



HAL
open science

Continual Repeated Annealed Flow Transport Monte Carlo

Alexander G D G Matthews, Michael Arbel, Danilo J Rezende, Arnaud Doucet

► **To cite this version:**

Alexander G D G Matthews, Michael Arbel, Danilo J Rezende, Arnaud Doucet. Continual Repeated Annealed Flow Transport Monte Carlo. International Conference on Machine Learning 2022, 2022, Baltimore, United States. hal-03869105

HAL Id: hal-03869105

<https://hal.science/hal-03869105v1>

Submitted on 24 Nov 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Continual Repeated Annealed Flow Transport Monte Carlo

Alexander G. D. G. Matthews¹ Michael Arbel² Danilo J. Rezende¹ Arnaud Doucet¹

Abstract

We propose Continual Repeated Annealed Flow Transport Monte Carlo (CRAFT), a method that combines a sequential Monte Carlo (SMC) sampler (itself a generalization of Annealed Importance Sampling) with variational inference using normalizing flows. The normalizing flows are directly trained to transport between annealing temperatures using a KL divergence for each transition. This optimization objective is itself estimated using the normalizing flow/SMC approximation. We show conceptually and using multiple empirical examples that CRAFT improves on Annealed Flow Transport Monte Carlo (Arbel et al., 2021), on which it builds and also on Markov chain Monte Carlo (MCMC) based Stochastic Normalizing Flows (Wu et al., 2020). By incorporating CRAFT within particle MCMC, we show that such learnt samplers can achieve impressively accurate results on a challenging lattice field theory example.

1. Introduction

There are few algorithmic problems richer or more fundamental than the task of drawing samples from an unnormalized distribution and approximating its normalizing constant. In this paper we combine two important methods for this task specifically Sequential Monte Carlo (SMC) (Del Moral et al., 2006) and variational inference with Normalizing flows (NFs) (Rezende and Mohamed, 2015).

The first of these components, SMC samplers, are an importance sampling based algorithm. They are a principled way to combine annealing, resampling and Markov chain Monte Carlo (MCMC) and as such enjoy continuing popularity (Dai et al., 2020). One potential disadvantage of

SMC samplers is that at each successive temperature step in a SMC sampler there is importance sampling- albeit between adjacent distributions. For high dimensional target distributions this can lead either to estimators with high error or require many temperatures. Another potential disadvantage is that for finite particle numbers, SMC estimates of expectation w.r.t. the target distribution are biased. This can be mitigated by using the SMC sampler inside a Particle MCMC outer loop (Andrieu et al., 2010) at the cost of adding repeated SMC sampler calls. It is particularly desirable in this Particle MCMC context for the base SMC sampler to be fast and output low variance estimates of the normalizing constant.

The other component of our method is variational inference with NFs. Flow methods learn a differentiable invertible transformation (diffeomorphism) typically with a tractable Jacobian (Papamakarios et al., 2021). The learnt mappings are both flexible and fast to evaluate but there are still some challenges around using flows. The tractability of training with reverse Kullback–Leibler divergence comes at the cost of its well known mode seeking behaviour. Further placing the full burden of modelling on a parameterized flow can lead to high memory usage in contrast to non-parametric sampling algorithms. Finally, as diffeomorphisms, flows preserve topological properties of their input. This property can represent a challenge whenever there is topological mismatch between the, typically, simple base distribution and the complex target (Cornish et al., 2020).

Methods combining annealed samplers with normalizing flows have the potential to fix the limitations of each component part. The use of flows can reduce the variance from importance sampling. The MCMC steps can reduce the topological and representational burden on the flows. Used correctly, annealing can reduce the mode seeking behaviour of the variational objective. Resampling can be used to focus computation on promising samples. The idea of combining *fixed* transformations with annealing goes back as far as (Vaikuntanathan and Jarzynski, 2011). It is the *training* of the flows that still presents a challenge and the reason that such combined approaches have not met their full potential. Two recent papers target this problem (Wu et al., 2020; Arbel et al., 2021), but as we shall relate, neither is entirely satisfactory or as well suited to the task as the algorithm we propose.

¹DeepMind ²Université Grenoble Alpes, Inria, CNRS. Correspondence to: Alexander G. D. G. Matthews <alexmatthews@google.com>, Arnaud Doucet <arnaud-doucet@google.com>.

CRAFT is a method for estimating an SMC sampler augmented with interleaved normalizing flows. Unlike many other approaches in this area that incorporate sampling ideas, it is able to cope with full MCMC steps with Metropolis accept/reject corrections and SMC resampling steps. The CRAFT objective is not a standard variational bound. Rather it uses a KL divergence for each transition between temperatures. We find that as well as outperforming other methods for estimating the flows in practice (Section 4.1 and Section 4.2), our learnt sampler is useful as a fast inner loop for a Particle MCMC sampler (Section 4.3).

2. Method

In this section we describe and motivate both the sampling method and how to learn the required flows. Section 2.1 describes the method for fixed flows, which is equivalent to an SMC sampler with added fixed normalizing flows. In Section 2.2, we discuss conventional evidence lower bound (ELBO) methods for learning the flows and their limitations. Finally, we present the CRAFT training method in Section 2.3, and also explain how we improve over the AFT method on which CRAFT builds, by solving what we call the *sample replenishment problem*.

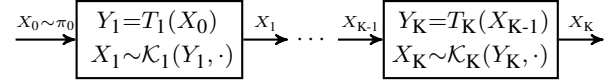
2.1. Sampler for fixed flows

We start by describing the sampler for fixed normalizing flows. In this case, it corresponds to a generalization of the standard SMC sampler which adds deterministic normalizing flows to the usual steps in SMC. This generalization reduces to a standard SMC sampler when the normalizing flows are the identity. Arbel et al. (2021) give a detailed history. The algorithm is described in Algorithm 3 which sequentially calls Algorithm 1.

We now describe the steps and derivation in more detail. We consider a sequence of distributions $(\pi_k(x))_{k=0}^K$ on \mathbb{R}^M where $\pi_k(x) = \frac{\gamma_k(x)}{Z_k}$ and $\gamma_k(x)$ can be evaluated pointwise. The final distribution $\pi_K(x)$ is the distribution with unknown normalizing constant Z_K that we wish to approximate. For the initial distribution $\pi_0(x)$ we assume that we can draw exact samples and that the normalizing constant $Z_0 = 1$ without loss of generality. The intermediate distributions enable us to transition smoothly (or in physics parlance ‘anneal’) from the tractable $\pi_0(x)$ to our goal $\pi_K(x)$. While they will sometimes be of interest themselves, they will often only be used to help us construct approximations. The method returns N weighted particles $(X_K^i, W_K^i)_{i=1}^N$ which can be used to provide an unbiased estimate of the normalizing constant Z_K and consistent estimates of expectations under the target π_K .

To build up intuition, consider a forward sampling process $\bar{\eta}$ producing a sequence $(X_k)_{k=0}^K$ defined on $\mathbb{R}^{M \times (K+1)}$

such that the final sample X_K approximates π_K . The process starts with an initial sample X_0 drawn from some distribution π_0 that is successively transformed using an interleaved sequence of K NFs $(T_k)_{k=1}^K$ and Markov transition kernels $(\mathcal{K}_k)_{k=1}^K$ with invariant distributions $(\pi_k)_{k=1}^K$:



Intuitively, the purpose of each flow T_k is to ‘transport’ samples from a density π_{k-1} corresponding to an annealing temperature $k-1$ towards the next density π_k at temperature k . Successive densities are usually similar, making the flow easier to learn compared to a flow directly transporting π_0 towards the target π_K . Since it is generally hard to find a flow T_k that *perfectly* transports between successive densities, the sampler employs two correction mechanisms: (1) the Markov transition kernel \mathcal{K}_k with invariant density π_k further diffuses the samples towards π_k and (2) Importance Sampling re-weights the samples to correct for any mismatch between π_k and the density $T_k^\# \pi_{k-1}$ obtained by transporting π_{k-1} using the flow T_k .

Importance Sampling. Ultimately, we are interested in re-weighting the final sample X_K to approximate the target π_K . This could be theoretically achieved by evaluating the target to marginal ratio π_K/η_K , where η_K is obtained by integrating the proposal distribution $\bar{\eta}$ over all previous variables $(X_k)_{k=0}^{K-1}$. To avoid computing such intractable integrals, we follow the standard approach in SMC and Annealed Importance Sampling (AIS) (Neal, 2001) of computing importance weights between the whole forward process $\bar{\eta}$ and an *augmented target* $\tilde{\pi}$ admitting π_K as marginal at time K . While there are multiple choices for the *augmented target* $\tilde{\pi}$, as discussed in (Del Moral et al., 2006), our choice of $\tilde{\pi}$ is made for tractability of the importance weights and corresponds to a formal *backward process* starting by sampling exactly from π_K and then sequentially generating samples backwards in time using the *reversal* transformations of each flow T_k and MCMC kernel \mathcal{K}_k . The reversal of each flow transformation is its inverse while the reversal of each forward Markov kernel is the Markov kernel $\tilde{\mathcal{K}}_k$ satisfying $\pi_k(x)\mathcal{K}_k(x, x') = \pi_k(x')\tilde{\mathcal{K}}_k(x', x)$. When choosing a reversible Markov kernel for the forward kernel \mathcal{K}_k , the reversal kernel is equal to the forward kernel, i.e. $\tilde{\mathcal{K}}_k = \mathcal{K}_k$. For our choice of augmented target, the *unnormalized* importance weights take the form:

$$w_K(x_{0:K-1}) = \prod_{k=1}^K G_k(x_{k-1}), \quad (1a)$$

$$G_k(x_{k-1}) := \frac{\gamma_k(T_k(x_{k-1}))}{\gamma_{k-1}(x_{k-1})} |\nabla T_k(x_{k-1})|. \quad (1b)$$

When the flow is the identity, equation (1b) reduces to the

ratio of successive densities, which is the standard annealed importance sampling expression. For non-identity flows the T_k dependent terms correct for the use of the flow. See Appendix A for a precise mathematical description of $\bar{\eta}$, $\bar{\pi}$ and the derivation of the weights (1a)-(1b).

The algorithm is then implemented sequentially to maintain a set of N particles $(X_k^{(i)}, W_k^i)_{i=1}^N$ consisting in pairs of samples and corresponding normalized IS weights at each time k where the samples are obtained using the forward generating process $\bar{\eta}$. The unnormalized IS weights w_k^i are computed recursively using $w_k^i = W_{k-1}^i G_k(X_{k-1}^i)$, thus allowing to compute the normalized IS weights W_k^i by normalizing over the set of particles $W_k^i = w_k^i / \sum_{j=1}^N w_k^j$. A consistent approximate expectation of a function f under the target π_k is then given by $\sum_{i=1}^N W_k^i f(X_k^{(i)})$. Moreover, an unbiased estimate Z_k^N of the normalizing constant Z_k is obtained sequentially from the previous one Z_{k-1}^N using the update $Z_k^N = Z_{k-1}^N (\sum_{i=1}^N w_k^i)$. If the NFs transport perfectly between the temperatures, this algorithm enjoys the property that the normalizing constant estimate Z_k^N has a zero variance and the corresponding approximate expectations are unbiased with a variance corresponding to the one of the true target.

Resampling. Up to this point, we have described AIS with additional fixed normalizing flows. We now go a step further to incorporate resampling, a key ingredient of SMC methods that has proved to be very beneficial (Arbel et al., 2021; Chopin, 2002; Del Moral et al., 2006; Hukushima and Iba, 2003). Resampling consists in randomly selecting a particle X_k^i with probability W_k^i and repeating the operation N times to construct a set of N particles approximately distributed according to π_k . This operation is equivalent to assigning a number of offspring N_k^i to each particle X_k^i and associating a uniform weight of $1/N$ to each offspring. The vector of all offspring $(N_k^i)_{i=1}^N$ is then drawn from a multinomial distribution with weights $(W_k^i)_{i=1}^N$ under the constraint that $\sum_{i=1}^N N_k^i = N$. The expectation under π_k of a function f is then approximated using $\sum_{i=1}^N \frac{N_k^i}{N} f(X_k^i)$ which is also an unbiased estimate of the weighted sum $\sum_{i=1}^N W_k^i f(X_k^i)$. Hence, resampling refocuses computational effort on promising particles (the ones falling in high density regions of π_k) whilst preserving the key analytic properties of the algorithm. However, to avoid the additional variance coming from the multinomial distribution, we only use it when the *effective sample size* $\text{ESS}_k^N := (\sum_{i=1}^N (W_k^i)^2)^{-1}$ of the particles falls beyond some predefined proportion $A \in [1/N, 1)$ (we use $A = 0.3$ in all our experiments) of the total number of samples N (Liu and Chen, 1995). Since resampling is followed by the Markov kernel step in each iteration of Algorithm 1 any degeneracy introduced during resampling can be reduced.

Algorithm 1 SMC-NF-step

- 1: **Input:** Approximations (π_{k-1}^N, Z_{k-1}^N) to (π_{k-1}, Z_{k-1}) , normalizing flow T_k , unnormalized annealed targets γ_{k-1} and γ_k and resampling threshold $A \in [1/N, 1)$.
 - 2: **Output:** Particles at iteration k : $\pi_k^N = (X_k^i, W_k^i)_{i=1}^N$, approximation Z_k^N to Z_k .
 - 3: Transport particles: $Y_k^i = T_k(X_{k-1}^i)$.
 - 4: Compute IS weights:
 $w_k^i \leftarrow W_{k-1}^i G_k(X_{k-1}^i)$ // unnormalized
 $W_k^i \leftarrow w_k^i / \sum_{j=1}^N w_k^j$ // normalized
 - 5: Estimate normalizing constant Z_k :
 $Z_k^N \leftarrow Z_{k-1}^N (\sum_{i=1}^N w_k^i)$.
 - 6: Compute effective sample size ESS_k^N .
 - 7: **if** $\text{ESS}_k^N \leq NA$ **then**
 - 8: Resample N particles denoted abusively also Y_k^i according to the weights W_k^i , then set $W_k^i = \frac{1}{N}$.
 - 9: **end if**
 - 10: Sample $X_k^i \sim \mathcal{K}_k(Y_k^i, \cdot)$. // MCMC
 - 11: Return (π_k^N, Z_k^N) .
-

Algorithm 2 CRAFT-training

- 1: **Input:** Initial NFs $\{T_k\}_{1:N}$, number of particles N , unnormalized annealed targets $\{\gamma_k\}_{k=0}^K$ with $\gamma_0 = \pi_0$ and $\gamma_K = \gamma$, resampling threshold $A \in [1/N, 1)$.
 - 2: **Output:** Learned flows T_k and length J sequence of approximations (π_K^N, Z_K^N) to (π_K, Z_K) .
 - 3: **for** $j = 1, \dots, J$ **do**
 - 4: Sample $X_0^i \sim \pi_0$ and set $W_0^i = \frac{1}{N}$ and $Z_0^N = 1$.
 - 5: **for** $k = 1, \dots, K$ **do**
 - 6: $\hat{h} \leftarrow \text{flow-grad}(T_k, \pi_{k-1}^N)$ using eqn (8).
 - 7: $(\pi_k^N, Z_k^N) \leftarrow \text{SMC-NF-step}(\pi_{k-1}^N, Z_{k-1}^N, T_k)$
 - 8: Update the flow T_k using gradient \hat{h} .
 - 9: **end for**
 - 10: Yield (π_K^N, Z_K^N) and continue for loop.
 - 11: **end for**
 - 12: Return learned flows $\{T_k\}_{k=1}^K$.
-

Algorithm 3 CRAFT-deployment

- 1: **Input:** Fixed/trained NFs $\{T_k\}_{k=1}^K$, number of particles N , unnormalized annealed targets $\{\gamma_k\}_{k=0}^K$ with $\gamma_0 = \pi_0$, resampling threshold $A \in [1/K, 1)$.
 - 2: **Output:** Approximations (π_K^N, Z_K^N) to (π, Z) .
 - 3: Sample $X_0^i \sim \pi_0$ and set $W_0^i = \frac{1}{N}$ and $Z_0^N = 1$.
 - 4: **for** $k = 1, \dots, K$ **do**
 - 5: $(\pi_k^N, Z_k^N) \leftarrow \text{SMC-NF-step}(\pi_{k-1}^N, Z_{k-1}^N, T_k)$
 - 6: **end for**
 - 7: Return (π_K^N, Z_K^N) .
-

2.2. Limitations of ELBO based training

The idea of learning parameters of a forward generating Markov process $\bar{\eta}$ to approximate a target distribution π_K and its normalizing constant Z_K has already been explored in the literature, in particular in the context of VAEs training (Salimans et al., 2015; Wu et al., 2020; Geffner and Domke, 2021; Thin et al., 2021; Zhang et al., 2021). In these references, the standard approach consisting of minimizing the KL divergence between the distribution $\bar{\eta}$ of the forward generating process and a suitable augmented target $\bar{\pi}$ admitting π_K as marginal at time K is adopted:

$$\text{KL}[\bar{\eta}||\bar{\pi}] = \log Z_K - \mathbb{E}_{\bar{\eta}}[\log w_K]. \quad (2)$$

However, minimizing this Kullback–Leibler divergence, equivalently maximizing the ELBO objective $\mathbb{E}_{\bar{\eta}}[\log w_K]$, requires differentiating through the forward sampler which can be challenging when MCMC kernels are included as we now discuss in more detail.

Discontinuity of the forward sampler. The most widely used MCMC transition kernels, such as Hamiltonian Monte Carlo (HMC) or Metropolis-adjusted Langevin algorithm (MALA), rely on an Metropolis accept/reject mechanism to ensure invariance. However, these mechanisms are discontinuous functions of the input random variables of the forward sampler. To avoid this issue and enable use of the reparameterization trick, a popular solution is to use approximate MCMC kernels without Metropolis correction which have continuous densities and then use an approximate reverse distribution (Salimans et al., 2015; Wu et al., 2020; Geffner and Domke, 2021; Thin et al., 2021; Zhang et al., 2021). The downside of this is the bias that accrues from not having the correction and the difficulty of approximating the reversal. This can mean many slow mixing proposals are required.

It is desirable to use in the forward generating process a combination of standard Metropolis-corrected MCMC samplers (e.g. HMC, MALA) and NFs as used by CRAFT, Stochastic Normalizing Flows (SNFs) (Wu et al., 2020) and AFT (Arbel et al., 2021). In this case, the proposition below proven in Appendix B and extending the results of (Thin et al., 2021) shows that the gradient of the ELBO in the flow parameters is the sum of two terms, one of which is a high variance score term.

Proposition 1. *Let U be the set of continuous r.v.s (used to sample proposals) of distribution independent of any parameter and A the set of discrete r.v.s (used for accept/reject steps) during the forward generating process. Let p denote the joint distribution of A, U . The gradient of the ELBO $\mathbb{E}_{\bar{\eta}}[\log w_K]$ in the flow parameters is:*

$$\mathbb{E}_p[\nabla \log(w_K \circ \Phi) + \log(w_K \circ \Phi)\nabla \log p], \quad (3)$$

where Φ is a differentiable re-parameterization of the trajectory $X_{0:K}$ in terms of U for any A , i.e. $X_{0:K} = \Phi(U, A)$. In particular, when the learned flows $(T_k)_{k=1}^K$ are optimal, i.e. $T_k^\# \pi_{k-1} = \pi_k$, it holds that:

$$\mathbb{E}_p[\nabla \log(w_K \circ \Phi)] = \mathbb{E}_p[\log(w_K \circ \Phi)\nabla \log p] = 0.$$

The first term on the r.h.s. of the expression (3) corresponds to a reparameterization trick term. The second term, however, requires computing a high variance score $\nabla \log p$. As pointed out by Thin et al. (2021), algorithms such as SNFs with Metropolis-adjusted MCMC kernels are omitting the second score term when optimizing the ELBO. While this term happens to vanish in the particular case where the flows are optimal, its contribution can, in general, be important especially when initializing the algorithm with sub-optimal flows.

Discontinuity of the resampling steps. The literature has mostly focused on scenarios where the forward generating process $\bar{\eta}$ is a simple Markov process. SMC forward generating processes as used in CRAFT could be exploited to obtain a tighter ELBO and have been used in the context of state-space models (Maddison et al., 2017; Le et al., 2018; Naesseth et al., 2018). However, the resampling steps of SMC correspond to sampling discrete distributions and lead to high variance gradient estimates. Omitting them can introduce significant bias (Corenflos et al., 2021). These difficulties are further exacerbated in our context since, as discussed earlier, the MCMC steps are not differentiable either.

Emergence of annealing requires mixing. In (Wu et al., 2020) the ELBO corresponding to equation (2) has the form of the standard VI with NFs objective, with additional weight terms entering into w_K due to the MCMC steps and annealing schedule. The additional log-weight corrections are

$$\Delta \log w_k^{\text{SNF MC}} = \log \gamma_k(x_k^{\text{SNF}}) - \log \gamma_k(\hat{x}_k^{\text{SNF}}), \quad (4)$$

where \hat{x}_k^{SNF} is the value of the sample after the MCMC at temperature k has been applied and x_k^{SNF} is the value of the sample before it (see Appendix A). This is the only way the annealed densities enter into the Stochastic Normalizing Flow ELBO. Since this additional term is zero in the limit where the Markov kernel does not mix, it follows that the value of the SNF ELBO reduces to the VI with normalizing flows objective in this case. Consequently, for the training objective to be significantly different from the one without MCMC and annealing, this method requires the MCMC kernels to mix sufficiently, which may not be the case in challenging examples.

2.3. CRAFT training

Here we describe a method that neatly sidesteps the issues described with the standard ELBO. We take the learning objective to have the following form:

$$H = \sum_{k=1}^K \text{KL}[T_k^\# \pi_{k-1} || \pi_k] \quad (5)$$

where the sum runs over each transition between temperatures. The minimum of the objective is zero and this is attained when each flow transports perfectly between successive temperatures. In general, breaking the annealing task down term-by-term means that we can reduce the mode seeking effect of the reverse KL divergence. Each KL term in the sum can be written as follows:

$$\text{KL}[T_k^\# \pi_{k-1} || \pi_k] = \mathbb{E}_{\pi_{k-1}} [D_k] + \log \left(\frac{Z_k}{Z_{k-1}} \right), \quad (6a)$$

$$D_k(x) := \log \frac{\gamma_{k-1}(x)}{\gamma_k(T_k(x))} + \log |\nabla_x T_k(x)|. \quad (6b)$$

The expectation over π_{k-1} on the RHS of the equation is approximated using a streaming SMC approximation of π_{k-1} which will always be based on the corresponding learnt sampler for the current optimization step:

$$\text{KL}[T_k^\# \pi_{k-1} || \pi_k] - \log \left(\frac{Z_k}{Z_{k-1}} \right) \approx \sum_i W_{k-1}^i D_k(X_{k-1}^i). \quad (7)$$

The terms inside the summation require that we can evaluate the output of the normalizing flow $T_k(X)$ and the log determinant of the Jacobian $\log |\nabla_x T_k(x)|$. Similarly the gradients with respect to flow parameters θ_k of flow T_k are approximated:

$$\frac{\partial H}{\partial \theta_k} \approx \sum_i W_{k-1}^i \frac{\partial D_k(X_{k-1}^i)}{\partial \theta_k}. \quad (8)$$

When the flows transport perfectly, the gradient $\frac{\partial H}{\partial \theta_k}$ will have true value zero and there will be no bias in this particle approximation to it. More generally the gradient estimate will be biased since it is based on a normalized importance estimator. We will analyze this aspect further in Section 3.2, but for now we note that the normal guarantees from the SMC literature imply the gradient is consistent in the number of particles and the variance and bias are of order $O(1/N)$ (Del Moral, 2013).

The proposed algorithm is described in Algorithm 2. Each training loop iteration of CRAFT looks like a iteration of the test time sampler in Algorithm 3 with gradient estimation and flow parameter updates added. Note that the parameter update is applied after the flow transport step for that temperature so the first effect it has will be on the particles that come through in the next step. This allows us

to view CRAFT as an optimization of the objective (5) and will allow stability analysis that we will detail later in the text. It also means that the normalizing constant estimators produced after each step are unbiased. Since all the constants produced in this way depend on the optimization sequence they are not independent. At test time when the parameters are fixed the independence is restored.

Relationship to AFT. We now describe the relationship between CRAFT and the AFT method of (Arbel et al., 2021). Relating them is made more challenging by the fact there are two variants of the AFT algorithm described in (Arbel et al., 2021). For clarity we include both in our supplement as Algorithms 5 and 7. We call Algorithm 5 *simple AFT*- it is easier to describe and analyse but is not actually used in practice by (Arbel et al., 2021) for reasons we will describe. Instead they use what we call *practical AFT* (Algorithm 7) a more complicated version that performs better in practice. Relative to CRAFT, both variants of AFT may be viewed as a one pass greedy optimization of the objective (5). The closest point of algorithmic comparison to CRAFT is simple AFT- we show the line difference to CRAFT in Algorithm 6. At each temperature step simple AFT optimizes the particle approximated loss to its numerical minimum, before then updating the particles using the new flow.

We now describe the *sample replenishment problem* in AFT. The learning of the flows is hampered by the fact that the sample complexity of estimation can be higher than the finite number of available particles. In simple AFT this manifests as over-fitting of the flows to the available particles and large biases in expectations. Regenerating new samples at each temperature separately would require computation quadratic in the number of temperatures, which is too slow. To mitigate the over-fitting in a manor consistent with the one pass paradigm of the paper, the authors added validation samples which were used for early stopping. Test samples were also added to remove finite sample residual bias. This significantly increased the complexity of the AFT method, changing simple AFT (Algorithm 5) to practical AFT (Algorithm 7). This gave substantial empirical improvements but the fundamental limitation of the sample replenishment problem remains, arising from the desire to use one temperature pass. Even when the particles are a perfect approximation to the desired distribution at each temperature, if the sample complexity of estimating the training gradients of the flows is high relative to the number of particles then the practical AFT method still struggles to train.

By having multiple independent passes instead of one, CRAFT is able to fully address the sample replenishment problem in a different way. Since in CRAFT the parameter update is applied after flow transport at each temperature the analysis of the algorithm is simpler than that required

for AFT, where sophisticated arguments must be invoked to ensure consistency of the particles. Later in Section 3.2 we will see that we can analyse CRAFT from the perspective of unbiased gradients. This interpretation does not apply to either variant of AFT.

From a practical perspective CRAFT performs better than AFT (Section 4.1). This is significant because (Arbel et al., 2021) already showed they could train samplers that performed well relative to strong baselines like SMC. CRAFT is easier to tune and more robust to hyperparameter misspecification than AFT. Conceptually it is much simpler, to the extent that it is easier to describe CRAFT as a standalone algorithm than in terms of the AFT algorithm on which it builds. CRAFT is closer to a traditional machine learning training paradigm with gradients applied to a single objective. This makes it easier to adopt and to scale through parallelism.

Recent work from (Zimmermann et al., 2021) investigates a framework that reduces to the AFT objective (5) for deterministic forward and backward transitions. Inspired by (Arbel et al., 2021) they investigate some toy examples with normalizing flows. There are good reasons to concentrate on the normalizing flow case as we have done. The reverse distribution is analytic and optimal for a flow but can be difficult to approximate otherwise (Thin et al., 2021). Further Zimmermann et al. (2021) do not use Markov transition kernels in practice, which we found to be essential for good performance.

2.4. Using CRAFT within Particle MCMC

Whilst the SMC-NF estimator of the normalizing constant is unbiased, in general estimates of the expectations w.r.t. the target are asymptotically consistent but exhibit a $O(1/N)$ bias. In situations requiring high accuracy expectations it is desirable to have a method returning asymptotically unbiased estimates. However, increasing N for the bias to be negligible will often not be feasible for challenging applications because of memory requirements. MCMC algorithms have the advantage of providing consistent estimate of these expectations as compute time increases without having to store an increasing number of samples. Particle MCMC methods (Andrieu et al., 2010) provide a way to bring such benefits to SMC samplers. In particular, the so-called Particle Independent Metropolis–Hastings is an MCMC sampler that uses an SMC sampler with N particles and provides consistent estimates of expectations for any N as the number of iterations increases; see Appendix D for details. Since a single pass of CRAFT with fixed parameters can be thought of as an SMC sampler with additional deterministic transformations it can be used here. Computational trade-offs for Particle MCMC are analyzed by Pitt et al. (2012).

3. Analysis of the training objective

3.1. Reformulating the CRAFT objective

The CRAFT objective (5) can be rewritten as a single KL divergence between certain product distributions:

$$H = \text{KL} \left[\prod_{k=1}^K T_k^\# \pi_{k-1} \parallel \prod_{k=1}^K \pi_k \right]. \quad (9)$$

We have adopted the notation that the product symbol \prod can be used to denote multi-variable product measures. The proof above effectively applies the multi-variable generalization of the result for the KL -divergence between two product measures, $\text{KL}[U \times V \parallel F \times G] = \text{KL}[U \parallel F] + \text{KL}[V \parallel G]$. Despite the non-standard form of the KL divergence, it is also possible to obtain a bound on the normalizing constant

$$\log Z_K \geq - \sum_{k=1}^K \mathbb{E}_{X \sim \pi_{k-1}} [D_k] \quad (10)$$

This bound can be obtained starting from $H \geq 0$, expanding each term using (6a) and noting that there is a telescoping cancellation between the intermediate normalizing constants. The bound is distinct from the one obtained if we take the logarithm of the unbiased estimate used in SMC samplers. In practice the bound is estimated using the particle estimate for each temperature and is effectively the training objective of CRAFT.

3.2. Re-thinking the KL criterion at the particle level

We discuss here a re-interpretation of the CRAFT training procedure in terms of KL-divergences between particle approximations and targets. An advantage of this interpretation is that it shows that we compute an unbiased gradient of a modified objective that is well-motivated even when the number of particles is moderate.

CRAFT training requires minimizing the objective (5). As we do not have access to π_{k-1} , CRAFT approximates the intractable gradient of this objective using equation (8). We will compactly refer to the particle approximation as a weighted sum of delta masses $\pi_{k-1}^N(dx) = \sum_{i=1}^N W_{k-1}^i \delta_{X_{k-1}^i}(dx)$. While the use of such an approach could be of concern when N is moderate or at initialization, the following proposition shows that such concerns are ill-founded and that this approximate SMC based gradient is also the *unbiased* gradient of a well-defined and intuitive objective.

Proposition 2. *Let $\hat{\pi}_{k-1}^N(dx) = \mathbb{E}[\pi_{k-1}^N(dx)]$ denote the expectation of the random SMC approximation π_k^N of π_{k-1} w.r.t. to the law of the SMC-NF algorithm. An unbiased*

gradient of the objective $\text{KL}[T_k^\# \hat{\pi}_{k-1}^N || \pi_k]$ is given by

$$\mathbb{E}_{X \sim \pi_{k-1}^N} [\nabla_{\theta_k} D_k(X)]. \quad (11)$$

The l.h.s of this KL is the pushforward of the ‘‘average’’ SMC approximation to π_{k-1} through the flow T_k .

To help understand this result and how it is derived, consider a simpler scenario where the random approximation π_{k-1}^N of π_{k-1} has been obtained by a batch parallel importance sampling method. That is we sample $X_{k-1}^i \stackrel{i.i.d.}{\sim} q_{k-1}$ for $i=1, \dots, N$, compute $w(X_{k-1}^i) = \pi_{k-1}(X_{k-1}^i) / q_{k-1}(X_{k-1}^i)$. We then define $W_{k-1}^i \propto w(X_{k-1}^i)$ such that $\sum_{i=1}^N W_{k-1}^i = 1$ and finally return $\pi_{k-1}^N(dx)$. If we average over the random particle locations $X_{k-1}^{1:N}$, we obtain

$$\hat{\pi}_{k-1}^N(dx) = \mathbb{E}_{X_{k-1}^i \stackrel{i.i.d.}{\sim} q_{k-1}} [\pi_{k-1}^N(dx)]. \quad (12)$$

Contrary to $\pi_{k-1}^N(dx)$ which is a discrete measure, the distribution $\hat{\pi}_{k-1}^N(dx)$ admits a density and its Kullback–Leibler divergence with $\mu_k = (T_k^{-1})^\# \pi_k$ is well-defined. It then follows directly from (12), diffeomorphism invariance, and iterated expectations that

$$\begin{aligned} \text{KL}[T_k^\# \hat{\pi}_{k-1}^N || \pi_k] &= \mathbb{E}_{X \sim \hat{\pi}_{k-1}^N} \left[\log \frac{\hat{\pi}_{k-1}^N(X)}{\mu_k(X)} \right] \\ &= \mathbb{E}_{X_{k-1}^i \stackrel{i.i.d.}{\sim} q_{k-1}} \left[\mathbb{E}_{X \sim \pi_{k-1}^N} \left[\log \frac{\hat{\pi}_{k-1}(X)}{\mu_k(X)} \right] \right]. \end{aligned} \quad (13)$$

A direct consequence of this identity is that an unbiased gradient of $\text{KL}[T_k^\# \hat{\pi}_{k-1}^N || \pi_k]$ is indeed given by (11). To prove Proposition 2, we extend this argument to the scenario where $\hat{\pi}_{k-1}^N(dx)$ has been obtained by using an SMC sampler where the particles are not independent because of resampling; see Appendix C.

Taking a step back and returning to the original objective we note that once we start considering the marginal particle distribution $\hat{\pi}_{k-1}^N$ instead of the target distribution π_{k-1} the flow parameters θ effect the KL divergences that follow them, since later particle distributions depend on them. This additional effect is not included in the CRAFT updates. Each flow takes whatever particle distribution it receives from the steps and is trained only to reduce the KL of the push-forward to the next target distribution π_k , without reference to what follows.

4. Experiments

In this section we empirically investigate the performance of CRAFT. First, in Section 4.1 we give a case study demonstrating the empirical benefit of CRAFT relative to AFT, then in Section 4.2 we show that CRAFT

outperforms Stochastic Normalizing flows in two challenging examples. We then show a compelling example use case for CRAFT as a learnt proposal for a particle MCMC sampler applied to lattice field theory. Code for the algorithms and examples can be found at https://github.com/deepmind/annealed_flow_transport. Further experiments and details are included in the Appendix.

4.1. Empirically comparing CRAFT and AFT

Our first point of comparison for CRAFT is with AFT (Arbel et al., 2021). We use what we call the practical AFT method in Section 2.3. To illustrate the benefit of solving the AFT sample replenishment problem we show the effect of varying the batch size in both algorithms. Both methods are given the same total budget of particles at train and test time. To highlight that the sample replenishment problem cannot be mitigated simply by using more MCMC we gave AFT 100 times more MCMC updates than CRAFT at train and test time. We use the 1024 dimensional log Gaussian Cox process (LGCP) example which is the most challenging from (Arbel et al., 2021). As in that paper, we used a diagonal affine flow.

Figure 1 shows the results. We see that CRAFT outperforms AFT for all number of particles considered. This can be understood in terms of our better solution to the sample replenishment problem as described in Section 2.3. Since many of the most challenging modern sampling problems also have high memory usage this is also of considerable practical significance.

4.2. Comparing CRAFT and SNFs

We now compare CRAFT training with that of the corresponding SNFs. For the Markov transition kernel both methods use full Metropolis corrected HMC. We use the same normalizing flow family for both methods. As such this constitutes a direct comparison of the two approaches. We ran timed training experiments for both methods computing the corresponding unbiased estimates for the normalizing constant as it progressed. Note that at test time these estimates could be obtained faster by running the learnt sampler. We consider two different target densities following (Arbel et al., 2021) namely the 30 dimensional variational autoencoder latent space and the LGCP example. For the VAE example we used an affine inverse autoregressive flow (Kingma et al., 2016). For the LGCP we again used a diagonal affine flow.

Figure 2 shows the results. In both cases we find that CRAFT converges to a better value than SNFs. We attribute the worse final value for SNFs to the issues with the training discussed in Section 2.2.

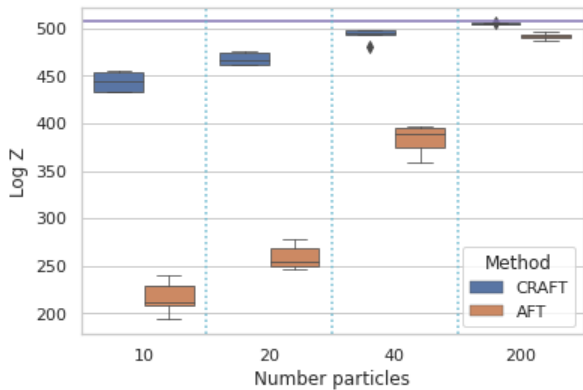


Figure 1. Comparison of normalizing constants between CRAFT and AFT for different numbers of particles. AFT is given 100 times more train/test MCMC steps than CRAFT. A gold standard value is shown in magenta.

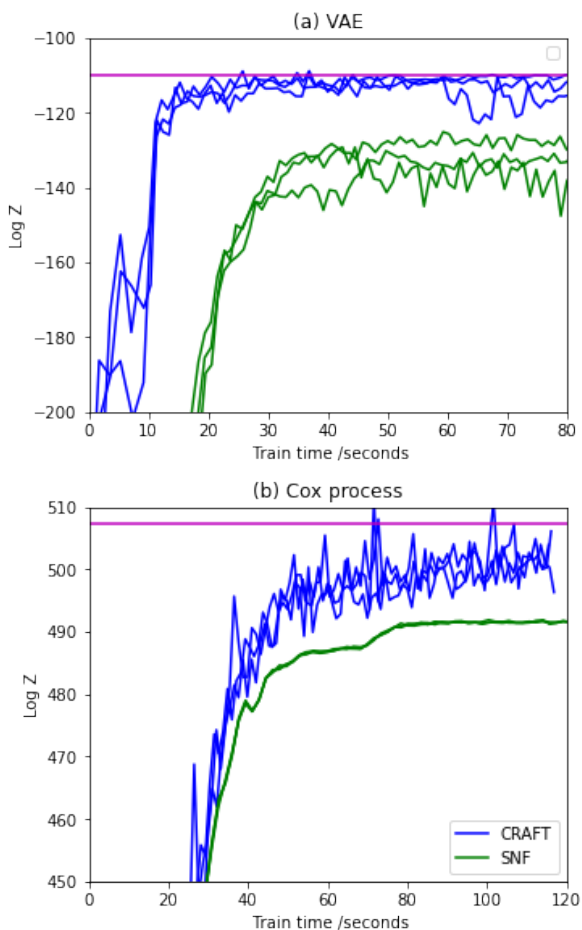


Figure 2. Timed comparison of normalizing constant estimates produced by CRAFT and SNF during training (higher is better). Three independent repeats are shown. A gold standard value is shown in magenta.

4.3. CRAFT based Particle MCMC for lattice ϕ^4 theory

A classic proving ground for sampling algorithms is in the area of lattice field theory. Indeed, it was this application that motivated Hamiltonian/Hybrid Monte Carlo (Duane et al., 1987). State of the art quantum chromodynamics (QCD) calculations (Borsanyi et al., 2021) use lattice sampling run on multiple supercomputers. There has been a recent surge of interest in using normalizing flows to improve sampling in such models starting with (Albergo et al., 2019), so it is natural to investigate the performance of CRAFT in this area.

Of the available lattice field theories we use the Euclidean ϕ^4 theory in two dimensions following (Albergo et al., 2019; 2021). We start from the continuum action:

$$S_{\text{cont}}[\phi] = \int \{ \|\nabla\phi(x)\|_2^2 + m^2\phi(x)^2 + \lambda\phi(x)^4 \} dx,$$

which we discretize on a 14×14 lattice, using lattice units, and periodic boundary conditions to obtain:

$$S_{\text{latt}}(\phi) = \sum_{\hat{x}} \{ \phi(\hat{x})\zeta(\hat{x}) + m^2\phi(\hat{x})^2 + \lambda\phi(\hat{x})^4 \}, \quad (14a)$$

$$\zeta(\hat{x}) := \sum_{\mu} [2\phi(\hat{x}) - \phi(\hat{x} + \hat{e}_{\mu}) - \phi(\hat{x} - \hat{e}_{\mu})]. \quad (14b)$$

Here the sum over \hat{x} runs over all lattice sites. The summation $\mu \in \{x, y\}$ runs over the dimensions of the lattice and \hat{e}_{μ} defines a single lattice step in the direction μ . $\lambda \geq 0$ defines the *coupling parameter* and m^2 is the *mass squared* which counter-intuitively may be positive or negative. A probability distribution over the values of the field at the lattice sites is then defined using $p(\phi) = \frac{1}{Z} \exp\{-S_{\text{latt}}(\phi)\}$. For positive λ and large ϕ the quartic term will dominate giving highly non-Gaussian tails. For negative m^2 , which will be our focus, the marginal distributions are bimodal.

As is typical of physical sciences applications, symmetry is crucial in this example. The target is invariant under any translation on the lattice. In Appendix H we show that the corresponding requirement on flows between successive temperatures is translation equivariance. We incorporate this in to the flow models under consideration using convolutional affine coupling layers (Dinh et al., 2017) with periodic boundary conditions.

Since high accuracy is required we are interested in methods that are asymptotically unbiased for general expectations. We take physically relevant expectations from (Albergo et al., 2019) namely the ‘two point susceptibility’ in the main text and the ‘Ising energy density’ in the supplement -both have similar results. We use Particle MCMC with three different independent proposal types. These are

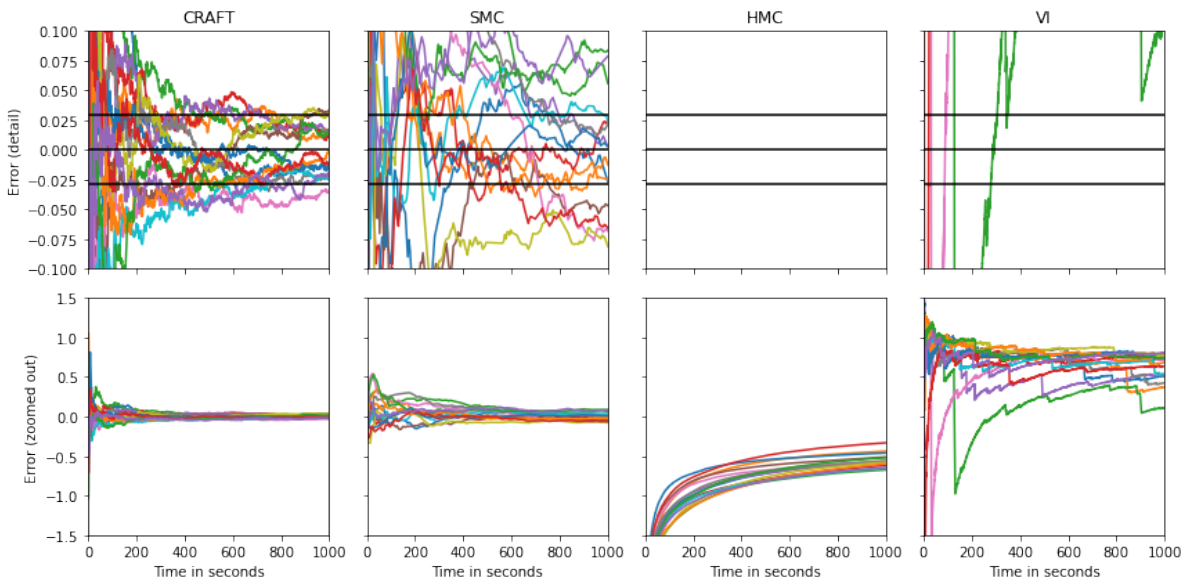


Figure 3. Timed comparison of MCMC methods for the ϕ^4 example, based on fifteen repeats. CRAFT, SMC and VI serve as proposal mechanisms for Particle MCMC. HMC is applied directly to the target. Error is in estimating two point susceptibility, an example, physically relevant expectation. Note HMC never reaches the detailed level of error in the top row.

SMC, VI with normalizing flows (and no annealing) and CRAFT. We also compare against HMC that directly targets the target density. The results are shown in Figure 3. We see that HMC struggles with this target density because it is unable to mix between modes. Despite producing samples quickly, the VI proposal fails to cover all of the modes because of the use of reverse KL divergence. The SMC proposal does much better. As an annealing method it is less susceptible to multi-modality. Finally CRAFT converges the fastest, improving on SMC without flows and avoiding the mode seeking behaviour of pure VI. Note that by scaling arguments if we accelerate an MCMC method by a factor of R we expect that errors in expectations will be reduced by a factor of \sqrt{R} on such error plots.

5. Conclusion

In this paper we have proposed and analyzed the CRAFT algorithm. At test time it is an SMC sampler with added flow transport steps. The training algorithm uses particle estimation to optimize a training objective that explicitly promotes transport between annealing temperatures. We have given a thorough analysis of the training including an insightful particle interpretation. We have shown that CRAFT compares well empirically and conceptually with existing methods to estimate flows in this context.

In terms of where such efforts fit into the bigger picture, there are many applications where high accuracy expecta-

tions are required and algorithms like Particle MCMC become necessary due to the challenging nature of the target. As such we believe the speed up we have achieved in the quantum field theory application is exciting both for that field and others.

6. Acknowledgements

The authors wish to thank Sébastien Racanière, Michalis Titsias, Phiala Shanahan, Jonas Köhler, Andy Ballard and the anonymous reviewers.

References

- Michael S. Albergo, Gurtej Kanwar, and Phiala E. Shanahan. Flow-based generative models for Markov chain Monte Carlo in lattice field theory. *Physical Review D*, 100:034515, 2019.
- Michael S Albergo, Denis Boyda, Daniel C Hackett, Gurtej Kanwar, Kyle Cranmer, Sébastien Racanière, Danilo Jimenez Rezende, and Phiala E Shanahan. Introduction to normalizing flows for lattice field theory. *arXiv preprint arXiv:2101.08176*, 2021.
- Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B*, 72(3): 269–342, 2010.

- Michael Arbel, Alex Matthews, and Arnaud Doucet. Annealed flow transport Monte Carlo. In *International Conference on Machine Learning*, 2021.
- Sz. Borsanyi, Z. Fodor, J. N. Guenther, C. Hoelbling, S. D. Katz, L. Lellouch, T. Lippert, K. Miura, L. Parato, K. K. Szabo, F. Stokes, B. C. Toth, Cs. Torok, and L. Varnhorst. Leading hadronic contribution to the muon magnetic moment from lattice QCD. *Nature*, 593(7857):51–55, 2021.
- James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018.
- Nicolas Chopin. A sequential particle filter method for static models. *Biometrika*, 89(3):539–552, 2002.
- Adrien Corenflos, James Thornton, George Deligiannidis, and Arnaud Doucet. Differentiable particle filtering via entropy-regularized optimal transport. In *International Conference on Machine Learning*, 2021.
- Rob Cornish, Anthony Caterini, George Deligiannidis, and Arnaud Doucet. Relaxing bijectivity constraints with continuously indexed normalising flows. In *International Conference on Machine Learning*, 2020.
- Chenguang Dai, Jeremy Heng, Pierre E Jacob, and Nick Whiteley. An invitation to sequential Monte Carlo samplers. *arXiv preprint arXiv:2007.11936*, 2020.
- Pim de Haan, Corrado Rainone, Miranda Cheng, and Roberto Bondesan. Scaling up machine learning for quantum field theory with equivariant continuous flows. *arXiv preprint arXiv:2110.02673*, 2021.
- Pierre Del Moral. *Mean Field Simulation for Monte Carlo Integration*. CRC press, 2013.
- Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B*, 68(3):411–436, 2006.
- Joshua V. Dillon, Ian Langmore, Dustin Tran, Eugene Brevdo, Srinivas Vasudevan, Dave Moore, Brian Patton, Alex Alemi, Matt Hoffman, and Rif A. Saurous. Tensor-Flow Distributions. *arXiv preprint arXiv:1711.10604*, 2017.
- Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *International Conference on Learning Representations*, 2017.
- Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, 1987.
- Tomas Geffner and Justin Domke. MCMC variational inference via uncorrected Hamiltonian annealing. In *Advances in Neural Information Processing Systems*, 2021.
- Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. In *International Conference on Machine Learning*, 2015.
- Jeremy Heng, Arnaud Doucet, and Yvo Pokern. Gibbs flow for approximate transport with applications to Bayesian computation. *Journal of the Royal Statistical Society Series B*, 83(1):156–187, 2021.
- Tom Hennigan, Trevor Cai, Tamara Norman, and Igor Babuschkin. Haiku: Sonnet for JAX, 2020.
- Matteo Hessel, David Budden, Fabio Viola, Mihaela Rosca, Eren Sezener, and Tom Hennigan. Optax: composable gradient transformation and optimisation, in *jax.*, 2020.
- Koji Hukushima and Yukito Iba. Population annealing and its application to a spin glass. In *AIP Conference Proceedings*, volume 690, pages 200–206. American Institute of Physics, 2003.
- Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.
- Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems*, 2016.
- Tuan Anh Le, Maximilian Igl, Tom Rainforth, Tom Jin, and Frank Wood. Auto-encoding sequential Monte Carlo. In *International Conference on Learning Representations*, 2018.
- Jun S Liu and Rong Chen. Blind deconvolution via sequential imputations. *Journal of the American Statistical Association*, 90(430):567–576, 1995.
- Chris J Maddison, John Lawson, George Tucker, Nicolas Heess, Mohammad Norouzi, Andriy Mnih, Arnaud Doucet, and Yee Teh. Filtering variational objectives. In *Advances in Neural Information Processing Systems*, 2017.
- Christian A Naesseth, Scott W Linderman, Rajesh Ranganath, and David M Blei. Variational sequential Monte Carlo. In *Artificial Intelligence and Statistics*, 2018.

- Radford M Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.
- George Papamakarios, Eric T Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal Machine Learning Research*, 22(57):1–64, 2021.
- Michael K Pitt, Ralph dos Santos Silva, Paolo Giordani, and Robert Kohn. On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *Journal of Econometrics*, 171(2):134–151, 2012.
- Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, 2015.
- Tim Salimans, Diederik Kingma, and Max Welling. Markov chain Monte Carlo and variational inference: Bridging the gap. In *International Conference on Machine Learning*, 2015.
- Saifuddin Syed, Alexandre Bouchard-Côté, George Deligiannidis, and Arnaud Doucet. Non-reversible parallel tempering: a scalable highly parallel MCMC scheme. *Journal of the Royal Statistical Society: Series B*, 84(2): 321–350, 2022.
- Achille Thin, Nikita Kotelevskii, Alain Durmus, Eric Moulines, Maxim Panov, and Arnaud Doucet. Monte Carlo variational auto-encoders. In *International Conference on Machine Learning*, 2021.
- Suriyanarayanan Vaikuntanathan and Christopher Jarzynski. Escorted free energy simulations. *The Journal of Chemical Physics*, 134(5):054107, 2011.
- Hao Wu, Jonas Köhler, and Frank Noé. Stochastic normalizing flows. In *Advances in Neural Information Processing Systems*, 2020.
- Guodong Zhang, Kyle Hsu, Jianing Li, Chelsea Finn, and Roger Grosse. Differentiable annealed importance sampling and the perils of gradient noise. In *Advances in Neural Information Processing Systems*, 2021.
- Heiko Zimmermann, Hao Wu, Babak Esmaeili, and Jan-Willem van de Meent. Nested variational inference. In *Advances in Neural Information Processing Systems*, 2021.

Continual Repeated Annealed Flow Transport Monte Carlo: Supplementary Material

A. Extended proposal and target distributions

Assume the transport maps T_l are fixed. We detail the extended proposal and target distributions used by Stochastic Normalizing Flows (Wu et al., 2020). These were introduced early on in (Vaikuntanathan and Jarzynski, 2011). The same target and proposal are also used by CRAFT and Annealed Flow Transport (Arbel et al., 2021) and the Gibbs flow algorithm (Heng et al., 2021) when resampling is omitted. We include here these distributions for sake of completeness.

We sample $X_0 \sim \pi_0(\cdot)$ at $k = 0$ then use $Y_k = T_k(X_{k-1})$ followed by $X_k \sim \mathcal{K}_k(Y_k, \cdot)$ at time $k \geq 1$. To simplify presentation, we do not use measure-theoretic notation. Hence, using the notation $M_l^{\text{trans}}(x, x') = \delta_{T_l(x)}(x')$ and $M_l^{\text{mut}}(x, x') = \mathcal{K}_l(x, x')$, we obtain the following proposal at time k after the transport step

$$\bar{\eta}_k(x_{0:k-1}, y_{1:k}) = \pi_0(x_0) \left(\prod_{l=1}^{k-1} M_l^{\text{trans}}(x_{l-1}, y_l) M_l^{\text{mut}}(y_l, x_l) \right) M_k^{\text{trans}}(x_{k-1}, y_k), \quad (15)$$

and the target is

$$\bar{\pi}_k(x_{0:k-1}, y_{1:k}) = \pi_k(y_k) L_{k-1}^{\text{trans}}(y_k, x_{k-1}) \left(\prod_{l=1}^{k-1} L_{l-1}^{\text{mut}}(x_l, y_l) L_{l-1}^{\text{trans}}(y_l, x_{l-1}) \right), \quad (16)$$

where $L_{l-1}^{\text{trans}}(x, x') = \delta_{T_{l-1}^{-1}(x)}(x')$ and $L_{l-1}^{\text{mul}}(x, x') = \pi_l(x') M_l^{\text{mul}}(x', x) / \pi_l(x)$. After the mutation step at time k , the proposal is

$$\bar{\eta}_k(x_{0:k}, y_{1:k}) = \bar{\eta}_k(x_{0:k-1}, y_{1:k}) M_k^{\text{mut}}(y_k, x_k) = \pi_0(x_0) \left(\prod_{l=1}^k M_l^{\text{trans}}(x_{l-1}, y_l) M_l^{\text{mut}}(y_l, x_l) \right) \quad (17)$$

and the target is

$$\bar{\pi}_k(x_{0:k}, y_{1:k}) = \pi_k(x_k) \left(\prod_{l=0}^{k-1} L_l^{\text{mut}}(x_{l+1}, y_{l+1}) L_l^{\text{trans}}(y_{l+1}, x_l) \right). \quad (18)$$

Hence the incremental weight after a transport term at time k is of the form

$$\frac{\bar{\pi}_k(x_{0:k-1}, y_{1:k})}{\bar{\eta}_k(x_{0:k-1}, y_{1:k})} = \frac{\bar{\pi}_{k-1}(x_{0:k-1}, y_{1:k-1})}{\bar{\eta}_{k-1}(x_{0:k-1}, y_{1:k-1})} \underbrace{\frac{\pi_k(y_k) L_{k-1}^{\text{trans}}(y_k, x_{k-1})}{\pi_{k-1}(y_{k-1}) M_k^{\text{trans}}(x_{k-1}, y_k)}}_{\text{incremental weight} = \frac{Z_{k-1}}{Z_k} G_{k, T_k}(x_{k-1})}, \quad (19)$$

while after the mutation step it is of the form

$$\frac{\bar{\pi}_k(x_{0:k}, y_{1:k})}{\bar{\eta}_k(x_{0:k}, y_{1:k})} = \frac{\bar{\pi}_k(x_{0:k-1}, y_{1:k})}{\bar{\eta}_k(x_{0:k-1}, y_{1:k})} \underbrace{\frac{\pi_k(x_k) L_{k-1}^{\text{mul}}(x_k, y_k)}{\pi_k(y_k) M_k^{\text{mul}}(y_k, x_k)}}_{\text{incremental weight} = 1}. \quad (20)$$

As shown in Appendix B.2 of (Arbel et al., 2021), we can also integrate out the random variables $Y_{1:k}$ in these expressions

as we can collapse the transport step and mutation step into one single Markov kernel

$$\begin{aligned} M_l^{\text{col}}(x_{l-1}, x_l) &= \int M_l^{\text{trans}}(x_{l-1}, y_l) M_l^{\text{mut}}(y_l, x_l) dy_l \\ &= \int \delta_{T_l(x_{l-1})}(y_l) \mathcal{K}_l(y_l, x_l) dy_l \\ &= \mathcal{K}_l(T_l(x_{l-1}), x_l). \end{aligned} \quad (21)$$

Similarly we can collapse the backward kernels that have been used to defined the extended target distributions $\bar{\pi}_k$

$$\begin{aligned} L_{l-1}^{\text{col}}(x_l, x_{l-1}) &= \int L_{l-1}^{\text{mut}}(x_l, y_l) L_{l-1}^{\text{trans}}(y_l, x_{l-1}) dy_l \\ &= \frac{\pi_l(T_l(x_{l-1})) |\nabla T_l(x_{l-1})| \mathcal{K}_l(T_l(x_{l-1}), x_l)}{\pi_l(x_l)}, \end{aligned} \quad (22)$$

so the collapsed proposal and target can be written as

$$\bar{\eta}_k(x_{0:k}) = \pi_0(x_0) \prod_{l=1}^k M_l^{\text{col}}(x_{l-1}, x_l), \quad (23)$$

$$\bar{\pi}_k(x_{0:k}) = \pi_k(x_k) \prod_{l=0}^{k-1} L_l^{\text{col}}(x_{l+1}, x_l). \quad (24)$$

For $k = K$, we write $\bar{\eta}_K = \bar{\eta}$ and $\bar{\pi}_K = \bar{\pi}$.

It follows that

$$\begin{aligned} \frac{Z_K \bar{\eta}(x_{0:K})}{\bar{\pi}(x_{0:K})} &:= w_K(x_{0:K-1}) = \frac{\gamma_K(x_K)}{\gamma_0(x_0)} \prod_{l=1}^K \frac{L_{l-1}^{\text{col}}(x_l, x_{l-1})}{M_l^{\text{col}}(x_{l-1}, x_l)} \\ &= \frac{\gamma_K(x_K) \gamma_{K-1}(x_{K-1}) \cdots \gamma_1(x_1)}{\gamma_0(x_0) \gamma_1(x_1) \cdots \gamma_{K-1}(x_{K-1})} \prod_{k=1}^K \frac{L_{k-1}^{\text{col}}(x_k, x_{k-1})}{M_k^{\text{col}}(x_{k-1}, x_k)} \\ &= \prod_{k=1}^K \frac{\gamma_k(x_k) L_{k-1}^{\text{col}}(x_k, x_{k-1})}{\gamma_{k-1}(x_{k-1}) M_k^{\text{col}}(x_{k-1}, x_k)} \\ &= \prod_{l=1}^K G_l(x_{l-1}), \end{aligned} \quad (25)$$

where, from (21) and (22), we obtain the so-called incremental weights

$$G_k(x_{k-1}) = \frac{\gamma_k(x_k) L_{k-1}^{\text{col}}(x_k, x_{k-1})}{\gamma_{k-1}(x_{k-1}) M_k^{\text{col}}(x_{k-1}, x_k)} = \frac{\gamma_k(T_k(x_{k-1})) |\nabla T_k(x_{k-1})|}{\gamma_{k-1}(x_{k-1})} \quad (27)$$

as given in (1b). In (Wu et al., 2020), the weight $w_K(x_{0:K-1})$ is instead obtained by first computing recursively the product appearing on the RHS of (25), thus they consider formally¹ in equation (12) of their paper a log-weight update at iteration k of the form

$$\begin{aligned} \log \left(\frac{L_{k-1}^{\text{col}}(x_k, x_{k-1})}{M_k^{\text{col}}(x_{k-1}, x_k)} \right) &= \log(\gamma_k(T_k(x_{k-1})) |\nabla T_k(x_{k-1})|) - \log \gamma_k(x_k) \\ &= \log \gamma_k(x_k^{\text{SNF}}) - \log \gamma_k(\hat{x}_k^{\text{SNF}}) \end{aligned} \quad (28)$$

where \hat{x}_k^{SNF} is the value of the sample after the MCMC at temperature k has been applied and x_k^{SNF} is the value of the sample before it as described in (4).

¹Note that while $G_k(x_{k-1})$ corresponds to the Radon-Nikodym derivative between the measures $\gamma_k(dx_k) L_{k-1}^{\text{col}}(x_k, dx_{k-1})$ and $\gamma_{k-1}(dx_{k-1}) M_k^{\text{col}}(x_{k-1}, dx_k)$, the ratio $L_{k-1}^{\text{col}}(x_k, dx_{k-1}) / M_k^{\text{col}}(x_{k-1}, dx_k)$ is not defined and taking its logarithm can only be interpreted as a shorthand notation for $\log G_k(x_{k-1}) - \log(\gamma_k(x_k) / \gamma_{k-1}(x_{k-1}))$.

B. Gradient expression for ELBO based approaches

Assume we consider the ELBO objective as in (Wu et al., 2020), that is we are interested in maximizing w.r.t. to the NFs parameters and the stochastic kernel parameters

$$\mathcal{L} = \mathbb{E}_{X_{0:K} \sim \bar{\eta}_K} \left[\log w_K(X_{0:K-1}) \right], \quad (29)$$

where $\bar{\eta}(x_{0:K}) = \bar{\eta}_K(x_{0:K})$ is defined in (23) and, as shown in (26), we have

$$w_K(x_{0:K-1}) = \frac{Z_K \bar{\pi}_K(x_{0:K})}{\bar{\eta}_K(x_{0:K})} = \prod_{l=1}^K G_l(x_{l-1}) \quad (30)$$

for $\bar{\pi}_K(x_{0:K})$ defined in (24). It is thus trivial to check that $w_K(X_{0:K-1})$ is an unbiased estimate of Z_K when $X_{0:K} \sim \bar{\eta}_K$.

Thin et al. (2021) provide a gradient estimate of the ELBO for annealed importance sampling, which corresponds to the case where $T_l(x) = x$ for all l , when the MCMC kernels \mathcal{K}_l are of the Metropolis–Hastings type and we use a reparameterized proposal

$$\mathcal{K}_l(x, x') = \int Q_l((x, u), x') g(u) du \quad (31)$$

for

$$Q_l((x, u), x') = \alpha_l(x, u) \delta_{S_l(x, u)}(x') + \{1 - \alpha_l(x, u)\} \delta_x(x'), \quad (32)$$

We show here how one can directly extend these results to our settings. We will use the convention $\alpha_l^1(x, u) = \alpha_l(x, u)$, $S_l^1(x, u) = S_l(x, u)$, $\alpha_l^0(x, u) = 1 - \alpha_l(x, u)$ and $S_l^0(x, u) = x$ so that we can rewrite

$$Q_l((x, u), x') = \sum_{a_l=0}^1 \alpha_l^{a_l}(x, u) \delta_{S_l^{a_l}(x, u)}(x'). \quad (33)$$

By combining (21), (23), (31) and (32), we can rewrite the distribution $\bar{\eta}_k(x_{0:K})$ as

$$\begin{aligned} \bar{\eta}_K(x_{0:K}) &= \pi_0(x_0) \prod_{l=1}^K \mathcal{K}_l(T_l(x_{l-1}), x_l) \\ &= \pi_0(x_0) \prod_{l=1}^K \int Q_l((T_l(x_{l-1}), u_l), x_l) g(u_l) du_l \\ &= \pi_0(x_0) \int \cdots \int \sum_{a_{1:K}} \prod_{l=1}^K \left(g(u_l) \alpha_l^{a_l}(T_l(\Phi_l(x_0, u_{1:l-1}, a_{1:l-1}), u_l)) \delta_{S_l^{a_l}(T_l(\Phi_l(x_0, u_{1:l-1}, a_{1:l-1}), u_l))}(x_l) \right) du_{1:K} \\ &= \pi_0(x_0) \int \cdots \int \sum_{a_{1:K}} g(u_{1:K}) \beta(a_{1:K} | x_0, u_{1:L}) \prod_{l=1}^K \delta_{S_l^{a_l}(T_l(\Phi_l(x_0, u_{1:l-1}, a_{1:l-1}), u_l))}(x_l) du_{1:K} \end{aligned} \quad (34)$$

where, for a given realization of $u_{1:L}$, we draw sequentially the Bernoulli r.v. A_l with

$$\mathbb{P}(A_l = a_l | x_0, u_{1:l-1}, a_{1:l-1}) = \alpha_l^{a_l}(T_l(x_{l-1}), u_l) \quad (35)$$

and, given $x_0, u_{1:l}, a_{1:l}$, the state x_l is given by

$$x_l := \Phi_l(x_0, u_{1:l}, a_{1:l}) = S_l^{a_l}(T_l(x_{l-1}), u_l) = S_l^{a_l}(T_l(\Phi_l(x_0, u_{1:l-1}, a_{1:l-1}), u_l)) \quad (36)$$

with $\Phi_0(x_0, u_{1:0}, a_{1:0}) := x_0$. Finally we used the notation $g(u_{1:K}) = \prod_{l=1}^K g(u_l)$ and the joint distribution of $A_{1:K}$ is denoted

$$\beta(a_{1:K} | x_0, u_{1:K}) = \prod_{l=1}^K \alpha_l^{a_l}(T_l(\Phi_l(x_0, u_{1:l-1}, a_{1:l-1}), u_l)). \quad (37)$$

So we can rewrite the ELBO as

$$\begin{aligned}\mathcal{L} &= \mathbb{E}_{\bar{\eta}_K(x_{0:K})} \left[\log w_K(x_{0:K-1}) \right] \\ &= \mathbb{E}_{\pi_0(x_0)g(u_{1:K})\beta(a_{1:K}|x_0, u_{1:K})} \left[\log w_K(x_{0:K-1}) \right],\end{aligned}\quad (38)$$

where we recall that the states $x_{0:K}$ are deterministic functions of $x_0, u_{1:K}, a_{1:K}$ given by (36). Now when taking the gradient of the ELBO w.r.t. parameters of the NFs and/or stochastic kernels then we have

$$\nabla \mathcal{L} = \mathbb{E}_{\pi_0(x_0)g(u_{1:K})\beta(a_{1:K}|x_0, u_{1:K})} \left[\nabla \log w_K(x_{0:K-1}) + \nabla \log \beta(a_{1:K}|x_0, u_{1:K}) \cdot \log w_K(x_{0:K-1}) \right]. \quad (39)$$

Now set $u_0=x_0, u=u_{0:K}, u=a_{1:K}$. Hence, the random variables (A, U) admit a joint distribution p of the form $p(a, u) := \pi_0(u_0)g(u_{1:K})\beta(a_{1:L}|x_0, u_{1:K})$. Moreover, define Φ the re-parametrization that maps (A, U) to the samples $X_{0:K}$, i.e. $[\Phi(A, U)]_l = \Phi_l(U_{0:l}, A_{1:l}) = X_l$. Hence, we can express the gradient in Equation (39) as follows:

$$\mathbb{E}_p [\nabla \log(w_K \circ \Phi) + \log(w_K \circ \Phi) \nabla \log p]. \quad (40)$$

The first term on the r.h.s. corresponds to the reparametrization trick while the second term is a score term, which in general can have high variance.

The gradient estimator of (Wu et al., 2020) effectively uses an unbiased estimator of the first reparameterization term but neglects the second term altogether.

In the particular case where the flows transport exactly between each temperature, i.e. $T_l^\# \pi_{l-1} = \pi_l$, the importance weight $w_K(x_{0:K-1})$ has the constant value Z_K . Hence, the second term in the expectation is a constant multiplied by a score function and consequently it vanishes:

$$\mathbb{E}_p [\log(w_K \circ g) \nabla \log p] \propto \mathbb{E}_p [\nabla \log p] = \int \nabla p = 0. \quad (41)$$

Moreover, since the ELBO is maximized when using optimal flows, the gradient $\nabla \mathcal{L}$ must vanish, which directly implies that the first term must also vanish. Therefore, despite the missing term in the gradient, the SNF learning rule has the correct fixed point in expectation.

C. Particle Interpretation of CRAFT

CRAFT is attempting to minimize w.r.t. θ the KL divergence $\mathcal{KL}(T_k^\#(\theta)\pi_k || \pi_{k+1}) = \mathcal{KL}(\pi_k || (T_k^{-1}(\theta))^\# \pi_{k+1})$. We have shown in Section 3.2 that we could re-interpret the criterion minimized by CRAFT at time k as the KL divergence between the expectation of the random measure approximating π_k and the pullback of π_{k+1} by $T_k(\theta)$. This was established in the case where the random measure is obtained using importance sampling. We show here that it is also valid for the case considered by CRAFT where it arises from an SMC-NF algorithm.

For sake of simplicity, we assume resampling is performed at any time step. Then in this case, the particles generated by the SMC-NF algorithm are distributed according to

$$\bar{q}_k(dx_{0:k}^{1:N}, dy_{1:k}^{1:N}) = \prod_{i=1}^N \pi_0(x_0^i) \prod_{l=1}^k \left[\prod_{i=1}^N \delta_{T_l(x_{l-1}^i)}(dy_l^i) \prod_{i=1}^N W_l^{\alpha_{l-1}^i} \mathcal{K}_l(y_l^{\alpha_{l-1}^i}, dx_l^i) \right], \quad (42)$$

where $\alpha_{l-1}^i \in \{1, \dots, N\}$ is the ancestral index of particle x_l^i . Using random particles arising from the SMC-NF algorithm, the distribution π_k is then approximated by the random empirical measure

$$\pi_k^N(dx) = \frac{1}{N} \sum_{i=1}^N \delta_{X_k^i}(dx) \quad (43)$$

and its expectation w.r.t. (42) is denoted

$$\hat{\pi}_k^N(dx) = \mathbb{E}[\pi_k^N(dx)]. \quad (44)$$

In this case, we have

$$\begin{aligned}
 \mathcal{KL}(\hat{\pi}_k^N || T_k^{-1}(\theta) \# \pi_{k+1}) &= \mathbb{E}_{X \sim \hat{\pi}_k^N} \left[\log \left(\frac{\hat{\pi}_k^N(X)}{(T_k^{-1}(\theta) \# \pi_{k+1})(X)} \right) \right] \\
 &= \mathbb{E}_{(X_{0:k}^{1:N}, Y_{1:k}^{1:N}) \sim \bar{q}_k, X \sim \pi_k^N} \left[\log \left(\frac{\hat{\pi}_k^N(X)}{(T_k^{-1}(\theta) \# \pi_{k+1})(X)} \right) \right] \\
 &= \mathbb{E}_{(X_{0:k}^{1:N}, Y_{1:k}^{1:N}) \sim \bar{q}_k} \left[\mathbb{E}_{X \sim \pi_k^N} \left[\log \left(\frac{\hat{\pi}_k(X)}{(T_k^{-1}(\theta) \# \pi_{k+1})(X)} \right) \right] \right]. \tag{45}
 \end{aligned}$$

So an unbiased gradient estimate w.r.t. θ of $-\mathcal{KL}(\hat{\pi}_k || T_k^{-1}(\theta) \# \pi_{k+1})$ is given by

$$-\mathbb{E}_{X \sim \pi_k^N} \left[\nabla_{\theta} \log \left(T_k^{-1}(\theta) \# \pi_{k+1} \right) \right], \tag{46}$$

which is the gradient used by CRAFT to update the parameter of $T_k(\theta)$.

D. Particle MCMC

We describe here the particle independent Metropolis–Hastings algorithm (PIMH) introduced in (Andrieu et al., 2010). This is an MCMC algorithm using as an independent proposal an importance sampling or SMC-type algorithm such as an SMC combined with NFs learned by CRAFT (see Algorithm 3). It generates a Markov chain $((\pi_K^N(j), Z_K^N(j)))_{j \geq 1}$ where $\pi_K^N(j)(dx) = \sum_{i=1}^N W_K^i(j) \delta_{X_K^i(j)}(dx)$ such that for any test function f

$$\frac{1}{J} \sum_{j=1}^J \mathbb{E}_{X \sim \pi_K^N(j)} [f(X)] \rightarrow \mathbb{E}_{X \sim \pi_K} [f(X)], \quad \text{where} \quad \mathbb{E}_{X \sim \pi_K^N(j)} [f(X)] := \sum_{i=1}^N W_K^i(j) f(X_K^i(j)), \tag{47}$$

as the number J of MCMC iterations goes to infinity. This is described in detail in Algorithm 4.

Algorithm 4 Particle Independent Metropolis–Hastings step $P((\pi_K^N(j), Z_K^N(j)), (\cdot, \cdot))$

- 1: **Input:** Current approximations $(\pi_K^N(j), Z_K^N(j))$ to (π_K, Z_K) .
 - 2: **Output:** New approximations $(\pi_K^N(j+1), Z_K^N(j+1))$ to (π_K, Z_K) .
 - 3: Run an SMC algorithm to obtain $(\pi_K^{*,N}, Z_K^{*,N})$ approximating (π_K, Z_K) (such as Algorithm 3).
 - 4: Set $(\pi_K^N(j+1), Z_K^N(j+1)) = (\pi_K^{*,N}, Z_K^{*,N})$ with probability $\alpha(Z_K^N(j), Z_K^{*,N}) := \min \left\{ 1, \frac{Z_K^{*,N}}{Z_K^N(j)} \right\}$.
 - 5: Otherwise set $(\pi_K^N(j+1), Z_K^N(j+1)) = (\pi_K^N(j), Z_K^N(j))$.
-

For fixed computational efforts, one could either run many MCMC iterations with N small or few MCMC iterations with N large. Under regularity conditions, it was shown in (Pitt et al., 2012) that N should be selected such that the variance of the estimate of the log-normalizing constant is close to 1 in order to minimize the asymptotic variance of the estimator (47).

E. Learning model parameters using CRAFT

We do not consider in the paper examples where π_K depends on some parameters ϕ that one wants to learn. We explain here how this could be addressed.

Consider the following target

$$\pi_K^{\phi}(x) = \frac{\gamma_K^{\phi}(x)}{Z_K^{\phi}}. \tag{48}$$

To estimate ϕ , we propose to learn ϕ by maximizing the log-normalizing constant/evidence

$$\arg \min_{\phi} \log Z_K^{\phi}. \tag{49}$$

The gradient of this objective w.r.t. ϕ can be approximated as follows

$$\nabla \log Z_K^\phi = \mathbb{E}_{X \sim \pi_K^\phi} [\nabla \log \gamma_K^\phi(X)] \approx \mathbb{E}_{X \sim \pi_K^{N,\phi}} [\nabla \log \gamma_K^\phi(X)], \quad (50)$$

where $\pi_K^{N,\phi}$ is a particle approximation of π_K^ϕ . This approximation can be obtained using a learned SMC/normalizing flow sampler obtained using CRAFT.

F. Further comparison of CRAFT and AFT

In this section we investigate the effect of reducing the number of MCMC iterations in CRAFT and AFT. As a benchmark we use the log Gaussian Cox process- with a large lattice discretization of $40 \times 40 = 1600$. To make the example harder we reduce the number of HMC iterations per temperature from 10 to 2. The rest of the experimental setup is kept the same. The goal here is to learn a fast flow augmented sampler at test time with reasonable training time. Since the diagonal affine flow used constitutes a negligible overhead the change corresponds to a 5 times reduction in test compute time relative to the previous work.

Figure 4 shows the results. We see that for 10 and 30 transitions that AFT struggles to learn a good flow. This is because the expectation of the objective required in the greedy training is not well estimated with fewer MCMC iterations. CRAFT has the same difficulty at the beginning of optimization but due to the repeated nature of the optimization it is able to recover.

Having seen the solution found by CRAFT one might reasonably ask if we could use different numbers of HMC iterations at train and test time for AFT- though this is not suggested in the original work of (Arbel et al., 2021). After all, we know that with 10 iterations AFT performs well on the Cox process task, which suggests we could use that learnt flow with fewer HMC samples at test time. This is indeed the case. Training with extra MCMC iterations restores the behaviour of AFT to be similar to that of CRAFT in this case. However this is at the expense of introducing even more hyperparameters to AFT that require tuning. Relative to CRAFT, a user of AFT has to correctly tune the behaviour of three sets of particles (train/validation/test) and manage multiple separate optimization loops (one for each temperature). This extra modification of allowing different number of MCMC iterations is then a further complexity.

Note additionally that AFT is subject to the degradation in performance with batch size relative to CRAFT described in Section 4.1- we regard this case as the most direct demonstration of how CRAFT solves the sample replenishment problem in AFT.

G. Distributed and parallel computation

In this section we discuss the particular challenges we envisage scaling the algorithm to large scale compute infrastructure and potential remedies.

We have described the algorithm as a sequential algorithm but there is considerable scope for running it asynchronously in parallel. Each temperature can be considered to correspond to a node in a chain graph. Along the edges of the graph messages are passed. In the most direct implementation this corresponds to the particles and the weights. There is no need to run for a full pass along the graph before starting the next one. In some compute clusters communication locality is also important so it is helpful that the message passing is sparse and local.

An alternative to passing the particles and weights is to cycle the temperature at each node and pass gradient updates instead. The temperatures at each node can be offset by one so that the gradients are always received at the right time. Such a prescription is inspired by the distributed Algorithm 5 of (Syed et al., 2022). For CRAFT, this method will reduce communication overhead between nodes in the case where the gradient updates require less memory than passing particles and weights. Although the gradient updates for neural networks can in general be large, incorporating symmetries as in Sections H and 4.3 or more generally adding domain knowledge can reduce network sizes substantially.

Another consideration is the case where only a few particles will fit onto a worker, so that it is desirable to parallelize a single iteration across multiple workers. The resampling step means that the algorithm is not batch parallel. Therefore resampling could introduce a significant communication overhead between the parallel workers. One possible remedy is only to resample within a worker.

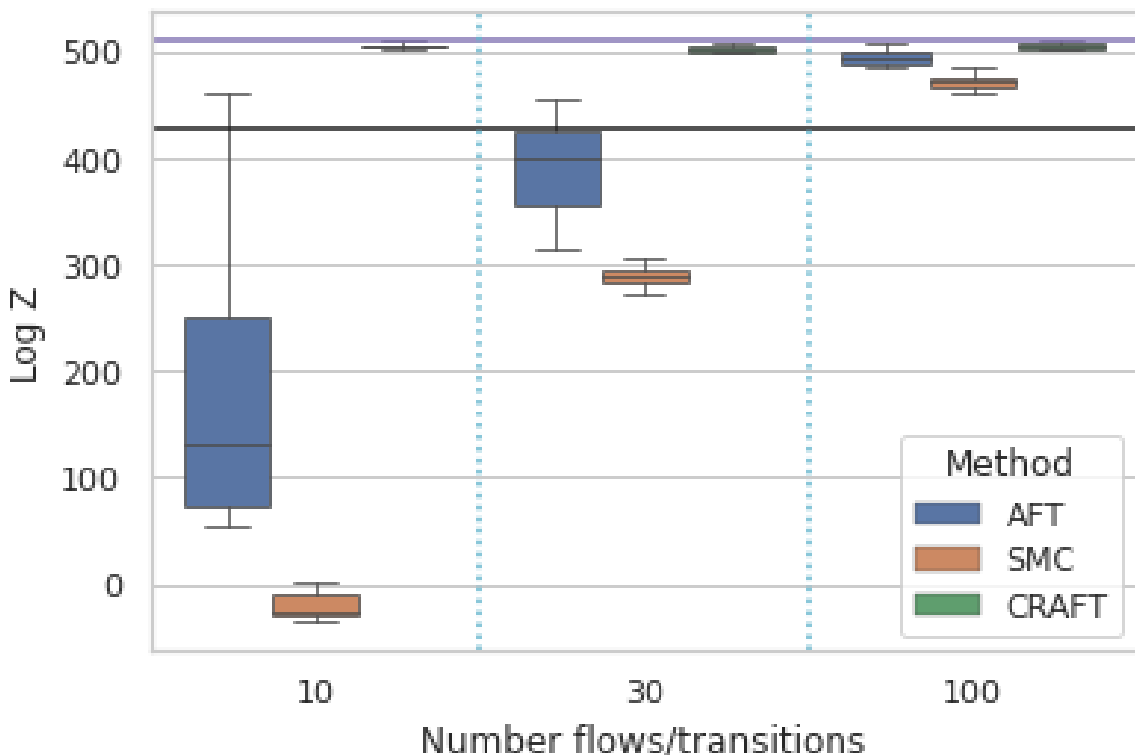


Figure 4. Comparison of CRAFT, AFT, SMC and VI on the pines task from (Arbel et al., 2021) with a reduced number of HMC iterations.

H. Incorporating symmetries

In applications to high dimensional target densities the target density often exhibits symmetries. Incorporating such symmetries in normalizing flows can give significant improvements in the performance and efficiency of normalizing flows.

Consider a case where we have a group \mathcal{G} with elements g and corresponding group representation members D_g . In this case we have a symmetric target density π_T which is symmetric under the target density.

$$\pi_T(D_g x) = \pi_T(x) \quad \forall g \in \mathcal{G} \quad (51)$$

We also take a base density π_B that is also symmetric under the group. We choose a normalizing flow that is equivariant under the action of the group so that:

$$T(D_g x) = D_g T(x) \quad \forall g \in \mathcal{G} \quad (52)$$

When we push forward the base density we end up with a distribution $T_{\#}\pi_B$ that is invariant under the action of the group.

$$T_{\#}\pi_B(D_g x) = T_{\#}\pi_B(x) \quad \forall g \in \mathcal{G} \quad (53)$$

In standard normalizing flow modelling the symmetric pushforward distribution is then tuned to match the symmetric target distribution using reverse KL-divergence $\text{KL}[T_{\#}\pi_B || \pi_T]$.

It is important to extend such a treatment of symmetries to the case of CRAFT. To this end we analyze the symmetry properties of a sequence of densities with a geometric annealing schedule. We consider an initial normalized density π_0

and unnormalized final density γ_1 . The unnormalized density of the geometric annealing schedule takes the form:

$$\log \gamma_\beta(x) = \log(1 - \beta) \log \pi_0(x) + \beta \log \gamma_1(x) \quad (54)$$

where the schedule time $\beta \in [0, 1]$. Clearly if both the initial density and final density are symmetric under the action of the group then so is $\gamma_\beta(x)$ and its normalized version π_β .

In CRAFT we use a sum of KL divergences between the pushforward of an initial annealed density and the next target density $\text{KL}[T\#\pi_\beta||\pi_{\beta'}]$. In terms of symmetry there is a direct analogy to the case of standard variational inference with normalizing flows with which we started the discussion. In particular π_β now takes the role of the symmetric base density and $\pi_{\beta'}$ takes the role of the target. The symmetry requirement on the normalizing flow T is thus the same. It needs to be equivariant under the action of the group so that the push forward $T\#\pi_\beta$ is symmetric in a manner that matches its target.

I. Additional experiment details

All experiments used a geometric (log-linear) annealing schedule. The initial distribution was always a standard multivariate normal. All experiments used HMC as the Markov kernel, which was tuned to get a reasonable acceptance rate based on preliminary runs of SMC. Normalizing flows were always initialized to the identity flow. Wherever a stochastic gradient optimizer was required we used the Adam optimizer (Kingma and Ba, 2015).

In terms of software dependencies for our code we used Python, JAX (Bradbury et al., 2018), Optax (Hessel et al., 2020), Haiku (Hennigan et al., 2020), and the TensorFlow probability JAX ‘substrate’ (Dillon et al., 2017).

All experiments were carried out using a single Nvidia v100 GPU. We used sufficient CPUs and CPU RAM such that these were not the bottleneck.

I.1. Comparing CRAFT and AFT batch size performance

Here we give more details of the experiment described in Section 4.1. We used the original software for AFT which is available at https://github.com/deepmind/annealed_flow_transport. We used practical AFT (Algorithm 7), which is the version using early stopping on a set of validation particles and a ‘hold out’ test set.

We used the same MCMC method for both approaches, which was pre-tuned full Metropolis adjusted Hamiltonian Monte Carlo (HMC) with 10 leapfrog steps per iteration. For CRAFT we used 10 HMC steps per temperature and for AFT we used 1000 HMC steps per temperature meaning that AFT had 100 times more HMC steps per temperature. The HMC step sizes were linearly interpolated with 0 corresponding to the initial distribution and 1 corresponding to the final distribution, the interpolation points were $[0, 0.25, 0.5, 1]$ and the corresponding step sizes were $[0.3, 0.3, 0.2, 0.2]$.

At test time AFT and CRAFT used the same number of particles for each experiment. To make it fair at train time, the total CRAFT particle budget was divided in to two halves for AFT, one half was used for the training particles and the other half was used for the validation particles. For each experiment the total number of train particles for either method was the same as the number used at test time and this is the number shown in Figure 1.

In terms of optimization both methods were well converged. AFT used an optimization step size of $1e-2$, and 500 optimization iterations per temperature. CRAFT used 200 total optimization iterations, and a step size of $5e-2$ which was reduced down to $1e-2$ after 100 iterations.

I.2. Comparing CRAFT and Stochastic Normalizing Flows

To the best of knowledge, the HMC based Stochastic Normalizing Flow described in (Wu et al., 2020) was not implemented or experimented with in that work. We implemented it and found in preliminary experiments that it outperformed the random walk Metropolis Hastings used in the original work. It was also the most commensurate with our own HMC usage. Therefore we used the HMC SNF for the comparisons in this work.

In the spirit of the original SNF paper, we performed preliminary experiments with a SNF ELBO that incorporated resampling in the forward computation. For the reverse computation we neglected the resampling contribution just as is done for the score term arising from the discrete Metropolis–Hastings correction. In other words we proceeded as if the forward computation was compatible with the reparameterization gradient although this does not in fact correspond to an unbiased

estimate of the gradient. These preliminary experiments indicated that this variant of SNF became unstable and challenging to train. The fact that CRAFT can cope with resampling is a benefit of the approach, and is expected from the form of the CRAFT training objective.

We observed that variational learning of MCMC step sizes is unstable in SNF, so we followed (Wu et al., 2020) who use a fixed step size during variational training for the majority of their experiments. Instead of a learnt step size, we tuned the step sizes separately in advance so that the corresponding MCMC would have a reasonable acceptance rate just as we do for CRAFT. Consequently we used the same Markov kernels for both CRAFT and SNFs.

For the SNF software we implemented them in JAX using similar modules to that of the CRAFT implementation. We verified this JAX implementation against the original publicly available SNF implementation at https://github.com/noegroup/stochastic_normalizing_flows. A benefit of using the same basic modules as the CRAFT implementation is that the timed comparison is much less subject to unwanted differences arising from the different software libraries used. We confirmed using a Python execution profiler that both the CRAFT and SNF code was spending time on the expected core algorithmic computations and not on unexpected code paths. We observed some variability in the CRAFT and SNF step times.

The JAX SNF implementation exploits the connection between SNFs and Annealed Importance Sampling with normalizing flows. This equivalence can be easily seen from the form of the overall unnormalized weights in each case (Appendix A). The gradient dynamics are unchanged by using this representation of the forward computation.

Relative to CRAFT we found that SNFs used significantly more memory in our experiments. This is due to the backward pass of the SNFs where gradients are passed through the whole forward computation whereas the gradients are local in CRAFT. To the benefit of SNFs, we reduced the batch size of both CRAFT and SNFs in the relevant experiments so that the SNF training would still fit on GPU and that the batch sizes would be comparable for the two algorithms.

The normalizing flow used for the VAE experiment was an Affine Inverse Autoregressive Flow. Similar to (Arbel et al., 2021) we use a masked network (Germain et al., 2015) to achieve the autoregressive dependency structure. The masked network had two hidden layers and unmasked would have 150 hidden units. The final weights and biases were initialized to give an identity flow. We used a leaky ReLU nonlinearity. Since the flows have high capacity in this case we only needed to use 3 temperatures in both cases. Since the cost of these inverse autoregressive flows is quadratic in the dimension they are prohibitively expensive for systems of higher dimensionality. We used 2 HMC steps per transition with 10 leapfrog steps in both cases. Both CRAFT and SNF had a batch size of 100.

For the LGCP SNF/CRAFT comparison we used 10 transitions in both cases. We used 1 HMC step per transition with 10 leapfrog steps in both cases. Both CRAFT and SNF had a batch size of 2000.

I.3. CRAFT based Particle MCMC for lattice ϕ^4 theory

Those readers who are less familiar with the background physics can gain intuition from the function form of equation (14). The terms over neighbouring lattice sites promote spatial correlation. The other terms all involve only a single lattice site.

The expectations we use to evaluate the algorithms are physically motivated observables described by (Albergo et al., 2019). In the main text we use the two point susceptibility. In Figure 5 we show that similar results show for the Ising energy density. The chains that were averaged to produce the expectations in 3 and 5 are shown in 6 and 7 respectively. The errors shown in the average value plots are computed by subtracting off the gold standard value.

We note that de Haan et al. (2021) incorporate additional symmetries of the target on top of the lattice translation symmetry we incorporate. This is at the expense of introducing numerical integration into the variational inference with normalizing flows. The convolutional affine coupling layers we use do not require such numerical integration, and are known to significantly outperform fully connected affine coupling layers in this context (Albergo et al., 2021).

The parameters we consider for the theory are based on those from (Albergo et al., 2019). In particular we take the hardest parameter set E5 and make it more difficult by reducing m^2 from -4 to -4.75 . As noted by (Albergo et al., 2019), the reason this makes the problem more challenging is it increases the gap between the positive and negative modes. The resulting parameters we use are therefore the following: Lattice size 14×14 ; $\lambda = 5.1$ and $m^2 = -4.75$.

In these ϕ^4 experiments, upon preliminary investigation of computation time using a Python execution profiler, we found

Continual Repeated Annealed Flow Transport Monte Carlo

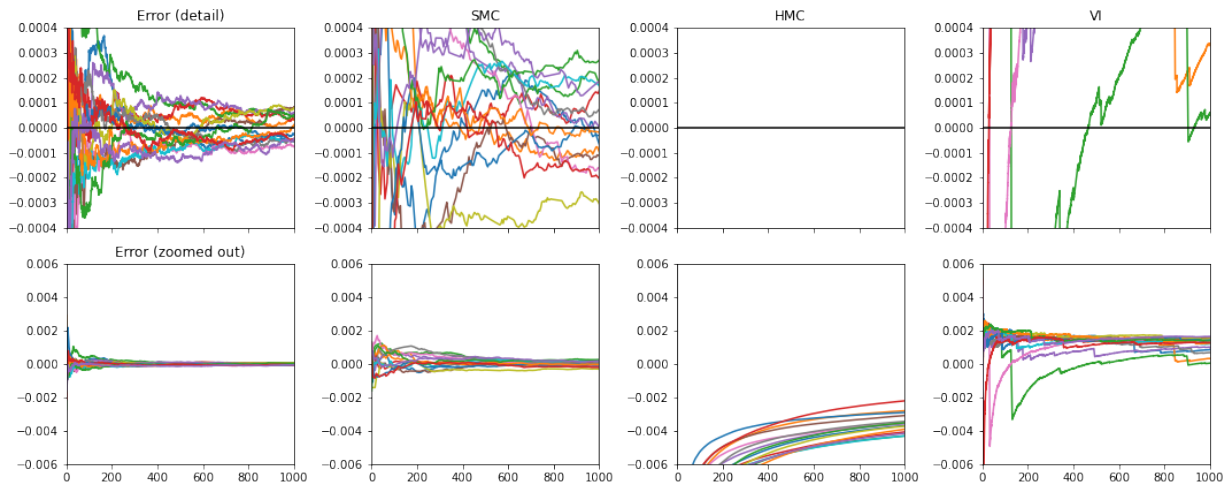


Figure 5. Timed comparison of MCMC methods for the ϕ^4 example, based on fifteen repeats. CRAFT, SMC and VI serve as proposal mechanisms for particle MCMC. HMC is applied directly to the target. Note HMC never reaches the detailed level of error in the top row.

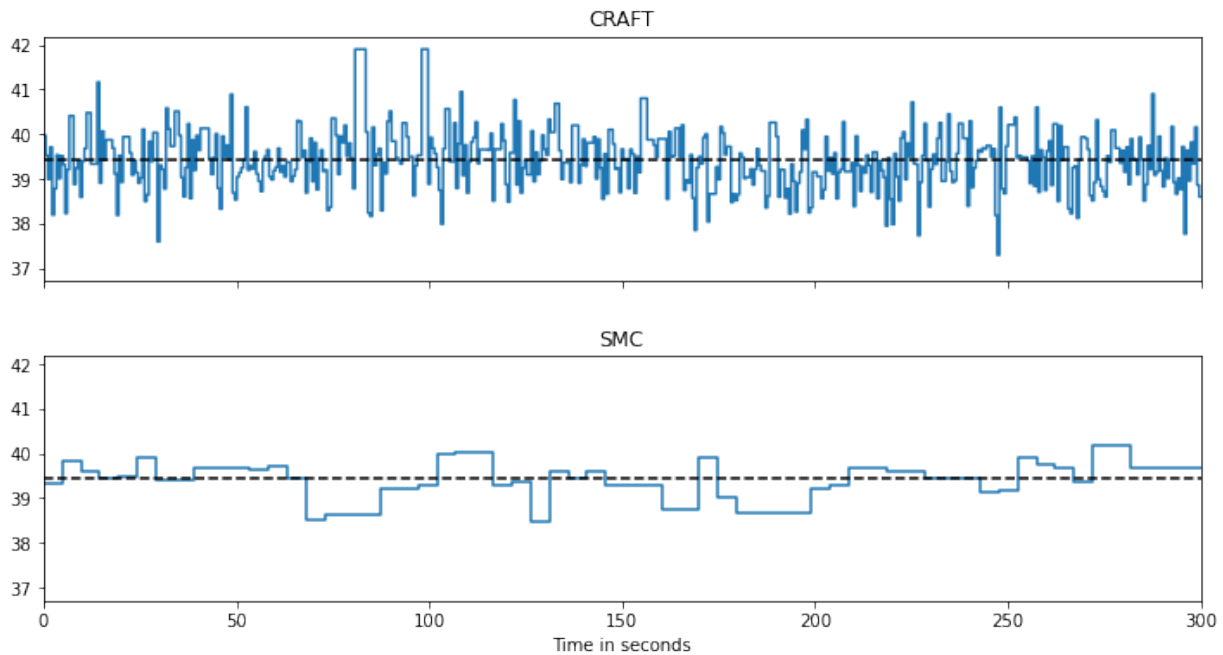


Figure 6. Five minutes of the raw chain used to compute the expectation averages for the SMC and CRAFT two point susceptibility. The estimated gold standard value is shown as the black dashed line. SMC and CRAFT are displayed because they are the two best algorithms. VI and HMC have much larger errors.

that time was sometimes being spent in parts of the code that might not be expected from algorithmic considerations alone. In particular for fast samplers like VI, we found that storing results and computing expectations were an (unoptimized) bottleneck. To make timed comparisons fairer, we therefore performed separate timing estimation on code that ran only the core sampling code and not on computing expectations and storing results. These adjusted timing estimates are what is

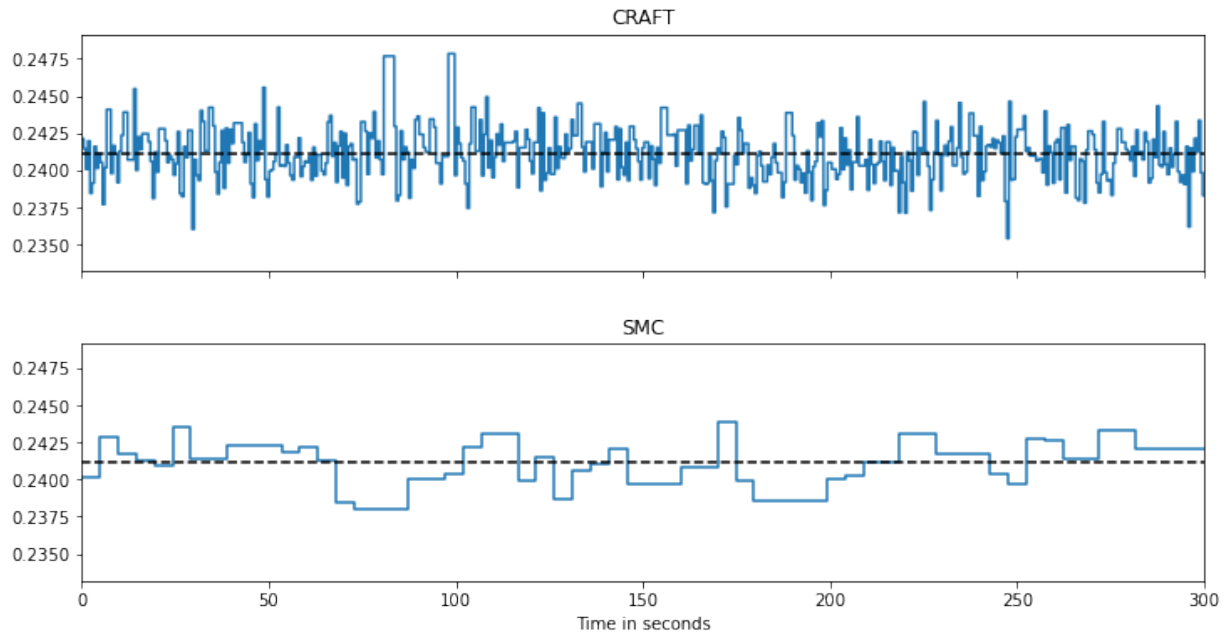


Figure 7. Five minutes of the raw chain used to compute the expectation averages for the SMC and CRAFT Ising energy density. The estimated gold standard value is shown as the black dashed line. SMC and CRAFT are displayed because they are the two best algorithms. VI and HMC have much larger errors.

reported in our results.

The gold standard values of the error for each reported expectation are computed using SMC. We verified in cases of simple known expectations that SMC and CRAFT gave the correct answers whereas HMC and VI gave large errors, just as for the physically motivated observables reported.

The normalizing flow used a convolutional affine coupling layer (Dinh et al., 2017) with periodic boundary conditions. We used alternating checkerboard masking patterns. The convolution kernels were of size 3×3 . The conditioner of each flow had one hidden convolutional layer with 10 channels. The final weights and the biases of the convolution were initialized to give an identity flow. We used a ReLU nonlinearity.

SMC used 90 transitions. Since the masking pattern required two coupling layers to change each site we used two coupling layers per transition in CRAFT. CRAFT used 10 transitions. VI, which has no annealing, used 20 coupling layers. We confirmed there was no benefit to increasing the number of VI coupling layers. We used 10 HMC steps per annealing transition with 10 leapfrog steps for SMC and CRAFT. Direct HMC used 10 leapfrog iterations per step.

J. CRAFT to AFT algorithm comparison.

In this section we describe the two different variants of AFT from (Arbel et al., 2021) for clarity and completeness. We also describe the line difference between CRAFT and simple AFT. Note that the full source code for each method is also available at https://github.com/deepmind/annealed_flow_transport. Based on equation (7) we define:

$$\mathcal{L}_k^N(T) := \text{KL}[T_k^\# \pi_{k-1} || \pi_k] - \log \left(\frac{Z_k}{Z_{k-1}} \right) \approx \sum_i W_{k-1}^i D_k(X_{k-1}^i) \quad (55)$$

which from equation (8) has approximate gradient given by

$$\sum_i W_{k-1}^i \frac{\partial D_k(X_{k-1}^i)}{\partial \theta_k}. \quad (56)$$

Algorithm 5 Simple AFT: from Arbel, Matthews and Doucet 2021. This corresponds to Algorithm 1 of the prior paper.

- 1: **Input:** number of particles N , unnormalized annealed targets $\{\gamma_k\}_{k=0}^K$ such that $\gamma_0 = \pi_0$ and $\gamma_K = \gamma$, resampling threshold $A \in [1/N, 1)$.
 - 2: **Output:** Approximations π_K^N and Z_K^N of π and Z .
 - 3: Sample $X_0^i \sim \pi_0$ and set $W_0^i = \frac{1}{N}$ and $Z_0^N = 1$.
 - 4: **for** $k = 1, \dots, K$ **do**
 - 5: Solve $T_k \leftarrow \operatorname{argmin}_{T \in \mathcal{T}} \mathcal{L}_k^N(T)$ using e.g. SGD.
 - 6: $(\pi_k^N, Z_k^N) \leftarrow \text{SMC-NF-step}(\pi_{k-1}^N, Z_{k-1}^N, T_k)$
 - 7: **end for**
-

Algorithm 6 CRAFT to simple AFT line difference: additions for CRAFT shown in green and removals from simple AFT shown in red

- 1: **Input:** number of particles N , unnormalized annealed targets $\{\gamma_k\}_{k=0}^K$ such that $\gamma_0 = \pi_0$ and $\gamma_K = \gamma$, resampling threshold $A \in [1/N, 1)$.
 - 2: **Output:** Length J sequence of approximations π_K^N and Z_K^N of π and Z .
 - 3: **for** $j = 1, \dots, J$ **do**
 - 4: Sample $X_0^i \sim \pi_0$ and set $W_0^i = \frac{1}{N}$ and $Z_0^N = 1$.
 - 5: **for** $k = 1, \dots, K$ **do**
 - 6: ~~Solve $T_k \leftarrow \operatorname{argmin}_{T \in \mathcal{T}} \mathcal{L}_k^N(T)$ using e.g. SGD.~~
 - 7: $\hat{h} \leftarrow \text{flow-grad}(T_k, \pi_{k-1}^N)$ using eqn (8).
 - 8: $(\pi_k^N, Z_k^N) \leftarrow \text{SMC-NF-step}(\pi_{k-1}^N, Z_{k-1}^N, T_k)$
 - 9: Apply gradient update \hat{h} to flow T_k
 - 10: **end for**
 - 11: Yield (π_K^N, Z_K^N) and continue for loop.
 - 12: **end for**
-

Algorithm 7 Practical AFT: from Arbel, Matthews and Doucet 2021. This corresponds to Algorithm 2 of the prior paper.

- 1: **Input:** number of training, test and validation particles $N_{\text{train}}, N_{\text{test}}, N_{\text{val}}$, unnormalized annealed targets $\{\gamma_k\}_{k=0}^K$ such that $\gamma_0 = \pi_0$ and $\gamma_K = \gamma$, resampling thresholds $A_a \in [1/N_a, 1)$ for $a \in \{\text{train, test, val}\}$, number of training iterations J .
 - 2: **Output:** Approximations $\pi_{K,\text{test}}^{N_{\text{test}}}$ and $Z_{K,\text{test}}^{N_{\text{test}}}$ of π and Z .
 - 3: **for** $a \in \{\text{train, test, val}\}$ **do**
 - 4: Sample $X_0^{i,a} \sim \pi_0$ and set $W_0^{i,a} \leftarrow \frac{1}{N_a}$ and $Z_0^{N,a} \leftarrow 1$.
 - 5: **end for**
 - 6: **for** $k = 1, \dots, K$ **do**
 - 7: Learn the flow $T_k \leftarrow \text{EarlyStopLearnFlow}(J, X_{k-1}^{N_{\text{train}}}, W_{k-1}^{N_{\text{train}}}, X_{k-1}^{N_{\text{val}}}, W_{k-1}^{N_{\text{val}}})$
 - 8: **for** $a \in \{\text{train, test, val}\}$ **do**
 - 9: $(\pi_{k,a}^{N_a}, Z_{k,a}^{N_a}) \leftarrow \text{SMC-NF-step}(\pi_{k-1,a}^{N_a}, Z_{k-1,a}^{N_a}, T_k)$
 - 10: **end for**
 - 11: **end for**
-

Algorithm 8 Subroutine *EarlyStopLearnFlow* for practical AFT.

- 1: **Input:** Number of training iterations J , training and validation particles and weights $\left\{X_{k-1}^{i,\text{train}}, W_{k-1}^{i,\text{train}}\right\}_{i=1}^{N_{\text{train}}}$ and $\left\{X_{k-1}^{i,\text{val}}, W_{k-1}^{i,\text{val}}\right\}_{i=1}^{N_{\text{val}}}$.
 - 2: **Output:** Estimated flow T_k
 - 3: Initialize flow to identity $T_k = I_D$.
 - 4: Initialize list of flows $\mathcal{T}_{\text{opt}} \leftarrow \{T_k\}$.
 - 5: Initialize list of validation losses $\mathcal{E} \leftarrow \left\{\sum_{i=1}^{N_{\text{val}}} W_{k-1}^{i,\text{val}} D_k(X_{k-1}^{i,\text{val}})\right\}$
 - 6: **for** $j = 1, \dots, J$ **do**
 - 7: Compute training loss using (55) $\mathcal{L}_k^{N_{\text{train}}}(T_k) \leftarrow \sum_{i=1}^{N_{\text{train}}} W_{k-1}^{i,\text{train}} D_k(X_{k-1}^{i,\text{train}})$.
 - 8: Update T_k using SGD step on $\mathcal{L}_k^{N_{\text{train}}}(T_k)$ with approx. gradient (56).
 - 9: Update list of flows $\mathcal{T}_{\text{opt}} \leftarrow \mathcal{T}_{\text{opt}} \cup \{T_k\}$
 - 10: Update list of validation losses $\mathcal{E} \leftarrow \mathcal{E} \cup \left\{\sum_{i=1}^{N_{\text{val}}} W_{k-1}^{i,\text{val}} D_k(X_{k-1}^{i,\text{val}})\right\}$
 - 11: **end for**
 - 12: Return flow with smallest validation error from the list of flows \mathcal{T}_{opt} .
-