i_{IS})$ where $m^i$, $\Sigma^i_{IS}$ are defined in (3). Let $m^i = \begin{bmatrix} m^i_s \\ m^i_r \end{bmatrix}$ and $\Sigma^i_{IS} = \begin{bmatrix} \Sigma_{IS,s} & \Sigma_{IS,s,r} \\ \Sigma_{IS,r,s} & \Sigma_{IS,r} \end{bmatrix}$. This is equivalent to first sampling $X^i_{r,s} \sim \mathcal{N}(m^i_s, \Sigma^i_{IS,s})$ and then sampling $X^i_{r,r} \sim \mathcal{N}(m^{*i}_r, \Sigma^i_{IS,r})$ where

$$m^{*i}_r \triangleq m^i_r + \Sigma^i_{IS,r,s} {\Sigma^i_{IS,s}}^{-1} (X^i_{r,s} - m^i_s),$$
$$\Sigma^*_{IS,r}{}^i \triangleq \Sigma^i_{IS,r} - \Sigma^i_{IS,r,s} {\Sigma^i_{IS,s}}^{-1} {\Sigma^i_{IS,r,s}}^T \tag{21}$$

Now, from (21), $\Sigma^*_{IS,r}{}^i \leq \Sigma^i_{IS,r}$. Also, since $m^i$ lies in a locally convex region of $E_Y(X^i_s, X_r)$, i.e. $\nabla^2 E_Y(X^i_s, m^i) \geq 0$ (by Theorem 1), $\Sigma^i_{IS} \leq \Delta_r$. This implies that $\Delta_{r,r} - \Sigma^i_{IS,r}$, which is a square sub-matrix of $\Delta_r - \Sigma^i_{IS}$, is also non-negative definite. Thus,

$$\Sigma^*_{IS,r}{}^i \leq \Sigma^i_{IS,r} \leq \Delta_{r,r} \tag{22}$$

If the maximum eigenvalue of $\Delta_{r,r}$ is small enough, any sample from $\mathcal{N}(m^{*i}_r, \Sigma^*_{IS,r}{}^i)$ will be close to $m^{*i}_r$ w.h.p. So we can set $X^i_{r,r} = m^{*i}_r$ with little extra error. *The algorithm is then called EIS-MT. It is summarized in Algorithm 2.* A more accurate, but also more expensive modification (need to implement it on-the-fly) would be do MT on the low eigenvalue directions of $\Sigma^i_{IS}$.

The IS-MT approximation introduces some error in the estimate of $X_{r,r}$ (error decreases with decreasing spread of $p^{**,i}(X_{r,r})$). But it also reduces the sampling dimension from $\dim(X)$ to $\dim([X_s; X_{r,s}])$ (significant reduction for large dimensional problems), thus improving the effective sample size. For carefully chosen dimension of $X_{r,r}$, this results in smaller total error, especially when the available number of particles, $N$, is small. This is observed experimentally, but proving it theoretically is an open problem. We say that the *IS-MT approximation is "valid"* for a given choice of $X_{r,r}$ if it results in smaller total error than if it were not used.

## 5. TEMPERATURE FIELD ESTIMATION

Consider the problem of estimating spatially varying temperature (temperature field) from a network of sensors, which obtain noisy observations of temperature and some of them could occasionally fail. Assume that we have sensors $S_1, \ldots, S_K$ in $K$ different spatial locations. The corresponding true temperature is $C = [C_1, \ldots, C_K]^T$ and the sensor observations are $Y = [Y_1, \ldots, Y_K]^T$. Define $V \triangleq [V_1, \ldots, V_K]^T$ where, $V_i$ is the coefficient along the $i^{th}$ eigen direction of temperature variation. The relationship between $C$ and $V$ is given as, $C = m_c + BV$, where $m_c$ is the mean temperature vector and $B$ is a $K \times K$ orthonormal matrix with its columns as the eigen directions of temperature variation. Thus the state vector becomes, $X = [C^T, V^T]^T$. The prior on $V$ is given as, $p(V) = \mathcal{N}(V; \mathbf{0}, \Sigma_v)$.

We assume that any sensor fails with probability $(1 - p)$

| Sl no. | Importance Sampling method | Averaged Normalized RMSE ($N = 30$) |
|---|---|---|
| 1 | EIS-MT ($X_s = [V_1]$, $X_{r,s} = [V_2, V_3]$, $X_{r,r} = [V_4, V_5, V_6, V_7]$) | **0.0416** |
| 2 | EIS-MT ($X_s = [V_1, V_2]$, $X_{r,s} = [V_3, V_4]$, $X_{r,r} = [V_5, V_6, V_7]$) | 0.0593 |
| 3 | EIS ($X_s = [V_1]$, $X_{r,s} = [V_2, V_3, V_4, V_5, V_6, V_7]$, $X_{r,r} = empty$) | 0.0449 |
| 4 | IS-Gaussian ($X_s = empty$, $X_{r,s} = [V]$, $X_{r,r} = empty$) | 0.0610 |
| 5 | IS-prior ($X_s = [V]$, $X_{r,s} = empty$, $X_{r,r} = empty$) | 0.0733 |

| Sl no. | Importance Sampling method | Averaged Normalized RMSE ($N = 100$) |
|---|---|---|
| 1 | EIS-MT ($X_s = [V_1]$, $X_{r,s} = [V_2, V_3]$, $X_{r,r} = [V_4, V_5, V_6, V_7]$) | 0.0375 |
| 2 | EIS-MT ($X_s = [V_1, V_2]$, $X_{r,s} = [V_3, V_4]$, $X_{r,r} = [V_5, V_6, V_7]$) | 0.0420 |
| 3 | EIS ($X_s = [V_1]$, $X_{r,s} = [V_2, V_3, V_4, V_5, V_6, V_7]$, $X_{r,r} = empty$) | **0.0368** |
| 4 | IS-Gaussian ($X_s = empty$, $X_{r,s} = [V]$, $X_{r,r} = empty$) | 0.0587 |
| 5 | IS-prior ($X_s = [V]$, $X_{r,s} = empty$, $X_{r,r} = empty$) | 0.0599 |

**Fig. 1**. Comparing EIS-MT with EIS, IS-prior and IS-Gaussian for $N = 30$ (top) and $N = 100$ (bottom)

independent of all other sensors. When the sensor is working properly, the observation is a noise-corrupted scaled version of the original temperature. But when the sensor fails, the observation is independent of the true temperature at the sensor location. We model it as a large variance Gaussian. To summarize, the observation likelihood (OL) is given as follows:

$$p(Y|X) = p(Y|C) = \prod_{i=1}^{K} [p\mathcal{N}(\alpha_o C_i, \sigma_o^2) + (1-p)\mathcal{N}(\mathbf{0}, 10\sigma_o^2)] \tag{23}$$

where $\alpha_o$ is a scaling factor and $\sigma_o^2$ is the observation noise variance. Since $C$ is deterministic given $V$, we performed importance sampling on $V$ and computed $C = m_c + BV$.

We simulated the above system with $K = 7$ sensors, $p = 0.8$, $\alpha_o = 0.9$, $\sigma_o = 0.5$, $m_c = [25, ..., 25]^T$ and $\Sigma_v = diag([3^2, 5^2, 2^2, 2^2, 1, 1, 1])$ where $diag(a)$ denotes a diagonal matrix with $a$ as its diagonal. The performance measure of the system is given by averaging the normalized RMSE, $NE = \frac{||C - \hat{C}||}{||C||}$ over 50 monte-carlo simulations. Here, $\hat{C}$ is the importance sampling estimate of $\mathbb{E}[C|Y]$.

We computed $\hat{C}$ using the following IS techniques and compared the $NE$ values: EIS, EIS-MT, IS-prior and IS-Gaussian-approx. Notice that IS-prior can be interpreted as EIS-MT with $X_s = X$, while IS-Gaussian can be interpreted as EIS-MT with $X_{r,s} = X$. We used two different values of the sample size, $N = 30$ and $N = 100$. Also, while performing EIS-MT we tried two different case : 1) when $X_s = [V_1]$, $X_{r,s} = [V_2, V_3]$, $X_{r,r} = [V_4, V_5, V_6, V_7]$ and 2) when $X_s = [V_1, V_2]$, $X_{r,s} = [V_3, V_4]$, $X_{r,r} = [V_5, V_6, V_7]$. The results are summarized in Fig. 1. Notice that both EIS and EIS-MT significantly outperform IS-prior and IS-Gaussian. When $N$ is large, EIS has the best performance. But as explained in Sec. 4, when $N$ is small, EIS-MT outperforms EIS and and all other methods. This is because in EIS-MT we importance sample only on 3 dimensions (while computing conditional posterior mode for the rest) and thus its effective sample size is much larger.

## 6. CONCLUSIONS

We proposed two new techniques for large dimensional Bayesian importance sampling problems with frequently multimodal likelihoods. Significantly improved performance over sampling from prior and sampling from Gaussian approximation to posterior (both of which can be interpreted as special cases of our algorithm) was demonstrated, particularly when the number of samples used is small. We also derived sufficient conditions to test for posterior unimodality.

## 7. REFERENCES

[1] N. Vaswani, "Particle filtering for large dimensional state spaces with multimodal observation likelihoods," *IEEE Trans. Sig. Proc.*, October 2008.

[2] L. Tierney and J. B. Kadane, "Accurate approximations for posterior moments and marginal densities," *J. Amer. Stat. Assoc*, vol. 81 (393), pp. 82–86, March 1986.

[3] D. V. Lindley, "Approximate bayesian methods," in *Bayesian Statistics, University Press, Valencia, Spain*, J. M. Bernado, M. H. Degroot, D.V. Lindley, and A. F. M. Smith, Eds., 1980.

[4] F. Mosteller and D.L. Wallace, *Inference and Disputed Authorship: The Federalist*, Addison-Wesley, Reading, MA, 1964.

[5] M. S. Oh and J. O. Berger, "Integration of multimodal functions by monte carlo importance sampling," *J. Amer. Stat. Assoc*, vol. 88 (22), pp. 450–456, 1993.

[6] N. J. Gordon, D. J. Salmond, and A. F. M. Smith, "Novel approach to nonlinear/nongaussian bayesian state estimation," *IEE Proceedings-F (Radar and Signal Processing)*, pp. 140(2):107–113, 1993.

[7] A. Doucet, "On sequential monte carlo sampling methods for bayesian filtering," in *Technical Report CUED/F-INFENG/TR. 310, Cambridge University Department of Engineering*, 1998.

[8] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for on-line non-linear/non-gaussian bayesian tracking," *IEEE Trans. Sig. Proc.*, vol. 50, no. 2, pp. 174–188, Feb. 2002.

[9] M. Isard and A. Blake, "Condensation: Conditional Density Propagation for Visual Tracking," *Intl. Journal Comp. Vis.*, pp. 5–28, 1998.