Resampling in Particle Filters - an Optimal Transport approach

Adrien Corenflos
(Dated: December 30, 2019)

We use recent advances in optimal transport to tackle the challenge of resampling in particle filters. In particular we propose one approximate differentiable scheme and two exact schemes to transform a weighted empirical distribution into a similar unweighted distribution.

Keywords: differentiable, resampling, particle filters, smc, optimal transport, divergence

I. INTRODUCTION & PREVIOUS WORKS

Particle filters have emerged as a state of the art inference technique for non linear State Space Models.

include description of Bootstrap filter

The resampling in section I is included to fight degeneracy of the particles and some alternatives have been proposed to find an optimal re-indexing index $\sigma : \{1, \dots, N\} \to \{1, \dots, N\}$ (see include ref about different schemes). This resampling step notoriously makes the particle filter non-differentiable with respect to their hyper-parameters as it creates a non-differentiable mapping , $(w_i, X_i) \mapsto (1/N, X_{\sigma(i)})$.

Most of these approaches rely only on the weights $(w_i)_i$ to provide a probabilistic resampling variable $(\sigma_i)_i$ that verifies $\mathbb{E}[\sigma_i = i] = w_i$ (reference on coherence of resampling schemes). These don't take into account the geometry of the space and create duplicates of particles.

On the other hand some approaches have been tried that take into account the distance between the particles. The most similar work to our approach is [9] where the author explicitly creates an optimal transport plan T such that $\tilde{X} = NTX$ mapping from the weighted particles to the unweighted ones.

The paper is organised as follows:

1. Approximate reweighting using optimal transport mapping and differentiability
2. Exact reweighting using global divergence minimization
3. Fast exact reweighting
4. Particle filter examples (probably learn the parameters of Lorentz system in dimension 3, and track with known parameters in higher dim, using section 2 and 3)

II. APPROXIMATE REWEIGHTING USING OPTIMAL TRANSPORT MAPPING

A. The algorithm

In [9] the author defines the planning matrix $T \in \mathbb{R}^N$ as the solution to the optimal transport problem:

$$T = \operatorname{argmin} \sum_{i,j} T_{i,j} d(X_i, X_j) \text{ with}$$
$$\forall j, \sum_i T_{i,j} = w_i \text{ and } \forall i, \sum_j T_{i,j} = 1/N$$

In practice the distance d is chosen to be the euclidean distance or the Manhattan distance. However optimal transport calculation scales as $O(N^3)$, which makes the approach in [9] unfeasible for cases where a large number of particles need to be used.

Regularized optimal transport has emerged as way to solve such constraints [2] where the optimal transport problem is replaced with the following include definition of the optimisation problem. An alternative formulation that we will use in the rest of this article is given by [**?**], copy the algorithm. In their paper they moreover prove the differentiability of their scheme. We combine their approach and the one in [9] to provide a differentiable rebalancing scheme:

$$\tilde{X}_\epsilon = NT_\epsilon^t X$$

where T is the solution of $L_\epsilon(w, X, \frac{1}{N}, X)$.

We illustrate the behaviour of type of rebalancing in figure 1 The EMD in python POT doesn't work, need to raise an issue on the github, the "exact scheme" is sinkhorn with very small reg, where $X \sim \sum_i w_i, \delta_{X_i}$, the X_i's are i.i.d. $U([-2, 2])$ and $w_i \propto f(X_i)$ where f is the pdf of a normal distribution $\mathcal{N}(\text{loc} = 0.5, \text{scale} = 1)$ and $N = 500$, i.e. $\frac{1}{N} = 0.002$

In practice the resulting weighting $NT_\epsilon w$ will not be totally flat due to the fact that T is an approximation: section II A shows how the weights are distributed depending on the regularisation. Moreover, their value doesn't depend on the past. We will therefore set the value of $NT_\epsilon w \simeq 1/N$ to $\frac{1}{N}$ and its gradient w.r.t. w and X to $\mathbf{0}$ in the rest of the article.

B. Gradients

As discussed in [6] the gradients for $(w_i, X_i) \mapsto \alpha, \beta$ are readily available using the implicit function theorem.

ϵ	0.01	0.05	0.10	0.50
mean	1.995	1.998	1.998	2.001
std	0.152	0.099	0.089	0.049
quantile 10%	1.801	1.874	1.875	1.919
median	2.009	2.020	2.023	2.015
quantile 90%	2.197	2.106	2.096	2.050

TABLE I. distribution of rebalanced weights ($\times 1000$) as a function of epsilon

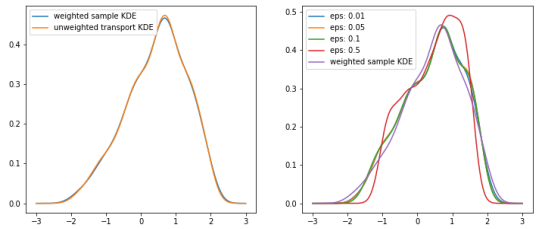

FIG. 1. Guaussian KDE for different transport schemes left: exact transport scheme, right: regularized transport scheme for different values of ϵ

Moreover the potentials and the transport matrix are linked via:

$$T_\epsilon = \exp \tfrac{1}{\epsilon^2}[\alpha \oplus \beta - \mathrm{C}] \cdot \alpha \otimes \beta$$

which in our case can be rewritten as:

$$T_\epsilon = \frac{\exp\left(\alpha^t + \beta - \frac{C^2}{2}\right)}{\epsilon^2} w^t / N$$

where $C = d(X_i, X_j)_{1 \le i,j \le N}$ and the exponential has to be understood element-wise.

This in turn means that the mapping $(w_i, X_i) \mapsto \tilde{X}_\epsilon$ is differentiable with gradient given by $\mathrm{d}\tilde{X}_\epsilon = N \mathrm{d}T_\epsilon X + N T_\epsilon \mathrm{d}X$

Some problems with the gradients coming from geomloss: they are only coded for the sake of differentiating the loss function, not the potentials. I'm chatting with Jean Feydy, I've patched by using eq (27) in [5] but need to check consistency formally

To illustrate the fact that our method works we compute the analytical gradient $\frac{\partial}{\partial \mathrm{loc}} \left(\frac{1}{N} \sum_{1 \le i \le N} w_i X_i \right)$ when w and X are given as in section II A for different values of the locations and the scale. The analytical gradient is given by:

$$\frac{1}{4} \int_{-2}^{2} \mathrm{pdf}(x, loc, scale) x \frac{x - loc}{\mathrm{scale}^2} dx$$

Numerically the result is shown in tables table III.

	-0.5	-0.25	0	0.25	0.5
0.25	0.25	0.25	0.25	0.25	0.25
0.5	0.25	0.25	0.25	0.25	0.24
1	0.16	0.18	0.18	0.18	0.16
1.5	0.09	0.09	0.10	0.09	0.09
2	0.05	0.05	0.05	0.05	0.05

TABLE II. Theoretical gradients, line: scale, column: location

	-0.5	-0.25	0	0.25	0.5
0.25	0.05	0.12	0.15	0.12	0.04
0.5	0.14	0.23	0.26	0.25	0.16
1	0.12	0.17	0.19	0.17	0.114
1.5	0.05	0.05	0.05	0.05	0.05
2	0.02	0.02	0.02	0.02	0.02

TABLE III. AutoDiff Gradients, line: scale, column: location

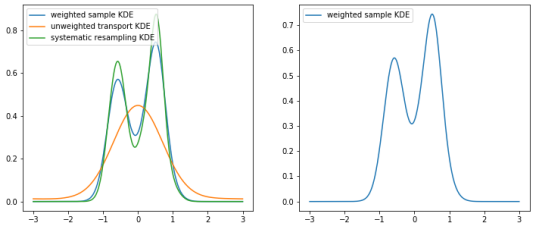

FIG. 2. Guaussian KDE for different transport schemes left: exact transport scheme and systematic resampling, right: regularized transport scheme for different values of ϵ

III. EXACT REWEIGHTING USING GLOBAL DIVERGENCE MINIMIZATION

While the scheme in section II has the advantage of providing a gradient for the resampling schemes, it suffers from the inconvenient of approximating the reweighted sample through a linear mapping only. While this can be useful for cases where the latent space regular enough, some more complex cases may arise where a linear mapping doesn't suffice anymore. This is true in particular when most of the particles are degenerate and when the distribution is multimodal Don't know if that's a good example, need Arnaud

This is illustrated in figure 2 where the example in section II has been modified in the following way: out of 500 particles distributed uniformly between -2 and 2, 250 have been chosen randomly and their weight set to the normal pdf with mean 0.5 and scale 0.1 and the rest to the normal pdf with mean -1 and scale 0.1. This is equivalent to a NESS of 41. As expected the linear transportation doesn't provide a good approximation of the degenerate distribution.

To alleviate this issue we propose to modify the algorithm above by using an approach similar to [8] and find unweighted $X_e\tilde{p}silon$ minimising a certain distance

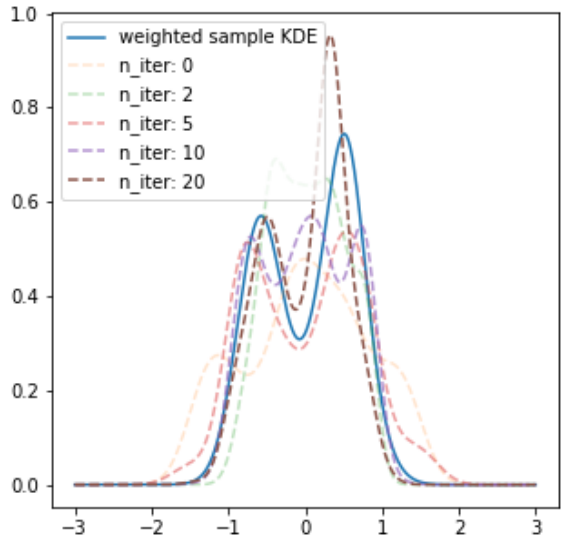

FIG. 3. Guaussian KDE for learnt unweighted distribution. Convergence of the unweighted distribution to the weighted one as the number of epochs increases

w.r.t. the degenerate sample (w, X). As discussed in [5] the entropic bias induced by the regularisation term in regularised Optimal Transport results in concentrated distributions, which is the opposite of our goal. Following [8] we therefore use the unbiased version of the Sinkhorn iterates: the Sinkhorn divergence. Include some background with feasibility [5] and proof of metrisation.

We therefore aim at minimising

$$\begin{aligned} S_\epsilon(\cdot)_{X,w} = & OT_\epsilon\left(\frac{1}{N}, \cdot, w, X\right) \\ & - 0.5OT_\epsilon\left(\frac{1}{N}, \cdot, \frac{1}{N}, \cdot\right) \\ & - 0.5OT_\epsilon(w, X, w, X) \end{aligned}$$

This is done via a gradient descent leveraging Adam. It converges but maybe not the best thing to do...

IV. FAST EXACT REWEIGHTING

I want to modify the scheme in section III by doing something similar:

Instead of going minimising the divergence by starting from the (1/N, X)

$$\begin{aligned} S_\epsilon(\cdot)_{X,w} = & OT_\epsilon\left(\frac{1}{N}, \cdot, w, X\right) \\ & - 0.5OT_\epsilon\left(\frac{1}{N}, \cdot, \frac{1}{N}, \cdot\right) \\ & - 0.5OT_\epsilon(w, X, w, X) algorithm \end{aligned}$$

It may be worth taking a lookahead approach: given a schedule $0 = t_1 < t_i < t_N = 1$, we can successively minimise for $k = 1, ..., N-1$:

$$X_{k+1} = \text{argmin}_z(S_\epsilon(X_k, w_k, z, w_{k+1})$$

Where $w_k = t_k \frac{1}{N} + (1 - t_i)w$ and $X_0 = X$. If the step size is taken small enough, X_k, w_k and X_{k+1}, w_{k+1} will be close enough to allow for an immediate minimisation at every step.

Algorithm 1 Incremental learning

Input: x, w, t, N, λ, K {Number of gradient descent loops},
Compute: $w_i := t_i \frac{1}{N} + (1 - t_i)w, i = 1..N$
Initialize $x_i = x$.
for $i = 1$ **to** N **do**
  $y_k = x_i$
  for $k = 1$ **to** K **do**
  $y_k = y_k - \lambda \nabla_y S_\epsilon(S_\epsilon(x_i, w_i, y_k, w_{i+1})$
  end for
$x_i = y_k$
end for

[1] Mathieu Blondel, Vivien Seguy, and Antoine Rolet. Smooth and sparse optimal transport, 2017.
[2] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transportation distances, 2013.
[3] Marco Cuturi and Gabriel Peyré. Semidual regularized optimal transport. *SIAM Review*, 60(4):941–965, Jan 2018.
[4] Tarek A. El Moselhy and Youssef M. Marzouk. Bayesian inference with optimal maps. *Journal of Computational*

Physics, 231(23):7815–7850, Oct 2012.

[5] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences.

[6] Jean Feydy and Alain Trouvé. Global divergences between measures: from Hausdorff distance to Optimal Transport. working paper or preprint, August 2018.

[7] J. F. G. de Freitas, M. Niranjan, A. H. Gee, and A. Doucet. Sequential monte carlo methods to train neural network models. *Neural Computation*, 12(4):955–993, 2000.

[8] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences, 2017.

[9] Sebastian Reich. A non-parametric ensemble transform method for bayesian inference, 2012.