Differentiable Particle Filter

Abstract

Sequential Monte Carlo (SMC or Particle Filters) methods are a set of powerful techniques for continuous state space models. Additionally to providing a belief for the state of an observed system, they also provide an estimate of its likelihood which is non-differentiable with respect to the model parameters. This is due to the fact that standard resampling schemes consist in a non-smooth re-indexing of the Monte Carlo sample. In this article we propose two new resampling schemes that rely on regularised optimal transport techniques and are respectively differentiable but biased, and non-differentiable but optimal in a certain metric. Furthermore we assess the behaviour of the methods in a linear setup by comparing with the Kalman Filter and discuss the influence of the hyper-parameters.

1 INTRODUCTION

Particle Filters (see Gordon, Salmond, and Smith 1993) offer an efficient way of performing posterior inference in otherwise intractable non-linear state space models and provide an unbiased estimate of the likelihood of the state space model parameters given observed data. Formally particle filters are interested in estimating state space hidden Markov models described by an unobserved state $X_t \in \mathbb{R}^{d_X}$ following $X_t|(X_{t-1} = x) \sim f_t(\cdot|x), t > 0$ and $X_0 \sim \mu(\cdot) \in \mathcal{M}(\mathbb{R}^{d_X})$ and an observed process $Y_t|(X_t = x) \sim g_t(\cdot|x) \in \mathcal{M}(\mathbb{R}^{d_Y})$. They do this by keeping track of X_t in the form of a weighted sample (w_t^i, X_t^i) through a method called Sequential Importance Sampling, however, this technique applied by itself leads to weight degeneracy that needs to be mitigated (see Doucet and Johansen 2009), a well accepted way to fight this is through the use of a resampling scheme that replaces low weight particles with high weights ones (see Hol, Schon, and Gustafsson 2006).

In this paper we focus on regularised optimal transport as a resampling technique in two different ways: in Section 3 we use the planning matrix as a direct map from a weighted sample $(w_i, X_i) \sim \mathbf{X}$ to an equally weighted sample $(\mathbb{1}_N, \mathbf{Z}_i^\epsilon) \sim \mathbf{X}^\epsilon$, where $\mathbb{1}_N$ is the vector of size N filled with $\frac{1}{N}$ and show $\mathbf{X}^\epsilon \xrightarrow[\epsilon \to 0]{\mathcal{L}} \mathbf{X}$, because the mapping we use is differentiable w.r.t (w_i, X_i), it will provide a biased but differentiable resampling scheme, then in Section 4 we provide a novel algorithm that learns the optimal equally-weighted sample $(\mathbb{1}_N, \mathbf{Z}_i^\epsilon)$ in the ϵ-Sinkhorn divergence sense.

Our main contributions are two fold: we introduce the planning matrix of the regularised optimal transport as a differentiable resampling scheme and we introduce a novel resampling algorithm that guarantees optimality of the resampled particles.

The rest of the paper is organised as follows: in Section 2 we give a brief recapitulation of the Particle Filter (Gordon, Salmond, and Smith 1993), the biased Sinkhorn distances (Cuturi 2013) and the Sinkhorn divergence (Feydy et al. 2018), in Section 3 we introduce the differentiable particle filter, discuss its behaviour and provide examples, in Section 4 we introduce an optimal resampling scheme and provide examples for it, and finally we conclude with future possible extensions and improvements.

2 BACKGROUND

2.1 Particle Filters

Particle filters have emerged as a standard technique for non linear data assimilation and approximates the filtering (posterior) distribution $X_t|(Y_t = y, X_{t-1}))$ using a weighted set of N samples (w_i, X_i) which are updated following a predict-update approach. Any given quantity of interest $\mathbb{E}[\phi(X_t)]|(Y_t = y, X_{t-1}))]$ can then estimated as a weighted average $\sum_i w_i \phi(X_i)$ with variance given by a central limit theorem (see Chopin et al. 2004). The "predict step" consists in proposing particles $X_t^i \sim p(\cdot|\mathbf{X}_{t-1} = X_{t-1}^i)$ whose weights will then be updated as per the bayes

formula:

$$w_t^i = p(X_t^i|\mathbf{Y}_t = \mathbf{y}, X_{t-1}^i) \quad (1)$$
$$\propto p(Y_t|X_t^i) \cdot p(X_t^i|X_{t-1}^i) \quad (2)$$
$$= p(Y_t|X_t^i) \cdot w_{t-1}^i \quad (3)$$

These will then be normalised to sum to 1. Additionally the marginal likelihood $P(\mathbf{Y}_{u=1..t}|\mathbf{X}_{u=1..t}, \theta)$ up to time time T where θ are model parameters can be estimated through Equation (2) prior to normalisation.

However, as time passes, the weights suffer a well-known degeneracy problem, that is all weights will converge to 0 apart from 1, hence not providing a good distributional estimate for the posterior. This has traditionally been mitigated by ancestry resampling: instead of proposing $X_t^i \sim p(\cdot|\mathbf{X}_{t-1} = X_{t-1}^i)$, we propose $X_t^i \sim p(\cdot|\mathbf{X}_{t-1} = X_{t-1}^{a_i})$ where $a_i \in 1, .., N$ is any index sampling such that $\mathbb{E}[\sum_i \mathbb{I}(a_i = j)|\mathbf{w}] = Nw_j, \forall j \in 1, ..., N$ (see Doucet and Johansen 2009). In practice this is only done when the efficient sample size (ESS) $\frac{1}{\sum_i w_i^2}$ is lower than a certain threshold (usually 50% of the sample size N).

This is summarised in Algorithm 1 and Algorithm 2. Because of the reparametrization trick (see Kingma and Welling 2013), only Algorithm 2 makes the particle filter non-differentiable with respect to its inputs.

Algorithm 1 Bootstrap Particle Filter

Input: X_i, w_i, y, N, X, L {Inputs at time $t > 0$},
if ESS $< N \cdot$ threshold **then**
  Resample
end if
for $i = 1$ **to** N **do**
  Propose: Sample X_i $p(X_{t+1}|X_t = X_i)$
  Update: Compute $w_i* = p(Y_t = y|X_t = X_i)$
end for
Compute log-likelihood update: $L+ = \log \sum_i w_i$
for $i = 1$ **to** N **do**
  $w_i = \frac{w_i}{\sum_i w_i}$
end for

Algorithm 2 Generic resampling

Input: X_i, w_i, N,
for $i = 1$ **to** N **do**
  Sample: a_i {satisfying hypotheses, for example $a_i \sim$ Multinomial($\mathbf{w}$)}
  Set: $w_i = \frac{1}{N}$, $X_i = X_{a_i}$
end for
output $\mathbf{w}, \mathbf{X}$

In this article, similarly to (Reich 2012; Graham and Thiery 2019) we consider an ensemble approach to particle filtering: instead of the resampling scheme $(w_i, X_i) \mapsto (\frac{1}{N}, X_{a_i})$ we provide two new mappings: we present a biased ensemble reweighting (Section 3) that comes from the planning matrix solution of a regularised Optimal Transport problem $(w_i, X_i) \mapsto (\frac{1}{N}, (M_\epsilon^{\mathbf{w},\mathbf{X}} X)_i)$, and we present a learnt optimal recentering of the particles learnt to minimize the Sinkhorn Divergence (see Genevay, Peyré, and Cuturi 2017; Feydy et al. 2018) $(w_i, X_i) \mapsto (\frac{1}{N}, Z_i^\epsilon)$.

2.2 REGULARISED OPTIMAL TRANSPORT

2.2.1 Optimal Transport

Optimal transport is interested in computing a distance between measures. Formally, given two empirical probability measures $\alpha, \beta \in \mathcal{M}_1^+(\mathcal{X})$, and given a symmetric positive cost function $C : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, it is interested in computing both the minimum and the minimising argument of the functional $\pi \mapsto \int_{\mathcal{X}^2} C(x, y) d\pi$ for π belonging to the simplex

$$S_{\alpha,\beta} = \left\{ \pi \in \mathcal{M}_1^+(\mathcal{X} \times \mathcal{X}) | \int \pi(\cdot, dy) = \alpha, \int \pi(dx, \cdot) = \beta \right\}$$

2.2.2 Sinkhorn distances

If the supports of α and β are of size N, this is known to scale in $O(n^3)$. Cuturi 2013 shows that a regularised version of the algorithm can be considered instead:

$$\mathbf{OT}_\epsilon := \min_{\pi \in S_{\alpha,\beta}} \int_{\mathcal{X}^2} C(x, y) d\pi + \epsilon KL(\pi || \alpha \otimes \beta)$$

Thanks to Fenchel-Rockafellar theorem, this can be rewritten (see Feydy et al. 2018; Peyré and Cuturi 2018) using dual functions $f, g \in \mathcal{C}(\mathcal{X})$:

$$\mathbf{OT}_\epsilon = \max_{f,g \in \mathcal{C}(\mathcal{X})} \langle \alpha, f \rangle + \langle \beta, g \rangle \quad (4)$$
$$- \epsilon \langle \alpha \otimes \beta, \exp\left(\frac{1}{\epsilon}(f \oplus g - C)\right) - 1 \rangle$$
$$\text{with: } \pi_\epsilon = \exp\left(\frac{1}{\epsilon}(f \oplus g - C)\right) \cdot \alpha \otimes \beta \quad (5)$$

This has been key for the recent development of computational optimal transport (Peyré and Cuturi 2018) as it translates a problem over a matrix to a problem over two related vectors. Moreover the optimality condition on f and g can be written in a fixed-point theorem form: if $\alpha = (w_i^X, X_i)_{1 \le i \le N}, \beta = (w_j^Y, Y_j)_{1 \le j \le M}$ (Feydy et al. 2018) which one only has to iterate successively to until convergence.

$\forall i, j$

$$f_i = -\epsilon \text{LSE}_k(\log w_k^Y + \frac{1}{\epsilon} g_k - \frac{1}{\epsilon} C(X_i, Y_k)) \quad (6)$$
$$g_j = -\epsilon \text{LSE}_k(\log w_k^X + \frac{1}{\epsilon} f_k - \frac{1}{\epsilon} C(X_k, Y_j)) \quad (7)$$

2.2.3 Sinkhorn divergences

Crucially $\mathbf{OT}_\epsilon(\cdot, \alpha)$ doesn't reach its minimum in α, which means that the solution of $\min_\beta \mathbf{OT}_\epsilon(\beta, \alpha)$ may actually lay far from α. This has been outlined first in Genevay, Peyré, and Cuturi 2017, and a solution is to consider instead the so-called "Sinkhorn Divergence"

$$\begin{aligned}\mathcal{W}_\epsilon(w_X, X, w_y, Y) := &\ \mathrm{OT}_\epsilon(w_X, X, w_y, Y) && (8)\\ &- 0.5\mathrm{OT}_\epsilon(w_X, X, w_X, X) && (9)\\ &- 0.5\mathrm{OT}_\epsilon(w_Y, Y, w_Y, Y) && (10)\end{aligned}$$

An important point is that the symmetric Optimal transport problems Equations (9) and (10) can be solved faster than the non-symmetric one, hence the computational burden is controlled by Equation (8)

3 DIFFERENTIABLE RESAMPLING

Given a weighted empirical distribution $(w_i, X_i)_{1\leq i\leq N} \in \mathcal{M}_1^+(\mathbb{R}^d$ we are interested in finding a "good enough" unweighted sample in the sense that for any $\phi := \mathbb{R}^d \to \mathbb{R}$, $\sum_{1\leq i\leq N} w_i\phi(X_i) \approx \frac{1}{N}\sum_{1\leq i\leq N}\phi(X_i)$ with small enough variance (Li et al. 2015). While most approaches consider a bootstrapping of the particle based on the weights: $(w_i, X_i) \mapsto (\frac{1}{N}, X_{a_i})$ this provides a non-differentiable mapping that prevents exact propagation of the gradient through the resampling step. Because of this and to prevent high variance in the gradient estimation, recent works by Maddison et al. 2017; Naesseth et al. 2017; Le et al. 2017 that link SMC and Variational Inference simply ignore the additional impact of resampling and tweak the AutoDiff scheme to propagate the gradient of particles that were not discarded at the resampling step only. As discussed in their papers this provides a biased estimator of the likelihood.

3.1 OPTIMAL TRANSPORT MAP RESAMPLING

To the best of our knowledge the first paper to have introduced Optimal Transport map as an ensemble technique for resampling is Reich 2012, and the method has since been applied in Graham and Thiery 2019 and Jacob, Lindsten, and Schön 2016 to provide a local mapping from prior to posterior and to couple same-seed particle filters trajectories in order to compute sensitivities with respect to its hyper-parameters.

The paradigm introduced in Reich 2012 can be phrased as follows. Let $(w, X)_i$ be a weighted sample before resampling and $C \in \mathbb{R}^d \times \mathbb{R}^d$ be a symmetric positive cost matrix, and lets consider the following Optimal Transport problem:

$$\mathbf{U} = \mathrm{argmin}_{M\in S_{\mathbf{w},\mathbf{1}_N}} \sum_{i,j} C_{i,j}, M_{i,j} \qquad (11)$$

As discussed in Reich 2012, this matrix constitutes a Markov Chain $\mathbf{P} := N\mathbf{U}$ and the mapping $\tilde{X} = \mathbf{P}\mathbf{X}$ provides a consistent estimate of $(w_i, X_i)_i$.

However the function $(w_i, X_i)_i \to \mathbf{P}\mathbf{X}$ is difficult to compute (see Cuturi 2013).

3.2 DIFFERENTIABLE RESAMPLING

Instead of considering the non-regularised problem in Equation (11), we instead use the mapping resulting from the regularised version of the Optimal Transport problem Equation (5) (Cuturi 2013; Feydy et al. 2018) $(\mathbf{w}, \mathbf{X}) \to \mathbf{P}_\epsilon\mathbf{X}$.

Proposition 1. *The mapping*

$$(\mathbf{w}, \mathbf{X}) \to \mathbf{P}_\epsilon = \exp\left(\frac{1}{\epsilon}\left(\mathbf{f}^T + \mathbf{g} - \mathbf{C}(\mathbf{X}, \mathbf{X})\right)\right) \cdot \mathbf{w}^T$$

where $\mathbf{f}$ *and* $\mathbf{g}$ *are given by Equations* (6) *and* (7) *is differentiable, with:*

$$\forall i, j$$

$$\frac{\partial f_i}{\partial \cdot} = \frac{\partial}{\partial \cdot} - \epsilon LSE_k(\log w_k^Y + \frac{1}{\epsilon}g_k - \frac{1}{\epsilon}C(X_i, Y_k)) \qquad (12)$$

$$\frac{\partial g_j}{\partial \cdot} = \frac{\partial}{\partial \cdot} - \epsilon LSE_k(\log w_k^X + \frac{1}{\epsilon}f_k - \frac{1}{\epsilon}C(X_k, Y_j)) \qquad (13)$$

Proof. The system of equations Equations (6) and (7) defines an system of implicit functions to which we can apply the implicit function theorem. In this case the application of it is trivial as the relationship is linear. □

In practice this means that the gradients can be propagated by automatic differentiation at the last step of the Sinkhorn iterates only provided that the algorithm has converged.

When using automatic differentiation, this can be summarised as:

Algorithm 3 Biased resampling

Input: X_i, w_i, N, n_{steps}
Stop registering gradients
Initialise: $\mathbf{f}$, $\mathbf{g}$

for $i = 1$ **to** $n_{\text{steps}} - 1$ **do**
  evaluate Equation (6) and Equation (7) simultaneously
end for
Register gradients
Set gradients of $\mathbf{f}$, $\mathbf{g}$ to 0
Evaluate Equation (6), Equation (7)
output $\frac{1}{N}\mathbb{1}_N$, $\mathbf{P}\mathbf{X}$ with P given by Equation (5)

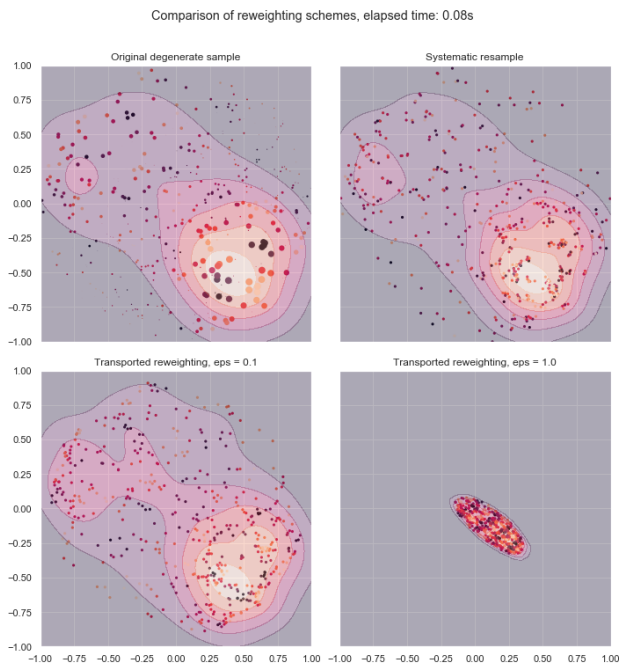

Figure 1: Biased transport comparison

3.2.1 Illustration

To illustrate the behaviour of the resampling scheme we consider a bimodal 2D distribution of 500 constructed as follows: 500 points $X_i \in \mathbb{R}^2$ are drawn uniformly within a circle of radius 1, then half the sample is randomly set to have weights proportional to $\mathcal{N}\left(\begin{pmatrix}-0.5\\0.5\end{pmatrix}, \begin{pmatrix}0.3 & 0.\\0. & 0.3\end{pmatrix}\right)$ and the other half $\mathcal{N}\begin{pmatrix}0.5 & 0.1 & 0.\\-0.5\text{'} & 0. & 0.1\end{pmatrix}$, see Figure 1. This corresponds tp a multimodal distribution $X|\,||X||_2^2 \leq 1$ where $X \sim \mathcal{N}\left(\begin{pmatrix}-0.5\\0.5\end{pmatrix}, \begin{pmatrix}0.3 & 0.\\0. & 0.3\end{pmatrix}\right) + \mathcal{N}\left(\begin{pmatrix}0.5\\-0.5\end{pmatrix}, \begin{pmatrix}0.1 & 0.\\0. & 0.1\end{pmatrix}\right)$

Figure 1 illustrates a well-known problem of the regularised Sinkhorn algorithm: as the regularisation increases, the resulting transporting plan will converge to the one that minimizes $KL(\pi||\alpha \otimes \beta)$, in this case $\mathbf{w_X} \otimes \mathbb{1_N}$, hence the resulting resampling of $\mathbf{X}$: $\mathbf{X}_\epsilon = \mathbf{P}_\epsilon \mathbf{X}$ collapses to the weighted mean of the sample $\mathbf{X}$.

3.2.2 Application To A State Space Model

We now consider the following noisy resonator:

$$x_{\text{true}}(t) = \sin(t) + \frac{1}{2}\mathcal{N}(0,1), t = 0., 0.1, ..., 20 \quad (14)$$

That we model by the following 2D state space model

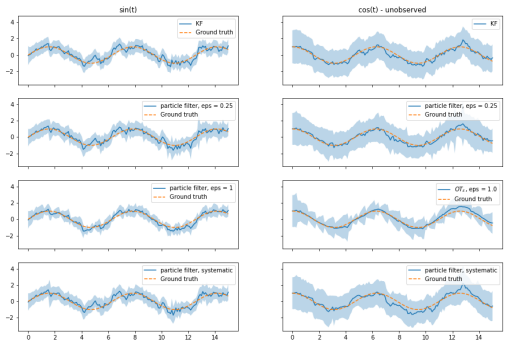

Figure 2: Filtering applied to Equation (15) Top: Kalman Filter, Second and Third: Biased Transport with $\epsilon = 0.25, 1$, Bottom: Systematic Resampling

$$x_0(t+dt) = x_0(t) + x_1(t)dt + N(0, \sigma_0^2) \quad (15)$$
$$x_1(t+dt) = x_1(t) - x_0(t)dt + N(0, \sigma_1^2) \quad (16)$$
$$y(t) = x_0(t) + N(0, \sigma_y^2) \quad (17)$$

Figure 2 compares Algorithm 3 for different regularisations with Algorithm 2 with systematic resampling (Li et al. 2015), all filters have 100 particles. The shaded zone corresponds to $\pm$2standard deviations

The collapsing phenomenon in the sample due to the regularisation as highlighted in Figure 1 is visible in Figure 2: each resampling step results in a decrease in variance of the sample.

3.2.3 Likelihood evaluation

Together with the resampling in Algorithm 2 the formula coming from Algorithm 1 provides an unbiased estimate of the likelihood for the state space model associated and also comes with a central limit theorem Chopin et al. 2004. However, the resulting function $\hat{\mathcal{L}}(\mathbf{y}|\theta)$ is not continuous and as a consequence not differentiable with respect to θ. On the other hand using Algorithm 3 provides a theoretically everywhere differentiable scheme for "recentering" the particles, albeit to the cost of biasedness in the resulting estimate.

4 OPTIMAL RESAMPLING

While the scheme in Section 3 has the advantage of providing a gradient for the resampling schemes, it suffers from the inconvenient of providing a collapsed estimate of the state post resampling, which in turns results in a biased estimate of the likelihood. Instead of learning a biased linear mapping, we can instead learn the best unweighted point

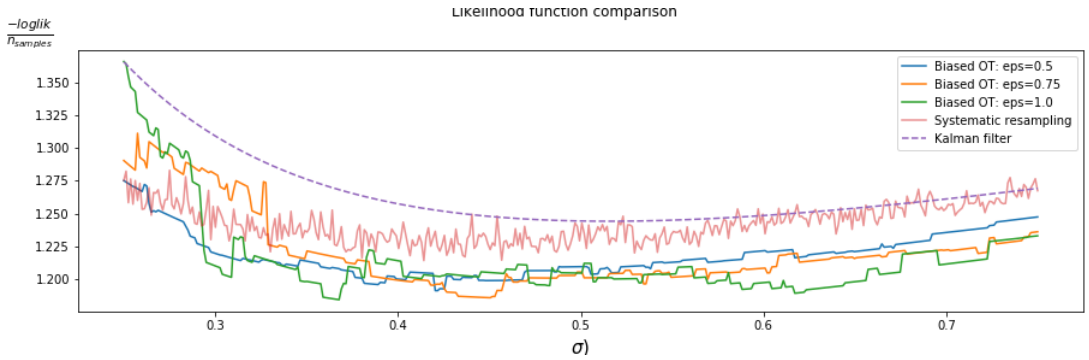

Figure 3: Likelihood estimate w.r.t. the observation log-error, 400 points

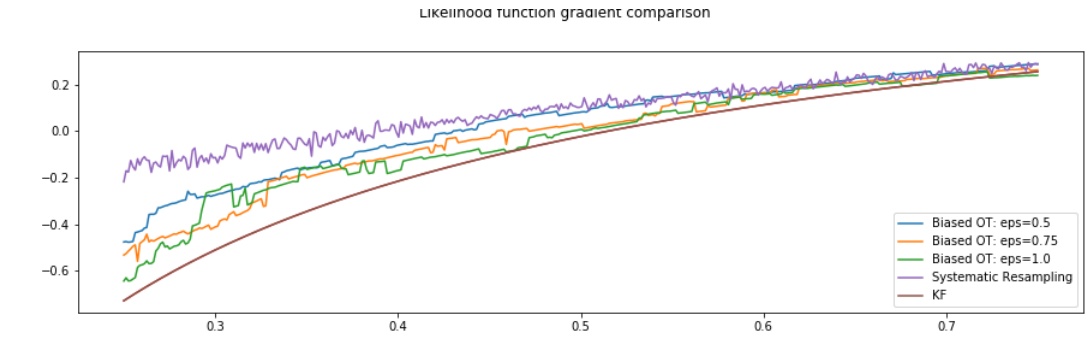

Figure 4: Gradient of the likelihood estimate w.r.t. the observation log-error

cloud minimising a distance from the weighted degenerate sample. To this end we consider the Sinkhorn divergence (see Section 2.2.3)

4.1 LEARNT POINTS CLOUD

As in Genevay, Peyré, and Cuturi 2017 we consider the optimisation problem given by the Wasserstein divergence $\mathcal{W}_\epsilon = 2OT_\epsilon(\alpha,\beta) - OT_\epsilon(\alpha,\alpha) - OT_\epsilon(\beta,\beta)$, where in our case α is the weighted degenerate sample $(w_i, X_i)_i$ and β is the target unweighted sample $(\frac{1}{N}, Z_i)_i$: this constitutes a gradient descent on Z with respect to the loss given by the Wasserstein divergence.

The resampling algorithm is therefore modified as in Algorithm 4

The behaviour of Algorithm 4 is shown in Figure 5

Algorithm 4 Optimal resampling

Input: X_i, w_i, N, n_{steps}, tolerance, λ {Learning rate}
Stop registering gradients
Initialise: f, g
$Z \leftarrow X$
for $i = 1$ **to** n_{steps} **do**
  if $\mathcal{W}_\epsilon(w, X, \frac{1}{N}, Z) <$ tolerance **then**
    Break
  end if
  $Z \leftarrow Z - \lambda \nabla_Z \mathcal{W}_\epsilon$
end for
output $\frac{1}{N}\mathbb{1}_N$, Z

5 CONCLUSION AND FUTURE WORKS

References

[Cho+04] Nicolas Chopin et al. "Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference". In: *The Annals of Statistics* 32.6 (2004), pp. 2385–2411.

[Cut13] Marco Cuturi. *Sinkhorn Distances: Lightspeed Computation of Optimal Transportation Distances*. 2013. arXiv: 1306.0895 [stat.ML].

[DJ09] Arnaud Doucet and Adam M Johansen. "A tutorial on particle filtering and smoothing: Fifteen years later". In: *Handbook of nonlinear filtering* 12.656-704 (2009), p. 3.

[Fey+18] Jean Feydy et al. *Interpolating between Optimal Transport and MMD using Sinkhorn Divergences*. 2018. arXiv: 1810.08278 [math.ST].

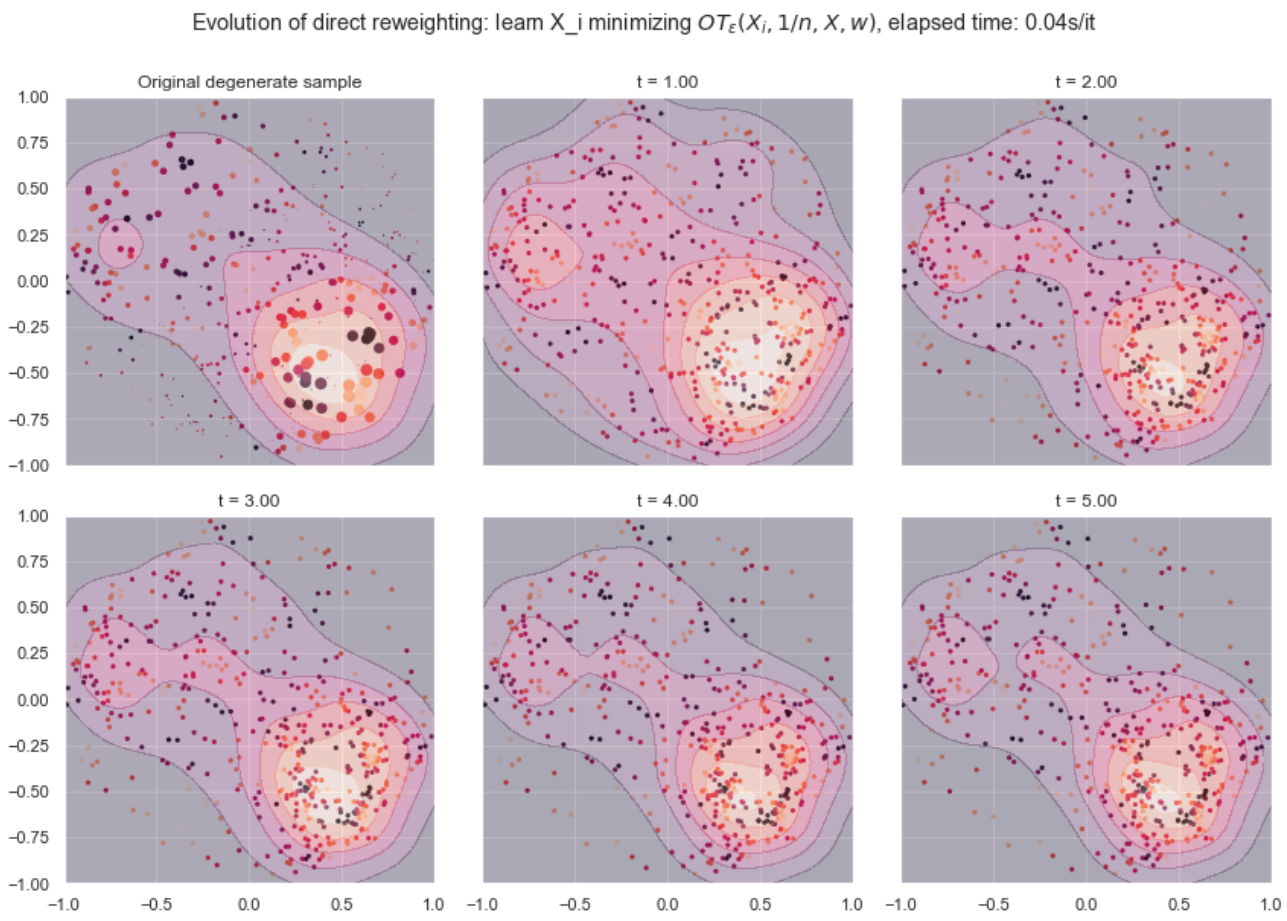

Figure 5: Gradient Flow for learning the unweighted sample
Top left: original sample, right and bottom: evolution of the gradient descent

[GPC17] Aude Genevay, Gabriel Peyré, and Marco Cuturi. *Learning Generative Models with Sinkhorn Divergences*. 2017. arXiv: 1706.00292 [stat.ML].

[GSS93] N. J. Gordon, D. J. Salmond, and A. F. M. Smith. "Novel approach to nonlinear/non-Gaussian Bayesian state estimation". In: *IEE Proceedings F - Radar and Signal Processing* 140.2 (1993), pp. 107–113. ISSN: 0956-375X. DOI: 10.1049/ip-f-2.1993.0015.

[GT19] Matthew M. Graham and Alexandre H. Thiery. *A scalable optimal-transport based local particle filter*. 2019. arXiv: 1906.00507 [stat.CO].

[HSG06] Jeroen D Hol, Thomas B Schon, and Fredrik Gustafsson. "On resampling algorithms for particle filters". In: *2006 IEEE nonlinear statistical signal processing workshop*. IEEE. 2006, pp. 79–82.

[JLS16] Pierre E. Jacob, Fredrik Lindsten, and Thomas B. Schön. *Coupling of Particle Filters*. 2016. arXiv: 1606.01156 [stat.ME].

[KW13] Diederik P Kingma and Max Welling. "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (2013).

[Le+17] Tuan Anh Le et al. "Auto-encoding sequential monte carlo". In: *arXiv preprint arXiv:1705.10306* (2017).

[Li+15] Tiancheng Li et al. "Resampling methods for particle filtering: identical distribution, a new method, and comparable study". In: *Frontiers of Information Technology & Electronic Engineering* 16 (Nov. 2015), pp. 969–984. DOI: 10.1631/FITEE.1500199.

[Mad+17] Chris J Maddison et al. "Filtering variational objectives". In: *Advances in Neural Information Processing Systems*. 2017, pp. 6573–6583.

[Nae+17] Christian A Naesseth et al. "Variational sequential monte carlo". In: *arXiv preprint arXiv:1705.11140* (2017).

[PC18] Gabriel Peyré and Marco Cuturi. *Computational Optimal Transport*. 2018. arXiv: 1803.00567 [stat.ML].

[Rei12] Sebastian Reich. *A non-parametric ensemble transform method for Bayesian inference*. 2012. arXiv: 1210.0375 [math.NA].