PML 5. Fusing Variational Inference and Markov Chain Monte Carlo

Probabilistic Machine Learning Reading Group

Max Hird

January 14, 2026

University of Waterloo, Canada

Variational Inference (VI)

VI is optimisation over the space of distributions

Find $q^* = \text{argmin}_{q \in \mathscr{P}(\mathbb{R}^d)} d\left(q, \pi\right)$

Purpose is to approximate $\mathbb{E}_\pi\left[f(X)\right]$ with $\mathbb{E}_{q^*}\left[f(X)\right]$

$\mathscr{P}\left(\mathbb{R}^d\right)$ is not parametrisable with any parameter that could fit on a computer. Instead we do:

Find $\theta^* = \text{argmin}_{\theta \in \Theta} d\left(q_\theta, \pi\right)$ and compute $\mathbb{E}_{q_{\theta^*}}\left[f(X)\right]$

So VI is **biased** (i.e. $q_{\theta^*} \neq \pi$ in general)

Often q_θ is `nice':

- Its properties (e.g. moments) can be read off
- Sampleable IID

So VI is **fast** (once we've found q_{θ^*})

VI cont.

Often $\pi = \pi\left(\cdot \mid y\right)$ is a Bayesian posterior with y as data

$$KL\left(q_\theta \| \pi\left(\cdot \mid y\right)\right) + \mathbb{E}_{q_\theta}\left[\log \frac{\pi\left(X, y\right)}{q_\theta\left(X\right)}\right] = \log \pi\left(y\right)$$

Define

$$\text{ELBO}\left(\theta\right) := \mathbb{E}_{q_\theta}\left[\log \frac{\pi\left(X, y\right)}{q_\theta\left(X\right)}\right]$$

Decompose

$$\text{ELBO}\left(\theta\right) = \mathbb{E}_{q_\theta}\left[\log \pi\left(X, y\right)\right] + \mathbb{E}_{q_\theta}\left[-\log q_\theta\left(X\right)\right]$$

Markov chain Monte Carlo (MCMC)

We can't easily access the properties of π by, say, sampling from it IID

Therefore MCMC forms a sequence of measures $\{\mu_t\}_{t=0}^{\infty}$ that tend to the target (in some sense)

In particular, measures are represented by states sampled from them $\{X_t\}_{t=0}^{\infty}$, and the dependencies between these states is Markovian

MCMC cont.

Markovian dependence (often) increases the variance of the estimators formed with the states $\{X_t\}_{t=0}^{\infty}$

$$\mathbb{E}_{\pi}\left[f(X)\right] \approx \frac{1}{T-T_0}\sum_{t=T_0+1}^{T} f\left(X_t\right)$$

Therefore MCMC is **slow** because it is inherently serial i.e. to get X_t we need X_{t-1} for which we need X_{t-2} etc.

But it is **asymptotically exact** e.g.

$$\frac{1}{T-T_0}\sum_{t=T_0+1}^{T} f\left(X_t\right) \rightarrow \mathbb{E}_{\pi}\left[f(X)\right] \text{ a.s.}$$

Markov Kernel Notation

A time-homogeneous Markov chain can be defined by a Markov kernel $K(x \to \cdot\,) \in \mathscr{P}\left(\mathbb{R}^d\right)$ for $x \in \mathbb{R}^d$

K can be viewed as an **operator** on $\mathscr{P}\left(\mathbb{R}^d\right)$: for all $\mu \in \mathscr{P}\left(\mathbb{R}^d\right)$ define $\mu K \in \mathscr{P}\left(\mathbb{R}^d\right)$ with

$$\mu K(A) = \int_{\mathbb{R}^d} \mu\left(\mathrm{d}x\right) K(x \to A) \text{ for all } A \in \mathscr{B}\left(\mathbb{R}^d\right)$$

i.e. to sample from $X \sim \mu K$ we simply sample $Y \sim \mu$ and then $X \sim K(Y \to \cdot\,)$

So $\left\{\mu_t\right\}_{t=0}^{\infty} = \left\{\mu_0 K^t\right\}_{t=0}^{\infty}$

If $\pi = \pi K$ then we call π an **invariant distribution** of K

Markov chain theory

We assume that $\mu_0 K^t \to \pi$ (in some sense) for all $\mu_0 \in \mathscr{P}\left(\mathbb{R}^d\right)$

[Meyn and Tweedie 1993 Proposition 13.2.2]: if π is an invariant measure of K then $||\mu_0 K^t - \pi||_{\mathrm{TV}}$ is non-increasing in t

Key insight 1: $\mu_0 K^t$ is closer to π than μ_0

Much of MCMC theory is the attempt to find conditions under which existing Markov kernels obey

$$d\left(\mu_0 K^t, \pi\right) \leq C\left(\mu_0\right) r(t) + b$$

Where $r(t)$ is monotonically decreasing, r and b depend on K (e.g. via its parametrisation/tuning) and π

Key insight 2: Efficiency of MCMC is sensitive to its parametrisation/tuning and initial distribution

MCMC within VI: General Idea

Using **Key Insight 1** we know that K will push a variational distribution q closer to π

Using **Key Insight 2** we know that how close will depend on q and K: therefore we can use variational methods to optimise over the q space and the K space

Markov chain VI [Salimans, Kingma, Welling 2015]

Main Idea: Use the Tth state in an MCMC chain as a variational approximation i.e. use $q_\theta = \mu_0 K^T$

Problem: ELBO needs access to the density of $\mu_0 K^T$

Solution:

$$\text{ELBO} = \log \pi(x) - KL\left(\mu_0 K^T \| \pi\right)$$

KL is wrt to a new distribution that is optimised over. Authors define

$$\text{ELBO}_{\text{aux}} = \text{ELBO} - \mathbb{E}_{\mu_0 K^T}\left[\text{KL}(\cdots)\right]$$

Where

$$KL(\cdots) = KL\left(\text{new variational distribution} \| \text{reverse Markov transition}\right)$$

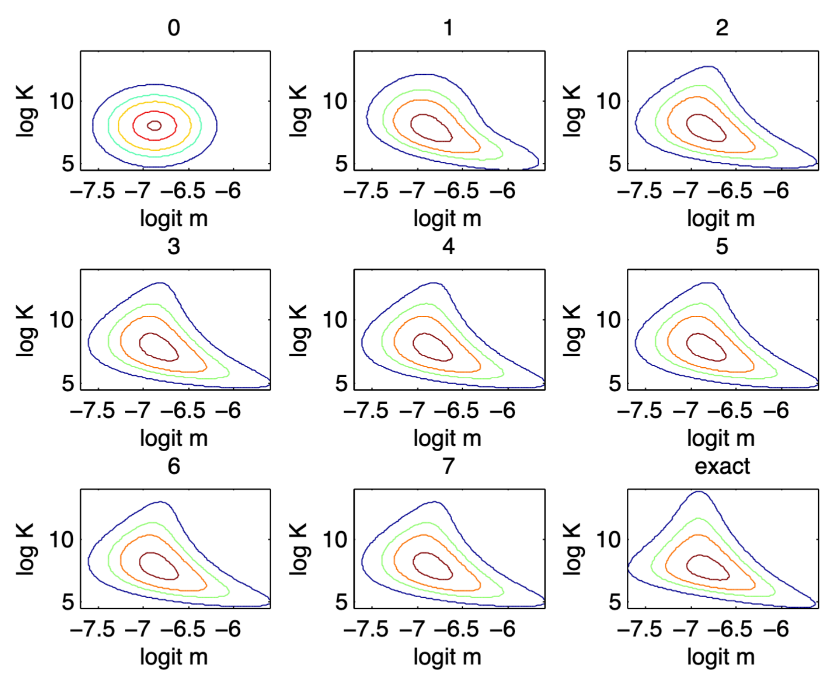
So to optimise ELBO_{aux} we're simultaneously optimising over K and the approximation to the reverse of K

Markov chain VI [Salimans, Kingma, Welling 2015] cont.

We can sample from $\mu_0 K^T$ and so we can get an unbiased estimate of ELBO_{aux}

Therefore we can use autodifferentiation to take the derivative of the method by which we get the estimate to give us an unbiased estimate of the gradient of ELBO_{aux}

Gradients are calculated wrt the variational approximation to the reverse Markov transition, and the parameters of K



Markov chain VI [Salimans, Kingma, Welling 2015] cont.

Problem: accept/reject chains mean ELBO_{aux} is no longer continuously differentiable (wrt some parameter)

Solution: Rao-Blackwellise ELBO_{aux} wrt that parameter

Problem: This operation is exponentially expensive in the chain length

Takeaways:

- The ELBO is no longer calculable if using a Markov kernel
- Accept\reject chains cause discontinuity in the objective
  - Although accept/reject chains are usually the only ones for which we can ensure π invariance
- According to the authors, improving K reduces the variance of gradient estimates

Amortised MCMC [Li, Turner, Liu 2017]

MCMC algorithms are constructed so that π is the unique solution to the fixed point equation $\pi = \pi K$ (this + other conditions ensures that $\mu_0 K^t \to \pi$)

Approximating π can therefore be done by approximating a solution to the fixed point equation:

$$\theta^* = \text{argmin}_{\theta \in \Theta} d\left(q_\theta K^T, q_\theta\right)$$

So do

$$\theta_t = \theta_{t-1} - \eta \nabla_\theta d\left(q_{\theta_{t-1}} K^T, q_{\theta_{t-1}}\right)$$

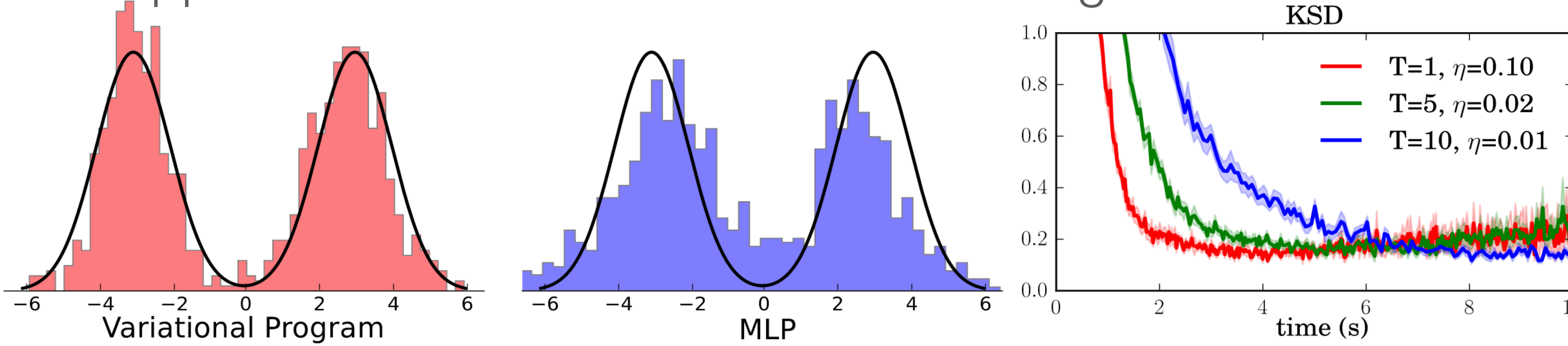
Problem: $d = KL$ needs density evaluation of $q_\theta K^T$

Solution: use a different d where

- We don't need to evaluate the density
- Gradients can be estimated using Monte Carlo

Amortised MCMC [Li, Turner, Liu 2017]

This approach can be viewed as the fine-tuning of a coarser model:



Takeaways:

- Again we have to reformulate or find a new objective due to effect of the Markov kernel on the approximation density
- Authors observe different dynamics for different T's which is interesting
- From the fixed point equation: $T = 1$ should in theory be fine
    - But practically K might be highly inefficient i.e. an accept/reject kernel with a high rejection rate

The Variationally Inferred Sampler [Gallego, Ríos Insua 2021]

Our variational approximation is $\mu_0 K^T$

In [Salimans, Kingma, Welling 2015] the authors optimise K

In [Li, Turner, Liu 2017] the authors optimise μ_0

In [Gallego, Ríos Insua 2021] the authors optimise both μ_0 and K

i.e. find

$$(\theta^*, \eta^*) = \text{argmin}_{\theta \in \Theta, \eta \in \Gamma} \mathscr{L}\left(q_\theta K_\eta^T, \pi\right)$$

As always, the entropy term in the ELBO is intractable

Questions

VI within MCMC

As we saw in [Gallego, Ríos Insua 2021], the parameters of K can be optimised

This idea has been explored in the subjects of `preconditioning' and `adaptive MCMC' that have been around for >20 years

However it's not easy to distil the efficiency of K down to a single quantity (like, say, the ELBO in VI)

According to folklore understanding, properties of K should look like properties of π

E.g. we might want $\text{Cov}_{K(x \to \cdot)}(X) = \text{Cov}_{\pi}(X)$ for all $x \in \mathbb{R}^d$

Otherwise, unless K has a distribution as a tuning parameter, it's not fully clear how to straightforwardly plug VI into MCMC

Nonlinear Preconditioning via Transport based VI

Transport based VI pushes a simple distribution ν through a diffeomorphism $T_\theta : \mathbb{R}^d \to \mathbb{R}^d$ to approximate π

i.e. find

$$\theta^* = \text{argmin}_{\theta \in \Theta} KL\left(T_\theta \sharp \nu || \pi\right)$$

If ν is an `easily sampleable' distribution and the VI is successful then $KL\left(T_\theta \sharp \nu || \pi\right) = KL\left(\nu || T_\theta^{-1} \sharp \pi\right)$ will be low and $T_\theta^{-1} \sharp \pi$ will be `easily sampleable'

1. Find θ^* using VI
2. Run an MCMC on a $T_{\theta^*}^{-1} \sharp \pi$ target
3. Transform states of the resulting Markov chain through T_{θ^*}

Nonlinear Preconditioning via Transport based VI

Methods developed according to this approach fall roughly into two categories:

1. Measure Transport (Papers from Youssef Marzouk, [Kim et al. 2013])
2. Normalising flows (Papers from Marylou Gabrié, [Hoffman et al. 2022], [Kanwar 2024])

These categories can be distinguished by the form of T_θ

1. T_θ has the form of a Knothe-Rosenblatt map
2. T_θ is a normalising flow

We need T_θ to be invertible, to have a Jacobian whose determinant is easily calculable, to be expressive.

Adaptive MCMC

MCMC Kernels usually have tuning parameters

`Good' values of these parameters are often calculable using expectations wrt π or K (notions of optimality are unclear)

E.g. the unadjusted Langevin algorithm with parameter $L \in \mathbb{R}^{d\times d}$ (full rank)

$$X_{t+1} = X_t + \frac{\sigma^2}{2} L L^\top \nabla \log \pi\left(X_t\right) + \sigma L \xi \text{ where } \xi \sim \mathcal{N}\left(0, \mathbf{I}_d\right)$$

There are various justifications for

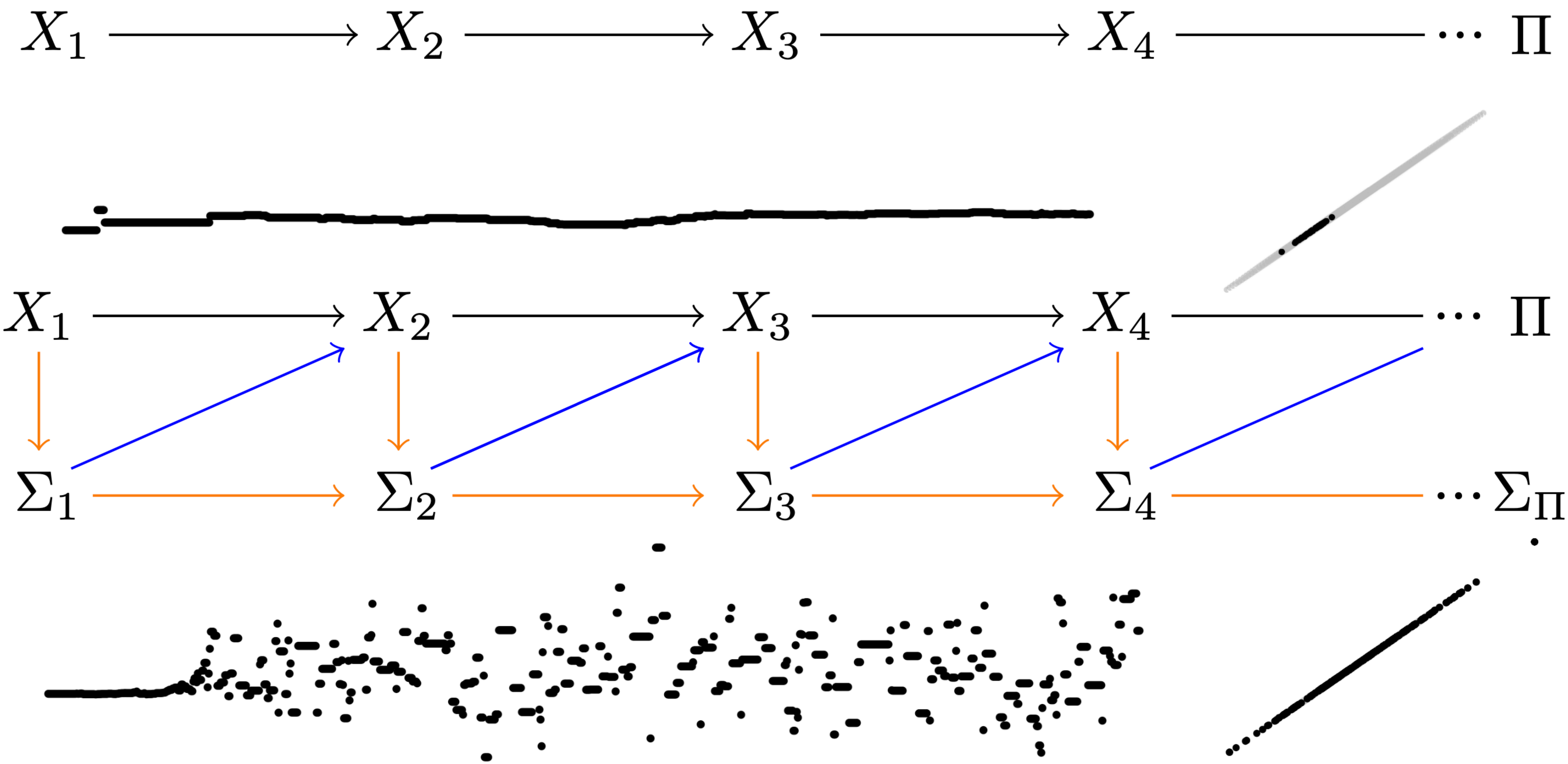
$$LL^\top = \text{Cov}_\pi(X) \text{ and } LL^\top = \text{Cov}_\pi\left(\nabla \log \pi(X)\right)^{-1}$$

See e.g. [Titsias 2023, Hird and Livingstone 2025]

Or we may want to maximise the expected acceptance probability of an accept/reject method (hence the expectation is wrt the proposal distribution)

Adaptive MCMC

Key Idea: Since we get approximate samples from π and exact samples from the proposal, we can optimise as the chain runs:



Fusion with VI: Use VI methods to optimise

Gradient Based Adaptive MCMC [Titsias and Dellaportas 2019]

Let K be an accept/reject kernel with proposal $q_\theta(x \to \cdot\,) \in \mathscr{P}\left(\mathbb{R}^d\right)$

Define the `speed measure':

$$s_\theta(x) := \exp\left(\beta \mathscr{H}_{q_\theta(x \to \cdot\,)}\right) \int_{\mathbb{R}^d} q_\theta\left(x \to \mathrm{d}y\right) \alpha\left(x \to y; \theta\right)$$

Derive a lower bound on $\log s_\theta(x)$:

$$\mathscr{F}_\theta(x) := \int_{\mathbb{R}^d} q_\theta\left(x \to \mathrm{d}y\right) \log \alpha\left(x \to y; \theta\right) + \beta \mathscr{H}_{q_\theta(x \to \cdot\,)}$$

And maximise at each step of the chain using a one sample Monte Carlo estimator

Note: Similarity with ELBO

IMH with Normalising Flows [Brofos, Gabrié et al. 2022]

Let $T_\theta : \mathbb{R}^d \to \mathbb{R}^d$ be a normalising flow and $\nu \in \mathscr{P}\left(\mathbb{R}^d\right)$ be a simple distribution

Authors want to maximise $\mathbb{E}_\pi\left[\log T_\theta \sharp \nu\left(X\right)\right]$ which is the same as wanting to minimise $KL\left(\pi || T_\theta \sharp \nu\right)$

$T_\theta \sharp \nu$ is then used as a proposal distribution in an Independent Metropolis Hastings kernel

So for each new state X_t in the Markov chain do:

$$\theta_{t+1} = \theta_t + \varepsilon_n \nabla \log T_\theta \sharp \nu\left(X_t\right)$$

Since we get approximate samples from π we can minimise the forward KL!

The difficulty of integrating VI into MCMC

This last example illustrates what I view to be a fundamental difficulty when trying to use VI in MCMC:

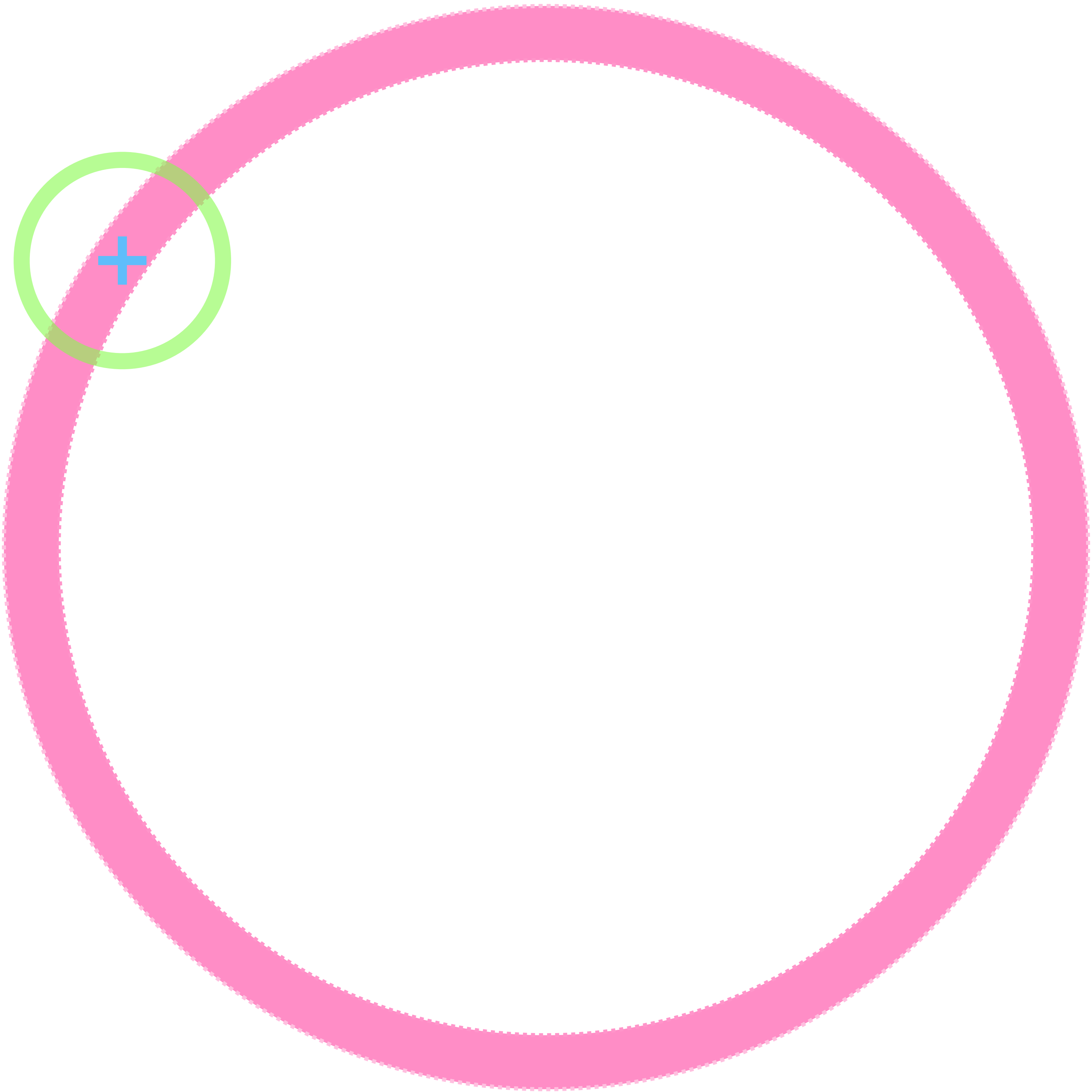
An MCMC Kernel (and its components e.g. proposal) is fundamentally **local**

A variational distribution is fundamentally **global**

Therefore either

- Integrate VI into a Monte Carlo method that is **global** in some sense
  - e.g. Rejection Sampler, Importance Sampler, Independent Metropolis Hastings
  - These are known to fail catastrophically
- Or work out some clever solution

The difficulty of integrating VI into MCMC



Summary

- MCMC in VI: qK^T is closer to π than q
  - The density of qK^T is incalculable so either a new objective must be found, or the ELBO must be approximated
  - Accept/reject chains are attractive to use but come with immediate drawbacks
  - The optimisation process is dependent on K
- VI in MCMC:
  - Transport based VI can be easily fit into the MCMC framework
  - Otherwise it's difficult to straightforwardly apply VI to MCMC because of the conflict between the locality of the Markov kernel and the global nature of the variational approximation