Bidirectional compositionality in inference and stochastic optimization

Moritz Schauer

May, 2022

Chalmers University of Technology and University of Gothenburg.
With Frank van der Meulen (TU Delft)

Illustration: Markov chain

Simple Markov chain $X_0, X_1, \dots, X_s$ taking values in finite sets $E_s = \{1, \dots, n_s\}$. Starting distribution as *row vector* q_0:

$$q_0[, i] = \mathrm{P}(X_0 = i) \quad \text{probability of starting in state } i$$

A transition *matrix* p_s (rows sum to 1) for step $s = 1, 2, \dots$:

$$p_s[i, j] = \mathrm{P}(X_s = j \mid X_{s-1} = i) \quad \text{probability to move to } j \text{ if in } i$$

Marginal distributions q_s after 0, 1, 2,... steps:

$$q_0, \quad q_1 := q_0 p_1, \quad q_2 := q_0 p_1 p_2, ...$$

Conditional probability to end in $X_t = l$ after starting from $X_s = i$:

$$h_s[i,] = (p_{s+1} \dots p_t)\,[i, l]$$

Illustration: Bayes

Observe $X_t = l$.

(Marginal) posterior $X_s \mid X_t = l$, $s < t$

$$\begin{aligned} \mathrm{P}(X_s = i \mid X_t = l) &= \frac{\mathrm{P}(X_s = i)\mathrm{P}(X_t = l \mid X_s = i)}{\mathrm{const}} \\ &= \frac{q_s[,i]h_s[i,]}{q_s h_s} \end{aligned}$$

with

$$\begin{aligned} q_0 p_1 \dots p_s =: q_s \\ h_s := p_{s+1} \dots p_t h_t \end{aligned}$$

and

$$h_t[i,] = \mathrm{P}(X_t = i \mid X_t = l) = \begin{cases} 0 & i \neq l \\ 1 & i = l \end{cases}.$$

Conditional Markov chain

Defining

$$\pi_s[,j] = \mathrm{P}(X_s = j \mid X_t = l)$$

we get an evolution for the conditional

$$\pi_s[,j] \ = \sum_i \pi_{s-1}[,i] \underbrace{\frac{h_s[j,]p_s[i,j]}{(p_s h_s)[i,]}}_{=:\ p_s^\star[i,j]}$$

or

$$\pi_s \ = \pi_{s-1} p_s^\star$$

where $p_s^\star$ is again a stochastic matrix.

Bidirectional machinery

Consuming a column vector h and a row vector π to produce kh and $\pi p^\star$

$$ph \xleftarrow{p} h$$

$$\pi \xmapsto{p^\star} \pi p^\star$$

where

$$(\pi p^\star)[,j] = \sum_i \pi[,i] \frac{h[j,]p[i,j]}{(ph)[i,]}$$

Generative model

```julia
sampled(rng, x, p) = rng, sample(rng, weights(p[x,:]))

function generate(rng, x, ps)
    xs = [x]
    for p in ps
        rng, x = sampled(rng, x, p)
        push!(xs, x)
    end
    return xs
end
xs = generate(rng, x0, ps)
```

1

Backward-forward transformed code

```
function backward(p, h)
    ph = p*h
    m = ph, h # needed in forward
    return m, ph
end

function forward(rng, x, p, m)
    ph, h = m # from backward
    pstarx = [p[x,j]*h[j]/ph[j] for j in 1:d]
    rng, sample(rng, weights(pstarx))
end
```

2

Backward-forward transformed code

```
function htransformed(rng, x, ps, h)
    xs = [x]
    ms = []
    for p in reverse(ps)
        m, h = backward(p, h)
        pushfirst!(ms, m)
    end
    for (p, m) in zip(ps, ms)
        rng, x = forward(rng, x, p, m)
        push!(xs, x)
    end
    return xs
end
h = ps[end][:, y]
posterior = htransformed(rng, x0, ps[1:end-1], h)
```

3

Contents

Table of contents

Category BorelStoch

Objects in $\textsc{BorelStoch}$ are standard Borel measure spaces $S = (E, \mathcal{B})$, $S' = (E', \mathcal{B}')$ (spaces equipped with σ-fields). $S \otimes S' = (E \times E', \mathcal{B} \otimes \mathcal{B}')$ defines a tensor product.

Take $I = (1, \{\varnothing, \{1\}\})$ the single element measure space to be formal unit of the tensor product $\otimes$

$$I \otimes S = S$$

Category BorelStoch

Arrows

$$p \colon S \rightharpoondown S'$$

in $\textsc{BorelStoch}$ are Markov kernels $p \colon E \times \mathcal{B}' \to [0,1]$ such that

$$p(x, \cdot) \text{ is a distribution parametrised by } x \in E,$$

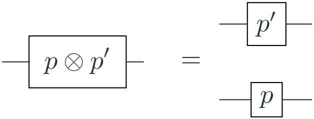
Familiar example of a Markov kernel:

$$p(x, A) = \int_A \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}(y-x)^2} \mathrm{d}y$$

$p(x, \cdot)$ roughly corresponds to the (Julia-) code `x -> Normal(x, 1)`.

Parallel composition

Parallel composition of arrows $p\colon S \rightarrow T$, $p'\colon S' \rightarrow T'$

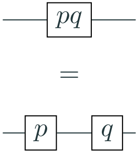


is by the tensor product

$$(p \otimes p')((x, x'), \mathrm{d}x \times \mathrm{d}y') = p(x, \mathrm{d}y)p'(x', \mathrm{d}y').$$

Composition

Sequential composition of $p\colon S \rightarrowtriangle T$, $q\colon T \rightarrowtriangle U$



by Chapman-Kolmogorov

$$pq\colon S \rightarrowtriangle U$$

$$(pq)(x, \mathrm{d}z) = \int_y q(y, \mathrm{d}z) p(x, \mathrm{d}y)$$

with identity $\mathrm{id}_S\colon S \rightarrowtriangle S$, $\mathrm{id}_S(x, \mathrm{d}y) = \delta_x(\mathrm{d}y)$ (Dirac).

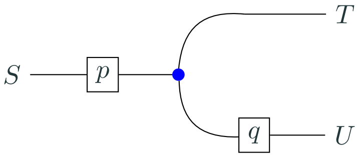
Joint laws

Model: $S \xrightarrow{p} T \xrightarrow{q} U$.

The Markov kernel

$$p \bullet q \colon S \to T \otimes U$$
$$(p \bullet q)(x, \mathrm{d}y \times \mathrm{d}z) := p(x, \mathrm{d}y) q(y, \mathrm{d}z)$$

represents the joint distribution on $T \otimes U$ given $x \in S$



$p \bullet q = p\Delta(\mathrm{id}_T \otimes q)$ with duplication kernel $\Delta \colon T \to T \otimes T$ with $\Delta(x, \mathrm{d}y \times \mathrm{d}z) = \delta_x(\mathrm{d}y)\delta_x(\mathrm{d}z)$.

Table of contents

Distributions

A Markov kernel $p\colon S \rightharpoonup T$.

Distributions π on S compose with p as

$$(\pi p)(\mathrm{d}y) = \int_x p(x, \mathrm{d}y)\pi(\mathrm{d}x) \quad \text{(Push forward)}$$

Distributions can be identified with Markov kernels $q\colon I \rightharpoonup S$ setting $\pi(\cdot) = q(1, \cdot)$.

When taking Markov kernels as maps $F(p)\colon \mathcal{P}(E) \to \mathcal{P}(E')$ acting on sets of distributions $(\mathcal{P}(E), \otimes)$

$$(\pi \otimes \pi')(p \otimes p') = (\pi p) \otimes (\pi' p')$$

Effects

A Markov kernel $p\colon S \multimap T$.

Effects or likelihoods i.e. positive random variables h on T

$$(ph)(x) = \int_y h(y)p(x, \mathrm{d}y) \quad \text{(Pullback)}$$

Dual pairing / scalar product of measures and effects

$$\pi h = \int_x h(x)\pi(\mathrm{d}x) = \mathbb{E}_\pi h$$

Measures and densities

Absolute continuity $q \ll p$ of two measures $p(A) = 0 \Rightarrow q(A) = 0$.

For two probability measures on $S = (E, \mathcal{B})$ this is equivalent to that q has a p-density $\frac{\mathrm{d}q}{\mathrm{d}p} \colon E \to [0, \infty)$

$$q(A) = \int_A \frac{\mathrm{d}q}{\mathrm{d}p} \mathrm{d}p \quad \text{or} \quad q = \frac{\mathrm{d}q}{\mathrm{d}p} \centerdot p$$

Example: $f \centerdot \lambda$ with $f(y) = \frac{1}{\sqrt{2\pi}} \exp(-(y-x)^2/2)$ and λ the Lebesgue measure defines the standard normal distribution with mean x.

Bayes rule

Give $I \overset{p}{\rightarrow} S \overset{q}{\rightarrow} T$ with $q(x, \cdot) \ll \lambda$ dominated by a reference measure.

Also pair of variables $(X, Y)\colon (\Omega, \mathcal{F}, \mathbf{P}) \to S \otimes T$ with joint distribution $p \bullet q$

Bayes rule: The posterior distribution $p^\star$ of X given $Y = y$ has a p-density

$$\frac{\mathrm{d}p^\star}{\mathrm{d}p} = \frac{h}{ph}, \qquad \text{where } h(x) = \frac{\mathrm{d}q(x, \cdot)}{\mathrm{d}\lambda}(y) \text{ is the likelihood}$$

▶ *The likelihood is the unnormalised posterior density (with respect to the prior)*

Table of contents

h-transform of a Markov kernel

Given a Markov kernel $p\colon S \rightarrowtriangle T$ and effect $h\colon T \to [0,\infty)$ we can define a new *Markov* kernel

$$p^\star(x, A) = \int_A \frac{h(y)}{(ph)(x)} p(x, \mathrm{d}y)$$

Here the normalization constant $(ph)(x)$ makes $p^\star$ Markov.

With $m(x,y) = \frac{h(y)}{(ph)(x)}$ we write short

$$p^\star = m \bullet p$$

Transport/Forcing

Given $I \overset{q}{\rightarrow} S \overset{p}{\rightarrow} T$ and a probability measure $\mu \ll pq$ on T. Then the h-transform of p with the effect

$$h = \frac{\mathrm{d}\mu}{\mathrm{d}(qp)}$$

transports q into the marginal μ:

$$qp^\star = \mu$$

$$\int_A q(\mathrm{d}x) \tfrac{\mathrm{d}\mu}{\mathrm{d}(qp)}(y) p(x, \mathrm{d}y) = \int_A \tfrac{\mathrm{d}\mu}{\mathrm{d}(qp)}(y)(qp)(\mathrm{d}y) = \mu(A)$$

h-transform synthetically

In a *non-causal* Markov category effects can take the form of a (non-Markov) kernel

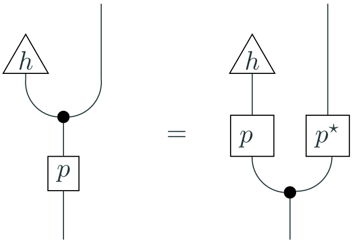
$$h \colon T \rightharpoonup I$$

where I is the terminal object. Think of $h(y, \{1\}) = h(y), h(y, \varnothing) = 0$.

The h-transform of $p \colon S \rightharpoonup T$ can be defined synthetically as $p^\star$ with

$$p\Delta(h \otimes \mathrm{id}) = \Delta((ph) \otimes p^\star).$$

This can be expressed as string diagram (bottom-to-top), with effects denoted by triangles:



Kullback-Leibler divergence

The Kullback-Leibler divergence

$$\mathrm{KL}(q \parallel p) = \begin{cases} \int \log \frac{\mathrm{d}q}{\mathrm{d}p} \mathrm{d}q & q \ll p \\ \infty & \text{otherwise} \end{cases}$$

For Markov kernels $p, q \colon S \rightharpoonup T$, KL is a function of x,

$$\mathrm{KL}(q \parallel p)|_x = \mathrm{KL}(q(x, \cdot) \parallel p(x, \cdot))$$

Donsker and Varadhan variational formula

Proposition

If $p\colon S \rightharpoonup T$ and h is an effect on T with $ph > 0$, then

$$\log ph = \sup_{q\colon q \ll p} \{q \log h - \mathrm{KL}(q \parallel p)\}$$

If $ph < \infty$, then the supremum on the right-hand side is attained if and only if $q = p^\star = \frac{h}{ph} \bullet p$ or

$$\frac{\mathrm{d}p^\star}{\mathrm{d}p} = \frac{h}{ph}$$

- *A posterior solves an optimisation problem!*

Variational formula

Proof.

Part 1: By Jensen's inequality, if $q \ll p$,

$$\log ph = \log \mathbb{E}_p\, h = \log \mathbb{E}_q\, \exp(\log h - \log \frac{\mathrm{d}q}{\mathrm{d}p})) \geq E_q \log h - \mathbb{E}_q \log \frac{\mathrm{d}q}{\mathrm{d}p}.$$

Part 2: $\log h - \log \frac{\mathrm{d}p^\star}{\mathrm{d}p} = \log h - (\log h - \log ph) = \log ph$ is constant. □

Bellman principle

Model: $S_0 \xrightarrow{p_1} S_1 \xrightarrow{p_2} S_2$. Fix x_0, so $p_1 = p_1|_{x_0}$ becomes a probability on S_1 and $p_{1,2} = p_1 \cdot p_2$ a joint probability on $S_1 \otimes S_2$.

Task: Given a likelihood $h_2(x_2)$,

$$\max_{q_{1,2} \ll p_{1,2}}! \qquad \mathbb{E}_{q_{1,2}} \log h_2 - \mathrm{KL}(q_{1,2} \parallel p_{1,2})$$

Setting $q_{1,2} = q_1 \cdot q_2$ where $S_0 \xrightarrow{q_1} S_1 \xrightarrow{q_2} S_2$ this can be rewritten

$$\sup_{q_1, q_2} \left\{ q_1 q_2 \log h_2 - q_1 \log \frac{\mathrm{d}q_1}{\mathrm{d}p_1} - q_1 q_2 \log \frac{\mathrm{d}q_2}{\mathrm{d}p_2} \right\}$$
$$= \sup_{q_1} \left\{ q_1 \sup_{q_2} \left\{ q_2 \log h_2 - q_2 \log \frac{\mathrm{d}q_2}{\mathrm{d}p_2} \right\} - q_1 \log \frac{\mathrm{d}q_1}{\mathrm{d}p_1} \right\}$$

▶ *Bellman: The best first step $q_1 = p_1^\star$ is the one which maximises the overall objective if it is followed by optimal remaining step(s) $q_2 = p_2^\star$.*

Bellman principle

Introducing the value functions V_i the supremum is found by backward recursion

$$
\begin{aligned}
V_2(x_2) &= \log h_2(x_2) \\
V_1(x_1) &= \sup_{q_2} \left\{ (q_2 V_2)(x_1) - \mathrm{KL}(q_2 \parallel p_2)\big|_{x_1} \right\} \\
V_0(x_0) &= \sup_{q_1} \left\{ (q_1 V_1)(x_0) - \mathrm{KL}(q_1 \parallel p_1)\big|_{x_0} \right\}
\end{aligned}
$$

Bellman principle

Optimal step q_2: Now taking the maximum of q_2 first

$$\begin{aligned} V_1(x_1) &= \sup_{q_2} \left\{ (q_2 \log h_2)(x_1) - \mathrm{KL}(q_2 \parallel p_2)|_{x_1} \right\}. \\ &= (\log p_2 h_2)(x_1) \\ &= (\log h_1)(x_1) \quad \text{(with } h_1 := p_2 h_2) \end{aligned}$$

is obtained in $q_2 = p_2^\star$ by

$$\frac{p_2^\star(x_1, \mathrm{d}x_2)}{p_2(x_1, \mathrm{d}x_2)} = \frac{h_2(x_2)}{h_1(x_1)}.$$

Bellman principle

Optimal step q_1: Plugging in the value $\log h_1 := \log p_2 h_2 = V_1(x_1)$ gives the objective

$$
\begin{aligned}
V_0 &= \sup_{q_1} \{q_1 \log(h_1) - \mathrm{KL}(q_1 \parallel q_2)\} \\
&= \log(p_1 h_1) \\
&= \log h_0, \quad \text{with } h_0 := p_1 h_1 = p_1 p_2 h_2
\end{aligned}
$$

found in $q_1 = p_1^\star$,

$$
\frac{p_1^\star(x_0, \mathrm{d}x_1)}{p_1(x_0, \mathrm{d}x_1)} = \frac{h_1(x_1)}{h_0(x_0)}.
$$

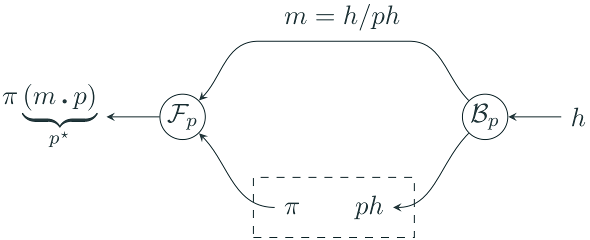
Table of contents

Structure

This is very suggestive,

$$(p_1 p_2)^\star = p_1^\star p_2^\star, \quad (p_1 \centerdot p_2)^\star = p_1^\star \centerdot p_2^\star$$

Note that here $p_1^\star$ has a "hidden" dependency on $h_1 = p_2 h_2$. To make $p \mapsto p^\star$ "functorial" we have to make the dependency explicit.

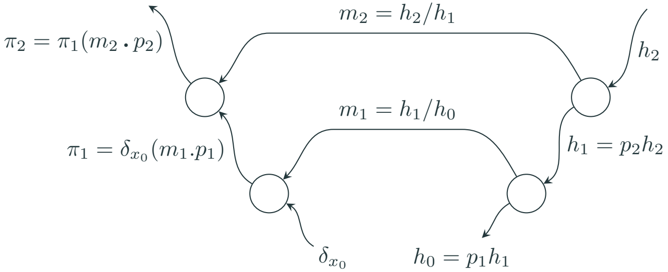


String diagram of composition

Model $S_0 \overset{p_1}{\to} S_1 \overset{p_2}{\to} S_2$. Given $h_2 \colon S_2 \to \mathbb{R}_{\geq 0}$.

Task: For fix x_0 compute π_1 and π_2, the marginal of the maximizer π of

$$\mathbb{E}_\pi h_2 - \mathrm{KL}(\pi \parallel p_1 \bullet p_2).$$



Directly or by the Bellman principle the h-transform composes optically.

Table of contents

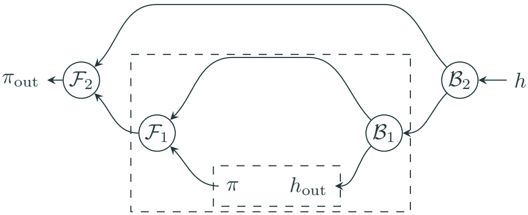
Key building block: optic

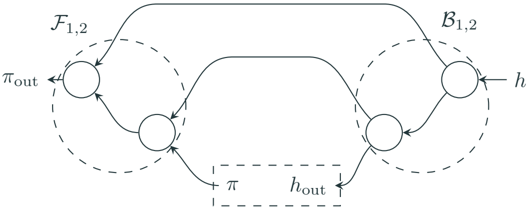
$\mathcal{P}(S)$ measures on S. $\mathcal{M}(S)$ functionals on S. $\mathbf{M}$ space of messages

- $\mathcal{F}\colon \mathcal{P}(S) \times \mathbf{M} \to \mathcal{M}(S')$
- $\mathcal{B}\colon \mathcal{M}(S') \to \mathbf{M} \times \mathcal{M}(S)$
- Compatible $\mathcal{F}_p$ and $\mathcal{B}_p$ work as pairs:

$$\langle \mathcal{F} \mid \mathcal{B} \rangle \colon \mathcal{P}(S) \times \mathcal{M}(S) \to \mathcal{P}(S') \times \mathcal{M}(S')$$

Composition of optics

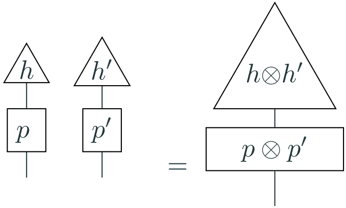




$$\langle \mathcal{F}_{1,2} \mid \mathcal{B}_{1,2} \rangle \cong \langle \mathcal{F}_1 \mid \mathcal{B}_1 \rangle \langle \mathcal{F}_2 \mid \mathcal{B}_2 \rangle$$

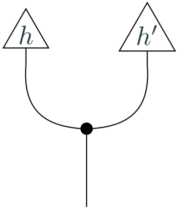
Effects and composition

For $p\colon S \rightarrow T$, $p'\colon S' \rightarrow T'$ and effects h, h' on T, T'.



Fusion

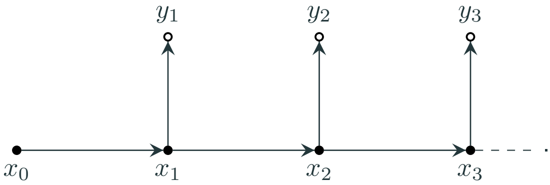
Fusion as pullback of product effects through duplication:



$$(\Delta(h \otimes h'))(x) = h(x)h'(x)$$

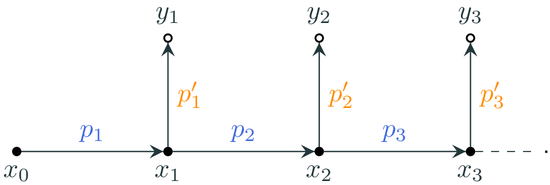
If you ever see $\left(\Sigma_1^{-1} + \Sigma_2^{-1}\right)^{-1} = \Sigma_1 - \Sigma_2 \left(\Sigma_1 + \Sigma_2\right)^{-1} \Sigma_2 \ldots$

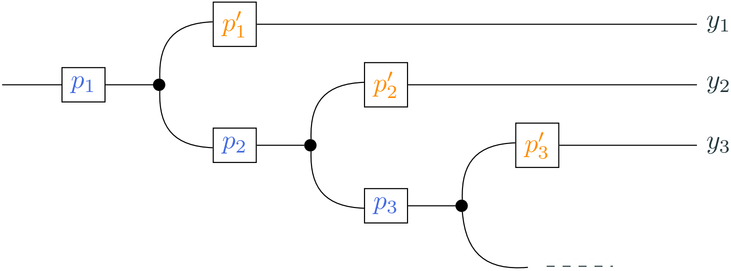
Example: State-space model



"Classic" diagram for a *state-space model*.

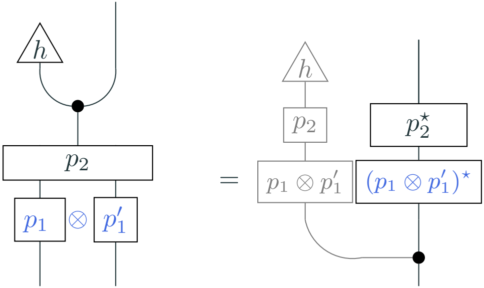
String diagram for a state-space model





Transform with h of product form gives the Kalman (RTS) smoother.

Collider



Conditioning on common effects makes (marginally) independent transitions dependent.

Table of contents

Concessions

Take $p\colon S \rightharpoonup T$. The maximizer $q = \frac{h}{ph}.p$ of

$$q \log h - \mathrm{KL}(q \parallel p)$$

will be hard to find. Hence we use variational methods or Monte Carlo methods guided by heuristics.

1. Choose

$$\widetilde{p} = \operatorname*{argmax}_{q \in \mathcal{Q}} \{q \log h - \mathrm{KL}(q \parallel p)\}$$

where $\mathcal{Q}$ is a variational class of Markov kernels $S \rightharpoonup T$.

   - Variational Bayes.

2. Use a heuristic $\widetilde{h} \approx h$ instead of the true cost/likelihood

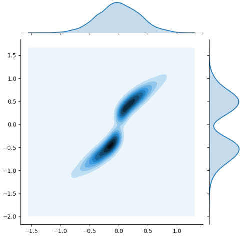
$$p^\circ = \frac{\widetilde{h}}{p\widetilde{h}} \cdot p, \quad w = \left(\frac{\widetilde{h}}{p\widetilde{h}}\right)^{-1}$$

   - Guided processes.

Table of contents

Optimal transport

Two distributions μ_0 on E_0 and μ_1 on E_1 and a cost function $c\colon E_0 \times E_1 \to \mathbb{R}$.



Optimal transport (Kantorovich formulation): Find a joint distribution q with marginals μ_0 and μ_1 minimising the average cost

$$qc \left(= \int c(x_0, x_1) q(\mathrm{d}x_0 \times \mathrm{d}x_1) \right)$$

$qc + \epsilon \, \mathrm{KL}(q \parallel \mu_0 \otimes \mu_1)$ (with entropy regularization.)

Optimal transport

The problem can be written as KL minimization task:

$$\inf_{q \ll p} \mathrm{KL}(q \parallel p) \quad \text{such that } q \text{ has marginals } \mu_0 \text{ and } \mu_1$$

with

$$p(\mathrm{d}x_0 \times \mathrm{d}x_1) \propto \exp(-c(x_0, x_1)/\epsilon)\lambda_0(\mathrm{d}x_0)\lambda_1(\mathrm{d}x_1)$$

where $\lambda_0 \gg \mu_0$ and $\lambda_1 \gg \mu_1$ reference measures

Entropy regularised optimal transport

p a joint distribution on $E_0 \times E_1$ and effects h_0 on E_0 and h_1 on E_1.

Proposition

Let $p^\star$ be the $h_0 h_1$ transformed probability measure

$$p^\star(\mathrm{d}x_0 \times \mathrm{d}x_1) \propto h_0(x_0) h_1(x_1) p(\mathrm{d}x_0 \times \mathrm{d}x_1).$$

- $q = p^\star$ maximises

$$\mathbb{E}_q(\log h_1(X_0) + \log h_2(X_1)) - \mathrm{KL}(q \parallel p)$$

  among all $q \ll p$

- $q = p^\star$ minimises

$$\mathrm{KL}(q \parallel p)$$

  among all $q \ll p$ *with the same marginals as* $p^\star$.

▶ *Whatever I get as optimiser, if it has the right marginals, its the optimal transport plan.*

h-transform

Disintegrate p into the marginal p_0 on S_0 and the conditional $p_1 \colon S_0 \rightarrow S_1$.

$$p = p_0 \cdot p_1 = p_0 \Delta(\mathrm{id}_T \otimes p_1)$$

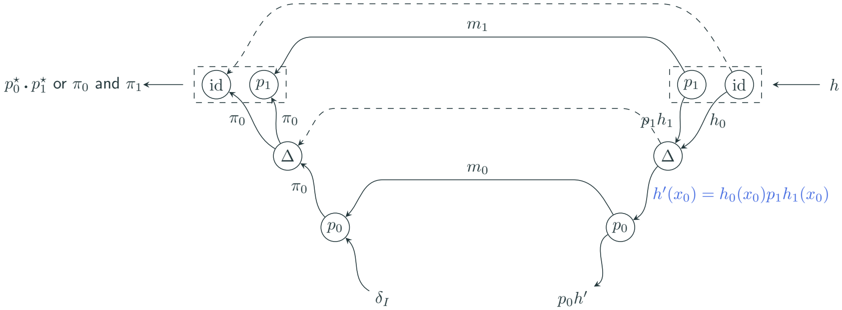
and h-transform by $h_0(x_0)h_1(x_1)$ gives the marginals of the optimiser given h_0, h_1.

$$(p_0 \cdot p_1)^\star = p_0^\star \cdot p_1^\star, \quad p_0^\star = \tfrac{h'}{p_0 h'} \cdot p_0, \quad p_1^\star = \tfrac{h_1}{p_1 h_1} \cdot p_1$$

with

$$h'(x_0) = h_0(x_0)(p_1 h_1)(x_0)$$

Message passing diagram



Sinkhorn

We need to find the *forcing*, the h-transform achieving the right marginals to find the the optimal transport plan. Sinkhorn algorithm uses coordinate descent on h_0 and h_1 to find the forcing.

Iterate until convergence:

$$h_0 = \frac{\mathrm{d}\mu_0}{\mathrm{d}\left(\frac{p_1 h_1}{p_0 p_1 h_1} \cdot p_0\right)} \quad \text{forcing} \quad p_0^\star = \mu_0$$

$$h_1 = \frac{\mathrm{d}\mu_1}{\mathrm{d}\left((h_0 \cdot p_0) p_1\right)} \quad \text{forcing} \quad p_0^\star p_1^\star = \mu_1$$

Table of contents

Continuous-time guided processes

Assume a continuous-time E-valued Markov process $X \equiv (X_u,\, u \in [s,t])$ starting in x_s.

The process is characterised by the space-time generator $\mathfrak{A}$: If for $f\colon [s,t] \times E \to \mathbb{R}$ there is $g\colon [s,t] \times E \to \mathbb{R}$ such that

$$M_\cdot = f(\cdot, X_\cdot) - f(s, X_s) - \int_s^\cdot g(u, X_u)\mathrm{d}u$$

is a local martingale, let $f \in \mathcal{D}(\mathfrak{A})$ (domain) and $\mathfrak{A}f = g$.

Implies a Markov transition kernel

$$p_{s\to t}(x_s, \cdot) = \mathbb{P}(X_t \in \cdot \mid X_s = x_s)$$

Change of measure

Define the *change of measure*

$$\mathrm{d}\mathbb{P}^\circ = D^h[s,t]\mathrm{d}\mathbb{P}$$

with

$$D^h[s,\cdot] = \frac{h(\cdot, X_\cdot)}{h(s,x_s)} \exp\left(-\int_s^\cdot \frac{\mathfrak{A}h}{h}(u,X_u)\mathrm{d}u\right)$$

and $h \in \mathcal{D}(\mathfrak{A})$ is a positive function such that $D^h[s,\cdot]$ is a martingale.

- *Solution to a Hamilton-Jacobi-Bellman equation*

Dynamics of the changed process

By *Palmowski-Rolski (2002)* the space-time generator of X under $\mathbb{P}^\circ$ is

$$\mathfrak{A}^\circ f = \frac{1}{h}\left[\mathfrak{A}(fh) - f\mathfrak{A}h\right] \tag{1}$$

Theorem

For a continuous-time process along an edge with

$$w(X) = \exp\left(\int_s^t \frac{\mathfrak{A}h}{h}(u, X_u)\mathrm{d}u\right)$$

we have

$$\frac{\mathbb{E}[f(X)h(t, X_t)]}{p_{s\to t}h(t,\cdot)} = \mathbb{E}^\circ f(X)w(X)$$

h can be a heuristic here or an actual h-transform with $\mathfrak{A}h = 0$.

Example: Conditional Brownian motion

W is a Brownian motion on $[0,1]$ under the measure p and $Y = X_1 + \epsilon$, $\epsilon \sim N(0,\sigma)$.

$$\mathfrak{A}f = \dot{f} + \frac{1}{2}f'' \quad \text{Space time generator of W}$$

The conditional likelihood of observing $Y = y$ is

$$h(1, x_1) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{1}{2}(y-x_1)^2/\sigma^2}$$

and h solves $\mathcal{A}h = 0$ with that boundary condition.

Then the conditional measure is

$$p^\star = \operatorname{argmax}\{qh - \mathrm{KL}(q \parallel p)\}$$

and the generator of the conditional process is

$$\mathfrak{A}^\star f = \frac{1}{h}\left[\mathfrak{A}(fh) - f\mathfrak{A}h\right] = \dot{f} + \nabla \log h f' + \frac{1}{2}f''$$

The conditional process has drift $\nabla_x \log h(t,x)$.

Table of contents

Large deviations: Sanov's theorem

Donsker-Varadhan variational characterisation

$$\sup_{\log h \in C_b} \{q \log h - \log ph\} = \mathrm{KL}(q \parallel p)$$

has maximiser in $h = \mathrm{d}q/\mathrm{d}p$ if $q \ll p$.

Empirical distribution of random sequence $X_i \overset{\text{i.i.d.}}{\sim} p, i \in \mathbb{N}$

$$\hat{p}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$$

Sanov's theorem: The empirical distribution satisfies the large deviation principle with good rate function $\mathrm{KL}(\cdot \parallel p)$,

$$\mathrm{P}(\hat{p}_n \in B) \asymp \exp\left(-n \inf_{q \in B} \mathrm{KL}(q \parallel p)\right)$$

Large deviations at the final time: h-transform

Let p be the Wiener measure on time span $[0,1]$. Take a independent sequence of canonical Brownian motions $w^{(i)} \sim p$ and fix the marginal measure μ_1 and let

$$B = \{q\colon q \circ w_1^{-1} = \mu_1\}$$

Taking $h = \mathrm{d}\mu_1/\mathrm{d}(q \circ w_1^{-1})$, h is the forcing h-transform such that the maximizer $q = p^\star$ of

$$q \log h - \mathrm{KL}(q \parallel p)$$

has marginal $q\colon q \circ w_1^{-1} = \mu_1$ thus

$$p^\star = \operatorname*{argmax}_{q \in B} \mathrm{KL}(q \parallel p)$$

Large deviations at the final time: h-transform

By h-transform with $h(s,\cdot)$ solving $\mathfrak{A}h = 0$ and $h(1, w_1) = h(w_1)$

$$\mathrm{d}w_t = \nabla \log h(t, w_t)\mathrm{d}t + \mathrm{d}w_t^\star, w_0 = 0$$

where $w_t^\star = w_t - \nabla \log h(t, w_t)\mathrm{d}t$ is a $p^\star$ Brownian motion.

Under the rare event B each $w^{(i)}$ looks like a Brownian motion with drift $\nabla \log h$.

Guiding for large deviations

Give me an approximation $\widetilde{h}$ with $\mathfrak{A}\widetilde{h} \approx 0$ and $\widetilde{h}(1, w_1) = h(w_1)$.

Then by Palmowski-Rolski with guiding process

$$\mathrm{d}w_t^\circ = \nabla \log \widetilde{h}(t, w_t^\circ)\mathrm{d}t + \mathrm{d}b_t$$

for some independent Brownian motion $(b_t)_{t\in[0,1]}$ we have

$$p^\star(A) = \frac{\mathbb{E}\mathbf{1}_A(w^\circ)\,\mathrm{weight}(w^\circ)}{\mathbb{E}\,\mathrm{weight}(w^\circ)}$$

with

$$\text{weight}(w^\circ) = \exp\left(\int_0^1 \frac{\mathfrak{A}\widetilde{h}}{\widetilde{h}}(t, w_t^\circ)\mathrm{d}t\right)$$

Thus sampling w° characterises large deviations in tractable way.

References

Frank van der Meulen, M.S.: Automatic Backward Filtering Forward Guiding for Markov processes and graphical models. https://doi.org/10.48550/arXiv.2010.03509, 2020.

Marcin Mider, M.S., Frank van der Meulen: Continuous-discrete smoothing of diffusions. *Electronic Journal of Statistics 15 (2)*, https://doi.org/10.1214/21-EJS1894, 2021.

Marc Corstanje, Frank van der Meulen, M.S.: Conditioning continuous-time Markov processes by guiding. https://doi.org/10.48550/ARXIV.2111.11377, 2022.

Frank van der Meulen: Introduction to Automatic Backward Filtering Forward Guiding. https://doi.org/10.48550/ARXIV.2203.04155, 2022.

Prelude: Markov process and discrete generator

Model: $p_i \colon S_{i-1} \rightharpoonup S_i$ for $i = 1, \ldots, t$

For fix $x_0 \in S_0$ this defines Markov process $X \equiv (X_i, i = 0, \ldots, t)$ with $X_0 = x_0$ an law $(\delta_{x_0} \bullet p_1 \bullet p_2 \bullet \cdots \bullet p_{t-1} \bullet p_t)$.

For time-dependent functionals $f(s, \cdot)$ on S_s, define the operator

$$(\mathfrak{A} f)(s, x_s) := (p_{s+1} f(s+1, \cdot))(x_s) - f(s, x_s)$$

Then

$$M_t = f(t, X_t) - f(0, X_0) - \sum_{s=0}^{t-1} (\mathfrak{A} f)(s, X_s)$$

is a martingale. ► *Martingales characterise Markov processes*

In particular, for $h(t, \cdot)$ given and $h(s, \cdot) = p_{s+1} h(s+1, \cdot)$

$$M_t = h(t, X_t) - h(0, X_0)$$

is a martingale.