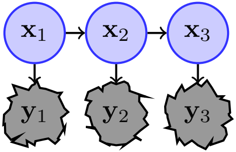
Differentiable Particle Filtering via Entropy-Regularized Optimal Transport

Adrien Corenflos, James Thornton, George Deligiannidis & Arnaud Doucet
Department of Statistics, Oxford University

ICML 2021

- Latent Markov process $\{X_t\}_{t\geq 1}$ and observations $\{Y_t\}_{t\geq 1}$ with $X_1 \sim \mu_\theta(\cdot)$,

$$X_t | (X_{t-1} = x_{t-1}) \sim f_\theta(\cdot | x_{t-1}), \qquad Y_t | (X_t = x_t) \sim g_\theta(\cdot | x_t).$$



- Ubiquitous in econometrics, statistics, machine learning, robotics.
- Interested in estimating parameter θ given observations $Y_{1:T} = y_{1:T}$

Inference in State-Space Models

- Given θ, sequential state inference based on optimal filter $p_\theta(x_t|y_{1:t})$

$$\textbf{Prediction: } p_\theta(x_t|y_{1:t-1}) = \int p_\theta(x_{t-1}|y_{1:t-1}) f_\theta\left(x_t|x_{t-1}\right) \mathrm{d}x_t$$

$$\textbf{Bayes update: } \; p_\theta(x_t|y_{1:t}) = \frac{g_\theta\left(y_t|x_t\right) p_\theta(x_t|y_{1:t-1})}{p_\theta(y_t|y_{1:t-1})},$$

- Log-likelihood function

$$\ell(\theta) = \log p_\theta(y_{1:T}) = \sum_{t=1}^{T} \log p_\theta(y_t|y_{1:t-1}).$$

- The optimal filter $p_\theta(x_t|y_{1:t})$ and log-likelihood $\ell(\theta)$ are intractable except for finite state-space and linear Gaussian models.

Particle Filter 101: The Bootstrap Filter

- **Sampling**: For $i = 1, ..., N$, sample $\widetilde{X}_t^i \sim f_\theta(\cdot|X_{t-1}^i)$ then

$$\widehat{p}_\theta(x_t|y_{1:t-1}) = \frac{1}{N}\sum_{i=1}^{N} \delta_{\widetilde{X}_t^i}$$

- **Weighting**: Set $\widetilde{p}_\theta(x_t|y_{1:t}) = \sum_{i=1}^N w_t^i \delta_{\widetilde{X}_t^i}, \quad w_t^i \propto g_\theta(y_t|\widetilde{X}_t^i), \ \sum_{i=1}^N w_t^i = 1.$

- **Resampling**: For $i = 1, ..., N$, sample $X_t^i \sim \widetilde{p}_\theta(x_t|y_{1:t})$ to obtain

$$\widehat{p}_\theta(x_t|y_{1:t}) = \frac{1}{N}\sum_{i=1}^{N} \delta_{X_t^i}$$

- From PF outputs, one gets a consistent estimate of $\ell(\theta)$

$$\widehat{\ell}(\theta) = \sum_{t=1}^{T} \log \widehat{p}_\theta(y_t|y_{1:t-1}), \quad \text{for } \widehat{p}_\theta(y_t|y_{1:t-1}) = \frac{1}{N}\sum_{i=1}^{N} g_\theta(y_t|\widetilde{X}_t^i)$$

- Likelihood estimate is unbiased for any $N \geq 1$: $\mathbb{E}_{\mathrm{PF}}[\exp \widehat{\ell}(\theta)] = \exp \ell(\theta)$

Variational inference meets particle filters

- As the likelihood estimate output by PF is unbiased,

$$\ell^{\mathrm{ELBO}}(\theta) = \mathbb{E}_{\mathrm{PF}}[\widehat{\ell}(\theta)] \leq \log \mathbb{E}_{\mathrm{PF}}[\exp \widehat{\ell}(\theta)] = \ell(\theta),$$

  which could be maximized using SGD.

- This was exploited in (Maddison et al., *NIPS* 2017; Naesseth et al., *AISTATS* 2018; Le et al., *ICLR* 2018).

- PF are attractive as the "variational gap" satisfies

$$\ell^{\mathrm{ELBO}}(\theta) - \ell(\theta) \approx -\frac{1}{2}\mathrm{var}\left[\frac{\exp \widehat{\ell}(\theta)}{\exp \ell(\theta)}\right].$$

- **Problem**: Unbiased estimates of $\nabla_\theta \ell^{\mathrm{ELBO}}(\theta)$ suffer from very high variance as resampling steps involve sampling from discrete distributions (high-variance REINFORCE estimators).

Variational inference meets particle filters

- Dropping resampling gradient terms has been used (Maddison et al., *NIPS* 2017; Naesseth et al., *AISTATS* 2018; Le et al., *ICLR* 2018) but can be problematic.

- **Proposition**. As $N \to \infty$, the expectation of the ELBO gradient estimate *dropping resampling terms* converges to

$$\sum_{t=1}^{T} \int \nabla_\theta \log p_\theta(x_t, y_t | x_{t-1}) p_\theta(x_{t-1}, x_t | y_{1:t}) \mathrm{d}x_{t-1} \mathrm{d}x_t$$

  whereas

$$\nabla_\theta \ell(\theta) = \sum_{t=1}^{T} \int \nabla_\theta \log p_\theta(x_t, y_t | x_{t-1}) p_\theta(x_{t-1}, x_t | y_{1:T}) \mathrm{d}x_{t-1} \mathrm{d}x_t.$$

- For slow mixing processes, those two quantities will differ significantly.

- Let $\mathcal{U}(\alpha, \beta) := \{\text{distributions with marginals } \alpha \text{ and } \beta\}$. Any $\mathcal{P} \in \mathcal{U}(\alpha, \beta)$ can "transport" α to β, i.e.

$$\beta(\mathrm{d}x') = \int \mathcal{P}(\mathrm{d}x, \mathrm{d}x') = \int \alpha(\mathrm{d}x)\mathcal{P}(\mathrm{d}x'|x).$$

- The Optimal Transport (OT) between α and β is given by

$$\mathcal{P}^{\mathrm{OT}} = \arg \min_{\mathcal{P} \in \mathcal{U}(\alpha,\beta)} \mathbb{E}_{(X,X') \sim \mathcal{P}}\left[||X - X'||^2\right],$$

  and $W_2^2(\alpha, \beta) = \mathbb{E}_{(X,X') \sim \mathcal{P}^{\mathrm{OT}}}\left[||X - X'||^2\right]$ is the squared 2-Wasserstein metric.

- If α, β have densities, $\mathcal{P}^{\mathrm{OT}}(\mathrm{d}x'|x) = \delta_{T(x)}(\mathrm{d}x')$ where T is the **Optimal Transport** map, i.e. if $X \sim \alpha$ then $X' = T(X) \sim \beta$.

- **Application to PF**: consider $\alpha = p_\theta(x_t|y_{1:t-1})$ and $\beta = p_\theta(x_t|y_{1:t})$. If we could compute T and differentiate it, we would have no resampling and a differentiable estimate of $\ell(\theta)$.

- In practice, N is finite and $\alpha_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{X_t^i}, \; \beta_N = \sum_{i=1}^{N} w_t^i \delta_{X_t^i}$.
- **Problem 1**: Computing $\mathcal{P}^{\mathrm{OT}}$ is $O(N^3 \log N)$, not parallelizable nor differentiable (Reich, SIAM Sci Comp 2013).
  - **Solution**: Use entropy-regularized OT (Cuturi, 2013).
- **Problem 2**: $\mathcal{P}^{\mathrm{OT}}(\mathrm{d}x'|x) \neq \delta_{T(x)}(\mathrm{d}x')$ for empirical measures.
  - **Solution**: Use ensemble transform (Reich, 2013; Cuturi & D., 2014): $X' = \int x' \mathcal{P}^{\mathrm{OT}}(\mathrm{d}x'|X)$ at the cost of introducing bias in $\widehat{\ell}(\theta)$.
- Combined to reparameterization trick, this provides differentiable PF.

Entropy Regularized Transport

- For any $\epsilon > 0$, define for $\mathbf{a} = (1/N, ..., 1/N)$, $\mathbf{b} = (w^1, ..., w^N)$ and $c_{i,j} = ||X_t^i - X_t^j||^2$

$$\mathrm{OT}_\epsilon(\alpha_N, \beta_N) = \min_{\mathbf{P} \in \mathcal{S}(\mathbf{a},\mathbf{b})} \sum_{i,j=1}^{N} p_{i,j} \left(c_{i,j} + \epsilon \log \frac{p_{i,j}}{a_i b_j} \right).$$

- Regularized OT can be solved using Sinkhorn's algorithm (Cuturi, *NIPS* 2013), linear convergence.
- Sinkhorn's recursion is differentiable (Genevay et al., *AISTATS* 2018): use implicit differentiation of fixed point.
- Differentiable Ensemble Transform (DET)

$$\bar{X}_t^i = N {\textstyle\sum_k} p_{k,i}^{\mathrm{OT},\epsilon} \, X_t^k.$$

- DET + reparam trick for transition $f_\theta(x_t|x_{t-1})$ = Differentiable Particle Filtering.

- **Ensemble Transform**: Let $\bar{\beta}_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{\bar{X}^i}$ where $\bar{X}^i$ are obtained using DET between α_N, β_N. α, β are two measures with λ-Lipschitz OT. Then for bounded 1-Lipschitz function ψ, we have

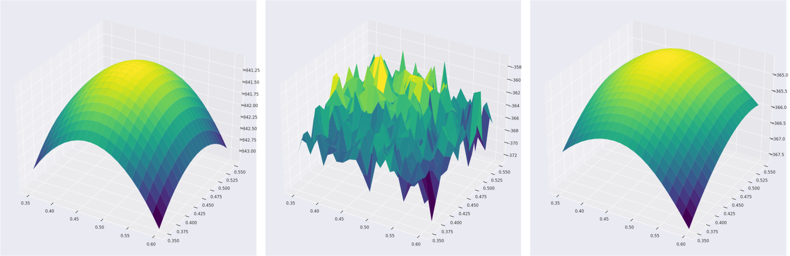
$$\left|\beta_N(\psi) - \bar{\beta}_N(\psi)\right| \leq 2\lambda^{1/2}\mathcal{E}^{1/2}\left[\mathfrak{d}^{1/2} + \mathcal{E}\right]^{1/2} + \max\{\lambda, 1\}\left[W_2(\alpha_N, \alpha) + W_2(\beta_N, \beta)\right] \quad (1)$$

  where $\mathfrak{d} := \sup_{x,y\in\mathcal{X}} |x - y|$ and $\mathcal{E} = W_2(\alpha_N, \alpha) + W_2(\beta_N, \beta) + \sqrt{2\epsilon \log N}$.

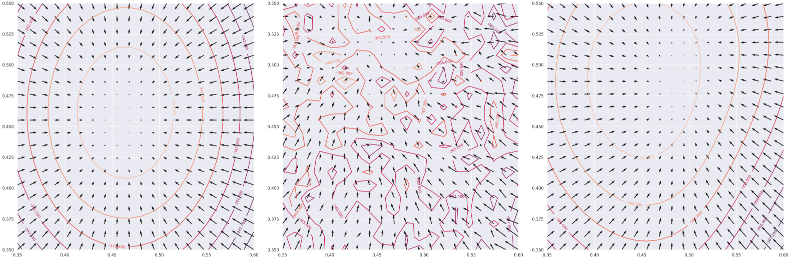
- **Consistency**: Under regularity assumptions, expectations w.r.t. filtering distributions and log-likelihood estimate converge as $N \to \infty$ and are consistent if $\epsilon = O(1/\log N)$.

Experiments: Linear Gaussian SSM

- $X_t|\{X_{t-1}\} \sim \mathcal{N}\left(\mathrm{diag}(\theta_1\ \theta_2)X_{t-1}, 0.5\mathbf{I}_2\right), \quad Y_t|\{X_t\} \sim \mathcal{N}(X_t, 0.1 \cdot \mathbf{I}_2)$



Log-likelihood $\ell(\theta)$, standard PF estimate $\widehat{\ell}(\theta; \mathbf{u})$ and differentiable PF estimate



Gradient $\nabla_\theta \ell(\theta)$, standard PF estimate $\nabla_\theta \widehat{\ell}(\theta; \mathbf{u})$ and differentiable PF estimate

Table 1: Mean & std of $\frac{1}{T}(\hat{\ell}(\theta; \mathbf{U}) - \ell(\theta))$

	Multinomial		DET	
θ_1, θ_2	mean	std	mean	std
0.25	-1.02	0.18	-1.02	0.18
0.50	-0.84	0.17	-0.85	0.17
0.75	-0.79	0.18	-0.79	0.18

Table 2: $10^3 \times$ RMSE[4] over 50 datasets

B	$\hat{\theta}^{\mathrm{MUL}}_{\mathrm{ELBO}}$	$\hat{\theta}^{\mathrm{DET}}_{\mathrm{ELBO}}$	$\hat{\theta}_{\mathrm{SMLE}}$
1	4.86	3.94	16.87
4	4.94	3.37	7.01
10	4.79	2.72	4.53
25	4.83	2.23	2.74

(left) Bias/std ELBO for standard PF & differentiable PF - (right) RMSE parameter estimates

Experiments: Robot Localization

- Given agent's initial state, S_1, and inputs a_t, one would like to infer its location given observations O_t.

- $S_t = (X_t^{(1)}, X_t^{(2)}, \gamma_t)$ where $(X_t^{(1)}, X_t^{(2)})$ are location coordinates and γ_t the robot's orientation. O_t are raw images, encoded to extract useful features using a NN E_θ, where $Y_t = E_\theta(O_t)$.

- Given actions $a_t = (v_t^{(1)}, v_t^{(2)}, \omega_t)$, we have

$$S_{t+1} = F_\theta(S_t, a_t) + \nu_t, \quad \nu_t \sim \mathcal{N}(\mathbf{0}, \Sigma_F),$$
$$Y_t = G_\theta(S_t) + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(\mathbf{0}, \sigma_G^2 I),$$

Experiments: Robot Localization

- Set up from (Jonschkowki et al. 2018) with DeepMind data: 'true' trajectories are given for each maze with state, action and raw 32×32 RGB pixel images O_t.

- As in (Wen et al., 2020), we consider a combination of losses

$$\hat{\mathcal{L}}_{\text{MSE}} = \frac{1}{T}\sum_{t=1}^{T} ||X_t^* - \sum_{i=1}^{N} w_t^i X_t^i||^2, \quad \hat{\mathcal{L}}_{\text{PF}} = -\frac{1}{T}\hat{\ell}(\theta),$$

$$\hat{\mathcal{L}}_{\text{AE}} = \sum_{t=1}^{T} ||D_\theta(E_\theta(O_t)) - O_t||^2,$$

  where $X_t^\star$ are the true states available from training data and $\sum_{i=1}^{N} w_t^i X_t^i$ are the PF estimate of $\mathbb{E}[X_t | y_{1:t}]$.

- The PF-based loss terms are not differentiable w.r.t. θ under traditional resampling schemes.

Experiments: Robot Localization

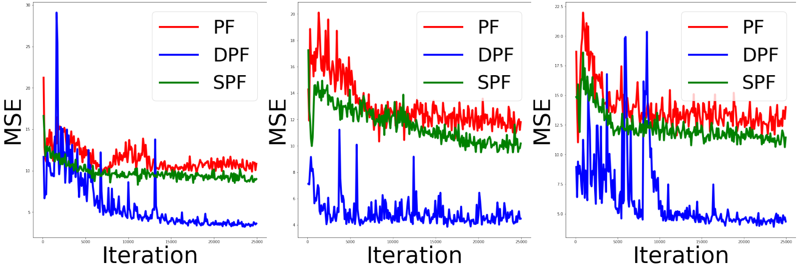


Figure: MSE of PF (red), SPF (green) and DPF (blue) estimates, evaluated on test data during training for 3 different mazes

Table: MSE and ± Standard Deviation evaluated on Test Data

	Maze 1	Maze 2	Maze 3
DET	$3.55_{\pm 0.20}$	$4.65_{\pm 0.50}$	$4.44_{\pm 0.26}$
MUL	$10.71_{\pm 0.45}$	$11.86_{\pm 0.57}$	$12.88_{\pm 0.65}$
SOFT	$9.14_{\pm 0.39}$	$10.12_{\pm 0.40}$	$11.42_{\pm 0.37}$

Discussion

- Differentiable particle filter = Regularized OT + reparameterization trick.
- End-to-end differentiable.
- Cost $O(N^2)$ vs $O(N)$ for available methods but additional cost negligible when used to train neural networks and regular PFs can be deployed once parameters have been estimated.
- DPF could be potentially sped up (Altschuler et al., 2019; Scetbon & Cuturi, 2020).
- Sharp quantitative results are still missing!

Main References

- A. Corenflos et al., Differentiable particle filtering using entropy-regularized optimal transport, arXiv:2102.07850 - ICML 2021.
- M. Cuturi, Sinkhorn distances: Lightspeed computation of optimal transport, *NIPS* 2013.
- R. Jonschkowski et al. Differentiable particle filters: End-to-end learning with algorithmic priors, *RSS* 2018.
- T. A. Le et al., Auto-encoding sequential Monte Carlo, *ICLR* 2018.
- X. Ma et al. Discriminative particle filter reinforcement learning for complex partial observations, *ICLR* 2020.
- C.J. Maddison et al., Filtering variational objectives, *NIPS* 2017.
- C. A. Naesseth et al., Variational sequential Monte Carlo, *AISTATS* 2018.
- S. Reich, A nonparametric ensemble transform method for Bayesian inference, *SIAM J. Scien. Comp.* 2013.
- H. Wen et al., End-To-End Semi-supervised Learning for Differentiable Particle Filters, arXiv:2011.05748.