Funsor: Functional Tensors for Probabilistic Programming

Fritz Obermeyer*, Eli Bingham*, Martin Jankowiak*, JP Chen, Du Phan

Introduction: We need 'autograd for integrals'

Probabilistic modelling and inference offer **a unifying approach to many machine learning tasks**, including quantifying uncertainty, learning structured generative models, producing interpretable explanations of data, and learning from weak or missing labels.

Probabilistic programming languages like Pyro allow **specification of probabilistic models in high-level programming languages.**

But many models that mix many different discrete and continuous variables still need custom inference strategies, and there is **no lower-level analogue of automatic differentiation software** intermediate between fully symbolic and fully black box integration.

Functional Tensors: a language for automatic integration over array-valued variables

Probabilistic programs generate lazy expressions with free variables. Inference algorithms integrate over free variables:

$$
\begin{array}{l|l}
\textbf{fun}\ \text{GenerativeModel}(x) & p \leftarrow 1 \\
\quad z \leftarrow \text{sample}(P_z) & p \leftarrow p \times P_z[v=z] \\
\quad y \leftarrow \exp(z) & \\
\quad \text{observe}(P_x[\theta=y],\, x) & p \leftarrow p \times P_x[\theta=y, v=x] \\
\textbf{end} & \text{maximize: } \sum_z p
\end{array}
$$

Approximate inference computations also generate lazy sum-product expressions in the same expression language.

$$
\begin{array}{l|l}
\textbf{fun}\ \text{GenerativeModel}(x) & p \leftarrow 1 \\
\quad z \leftarrow \text{sample}(P_z) & p \leftarrow p \times P_z[v=z] \\
\quad \text{observe}(P_x[\theta=z],\, x) & p \leftarrow p \times P_x[v=x, \theta=z] \\
\textbf{end} & \\
\textbf{fun}\ \text{InferenceModel}(x) & q \leftarrow 1 \\
\quad z \leftarrow \text{sample}(Q[\theta=x]) & q \leftarrow q \times Q[v=z, \theta=x] \\
\textbf{end} & \text{maximize: } \sum_z q \log(p/q)
\end{array}
$$

Pyro represents terms with discrete free variables as torch.Tensors. We extend this representation to other functions, encoding them as "tensors" where some of the "dimensions" have size "real":

$$
\begin{array}{rll}
\tau \in \text{Type} ::= & \mathbb{Z}_n & \text{"bounded integer"} \\
| & \mathbb{Z}_{n_1} \times \cdots \times \mathbb{Z}_{n_k} \to \mathbb{R} & \text{"real-valued array"}
\end{array}
$$

We define a set of specific terms that are closed under variable substitution, sum-product operations, and various approximations:

$$
\begin{array}{rll}
e \in \text{Funsor} ::= & \text{Tensor}(\Gamma, w) & \text{"discrete factor"} \\
| & \text{Gaussian}(\Gamma, i, P) & \text{"Gaussian factor"} \\
| & \text{Delta}(v, e) & \text{"point mass"} \\
| & \text{Variable}(v, \tau) & \text{"delayed value"} \\
| & \widehat{f}(e_1, \ldots, e_n) & \text{"apply function"} \\
| & e_1[v=e_2] & \text{"substitute"} \\
| & \sum_v e & \text{"marginalize"} \\
| & \prod_{v/s} e & \text{"Markov product"}
\end{array}
$$

We extend lazy tensor expressions to include **dimensions of size 'real'** and implement **semisymbolic integration**

On GitHub: pyro-ppl/funsor

Approximation and transformation via (re-)interepretation

Most integrals defined by Funsor terms cannot be computed directly. We rewrite lazy expressions by evaluating them with many different interpreters.

Some rules trigger PyTorch ops:

```
@eager.register (Binary, Op, Tensor, Tensor)
def eager_binary_tensor(op, lhs, rhs):
  inputs, (x, y) = align_tensors(lhs, rhs)
  data = op(x.data, y.data)
  return Tensor(data, inputs, lhs.dtype)
```

Some trigger further rewrites:

```
@eager.register(Binary, AddOp, Delta, Funsor)
def eager_add_delta(op, lhs, rhs):
  if lhs.name in rhs.inputs:
    rhs = rhs(**{lhs.name: lhs.point})
    return op(lhs, rhs)
  return None  # defer to default implementation
```

Some rewrite subexpressions into approximate versions. `monte_carlo` rewrites Tensor and Gaussian to Delta:

```
@dispatched_interpretation
def monte_carlo(cls, *args): ...

@monte_carlo.register(Integrate, Funsor, Funsor, set)
def monte_carlo_integrate(log_measure, integrand, vs):
  log_measure = log_measure.sample()
  return eager.dispatch(
    Integrate, log_measure, integrand, vs)
```

Funsor expressions are closed under these approximation rewrites.

Example: detecting EEG changepoints

We use a moment matching interpreter to fit a switching linear dynamical system:

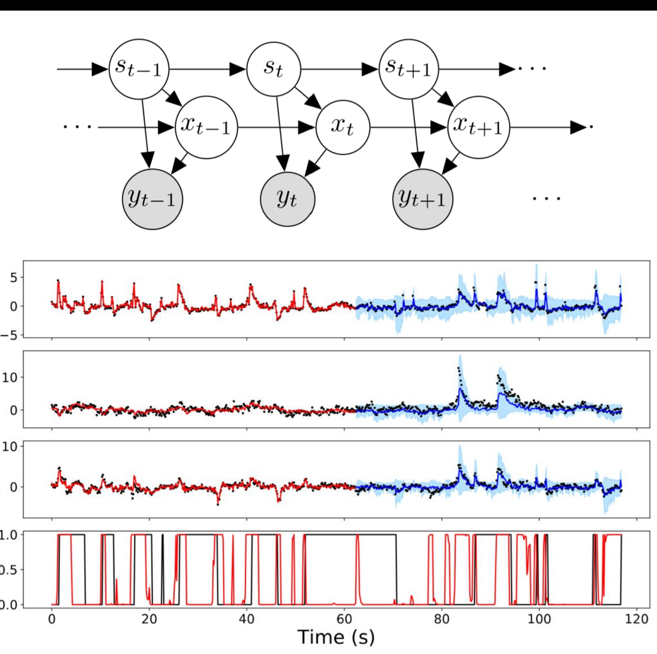
```
@interpretation(moment_matching_interpretation)
def marginal_log_prob(data, s_trans, x_trans, y_dist):
  ...
  for t, y in enumerate(data):
    ss[t] = Variable(f"s_{t}", bint(2))
    xs[t] = Variable(f"x_{t}", reals(5))
    ...
    log_prob += fdist.Categorical(
      s_trans(s=ss[t - 1]), value=ss[t])
    log_prob += x_trans(s=ss[t], x=xs[t - 1], y=xs[t])

    log_prob = log_prob.reduce(ops.logaddexp,
      {ss[t - 2].name, xs[t - 2].name})

    log_prob += y_dist(s=s_vars[t], x=x_vars[t], y=y)

  for t in range(2):
    log_prob = log_prob.reduce(ops.logaddexp,
      {s_vars[T - 2 + t].name, x_vars[T - 2 + t].name})

  return log_prob
```



Extending the language: parallel-scan over sequential structure

Many common sum-product expressions have linear chain structures.
We define a generic operation on atomic funsor terms for collapsing this structure:

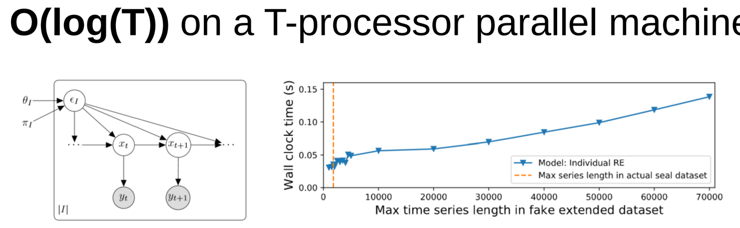
For Tensor terms, the MarkovProduct op corresponds to chain matrix multiplication:

$$\prod_{t/(i,j)} f = f[t=0] \bullet f[t=1] \bullet \cdots \bullet f[t=T-1]$$

i.e. the binary operation is a GEMM:

$$f \bullet g = \sum_k f[j=k] \times g[i=k]$$

Our parallel-scan algorithm computes this in **O(log(T))** on a T-processor parallel machine:

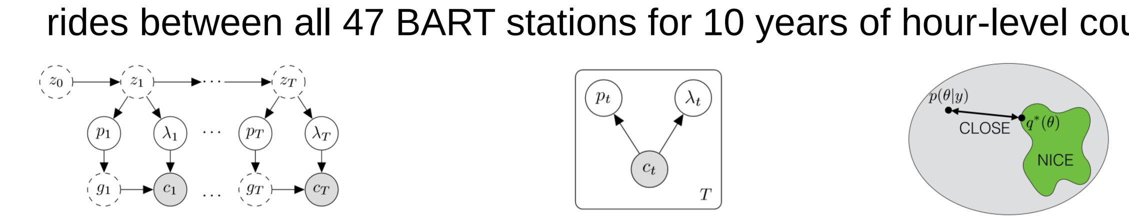


Algorithm 1 MarkovProduct

input a funsor f, a time variable $t \in \text{fv}(f)$,
 a step mapping $s \subseteq \text{fv}(f) \times \text{fv}(f)$.
output the Markov product funsor $\prod_{t/s} f$.
Create substitutions with fresh names (barred):
$s_e \leftarrow \{(y, \bar{x}) \mid (x, y) \in s\}$ to rename even factors, and
$s_o \leftarrow \{(x, \bar{x}) \mid (x, y) \in s\}$ to rename odd factors.
Let $v \leftarrow \{\bar{x} \mid (x, y) \in s\}$ be variables to marginalize.
Let $T \leftarrow |\Gamma_f[t]|$ be the length of the time axis.
while $T > 1$ **do**
 Split f into even and odd parts of equal length:
 $f_e \leftarrow f[s_e, t=(0, 2, 4, 6, ..., 2\lfloor T/2 \rfloor - 2)]$
 $f_o \leftarrow f[s_o, t=(1, 3, 5, 7, ..., 2\lfloor T/2 \rfloor - 1)]$
 Perform parallel sum-product contraction:
 $f' \leftarrow \sum_v f_e \times f_o$
 if T *is even* **then** $f \leftarrow f'$
 else $f \leftarrow \text{concat}_t(f', f[t=T-1])$;
 $T \leftarrow \lceil T/2 \rceil$
return $f[t=0]$

Example: forecasting BART ridership

We do collapsed variational inference in a neural Kalman filter to model rides between all 47 BART stations for 10 years of hour-level counts:




```
def model(features, counts):
  gr_t = funsor.Variable(
    "gr_t", ...)["time"]
  init, trans, obs, rate = ...
  ...
  prior = MarkovProduct(
    ops.logaddexp, ops.add,
    trans + obs(gate_rate=gr_t),
    "time", {"state": "state(time=1)"})

  llk = fdist.Poisson(rate["origin", "destin"])
  return prior + llk(value=counts)
```

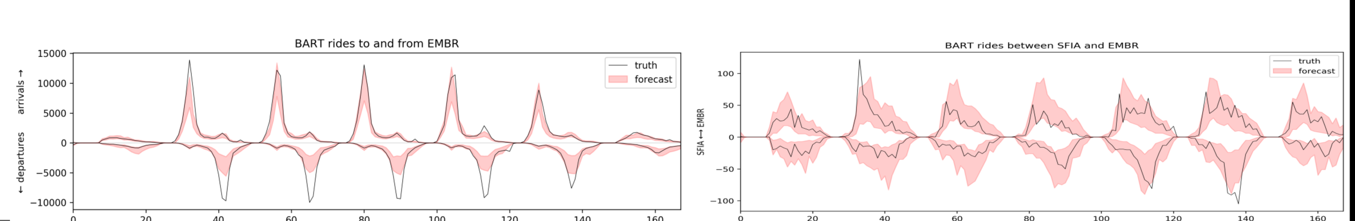
```
def guide(features, counts):
  loc, sc = ...
  diag_normal = fdist.Normal(
    loc, sc, value="gr_t")
  return diag_normal
```

```
def elbo_loss(features, counts):
  q = guide(features, counts)

  with interpretation(lazy):
    p = model(features, counts)
    pq = p - q

  with interpretation(monte_carlo):
    elbo = funsor.Integrate(q, pq)

  return elbo
```



Operational semantics: designing for performance and ease of implementation

The term language has an obvious default operational semantics backed by high-performance tensor libraries like JAX and PyTorch. This not only gives us differentiable, hardware-accelerated kernels, but provides multiple tracing JIT compilers that compile away all runtime pattern-matching overhead.

$\boxed{\text{Exact}}$

$$
\begin{array}{rcll}
\text{Delta}(v, e_1) \times e_2 & \Rightarrow & \text{Delta}(v, e_1) \times e_2[v \mapsto e_1] & \text{if } v \in \text{fv}(e_2) \\
\sum_v \text{Delta}(v, e) & \Rightarrow & 1 & \\
\sum_v \text{Delta}(v, e_1) \times e_2 & \Rightarrow & e_2 & \text{if } v \notin \text{fv}(e_2) \\
\text{Variable}((v{:}\mathbb{Z}_n)) & \Rightarrow & \text{Tensor}((v{:}\mathbb{Z}_n), \mathtt{arange}(n)) & \\
\text{Tensor}(\Gamma_1, w_1)[v \mapsto \text{Tensor}(\Gamma_2, w_2)] & \Rightarrow & \text{Tensor}((\Gamma_1 - (v{:}\tau)) \cup \Gamma_2, \mathtt{index}(w_1, v, w_2)) & \text{if } (v{:}\tau) \in \Gamma_1 \\
\widehat{f}(\text{Tensor}(\Gamma_1, w_1), \ldots, \text{Tensor}(\Gamma_n, w_n)) & \Rightarrow & \text{Tensor}(\cup_k \Gamma_k, f(w_1, \ldots, w_n)) & \\
\sum_v \text{Tensor}(\Gamma, \mathrm{w}) & \Rightarrow & \text{Tensor}(\Gamma\text{-}(\mathrm{v}{:}\tau), \mathtt{sum}(\mathrm{w}, \mathrm{v})) & \text{if } (v{:}\tau) \in \Gamma \\
\prod_v \text{Tensor}(\Gamma, \mathrm{w}) & \Rightarrow & \text{Tensor}(\Gamma\text{-}(\mathrm{v}{:}\tau), \mathtt{prod}(\mathrm{w}, \mathrm{v})) & \text{if } (v{:}\tau) \in \Gamma
\end{array}
$$

$\boxed{\text{Lazy}}$

$$
\begin{array}{rcll}
\text{Delta}(v_1, e_1)[v_2 \mapsto e_2] & \Rightarrow & \text{Delta}(v_1, e_1[v_2 \mapsto e_2]) & \text{if } v_1 \neq v_2 \text{ and } v_1 \notin \text{fv}(e_2) \\
\text{Variable}((v{:}\tau))[v \mapsto e] & \Rightarrow & e & \\
\text{Variable}((v{:}\tau))[v' \mapsto e] & \Rightarrow & \text{Variable}((v{:}\tau)) & \text{if } v \neq v' \\
e_1[v \mapsto e_2] & \Rightarrow & e_1 & v \notin \text{fv}(e_1) \\
\widehat{f}(e_1, \ldots, e_n)[v \mapsto e_0] & \Rightarrow & \widehat{f}(e_1[v \mapsto e_0], \ldots, e_n[v \mapsto e_0]) & \\
(\sum_{v_1} e_1)[v_2 \mapsto e_2] & \Rightarrow & \sum_{v_1} e_1[v_2 \mapsto e_2] & v_1 \notin \text{fv}(e_2) \text{ and } v_1, v_2 \text{ distinct} \\
(\prod_{v_1/s} e_1)[v_2 \mapsto e_2] & \Rightarrow & \prod_{v_1/s} e_1[v_2 \mapsto e_2] & v_1 \notin \text{fv}(e_2) \text{ and } v_1, s, v_2 \text{ all distinct}
\end{array}
$$

Gaussian terms are unnormalized block-structured multivariate Gaussian densities.

Closed under products and marginalization.

Numerically stable representation: information vector and precision matrix.

$$
\begin{array}{rcll}
\text{Gaussian}(\Gamma_1, i, \Lambda)[v \mapsto \text{Tensor}(\Gamma_2, w)] & \Rightarrow & \text{Gaussian}((\Gamma_1 - (v{:}\tau)) \cup \Gamma_2, i', \Lambda') & \text{if a } v_2 \neq v \in \Gamma_1 \text{ is real-valued} \\
\text{Gaussian}(\Gamma_1, i, \Lambda)[v \mapsto \text{Tensor}(\Gamma_2, w)] & \Rightarrow & \text{Tensor}((\Gamma_1 - (v{:}\tau)) \cup \Gamma_2, w') & \text{if only } v \in \Gamma_1 \text{ is real-valued} \\
\text{Gaussian}(\Gamma_1, i_1, \Lambda_1) \times \text{Gaussian}(\Gamma_2, i_2, \Lambda_2) & \Rightarrow & \text{Gaussian}(\Gamma_1 \cup \Gamma_2, i_1 + i_2, \Lambda_1 + \Lambda_2) & \\
\prod_v \text{Gaussian}(\Gamma, i, \Lambda) & \Rightarrow & \text{Gaussian}(\Gamma - (v{:}\mathbb{Z}_n), \mathtt{sum}(i, v), \mathtt{sum}(\Lambda, v)) & \text{if } v \text{ is a bounded integer variable} \\
\sum_v \text{Gaussian}(\Gamma, i, \Lambda) & \Rightarrow & \text{Tensor}(\Gamma_d, w) \times \text{Gaussian}(\Gamma - (v{:}\tau), i', \Lambda') & \text{if } v \in \Gamma \text{ is a real array variable} \\
\sum_v \text{Tensor}(\Gamma_1, w) \times \text{Gaussian}(\Gamma_2, i, \Lambda) & \Rightarrow & \text{Tensor}(\Gamma_1, w) \times \sum_v \text{Gaussian}(\Gamma_2, i, \Lambda) & \text{if } v \in \Gamma_2 \text{ is a real array variable}
\end{array}
$$

Operational semantics: term rewriting with a hierarchy of tagless final interpreters

Sometimes expressions are expensive to evaluate directly or need to be simplified for pattern-matching with existing rewrite rules.

We rewrite these expressions with interpreters that **preserve their exact semantics**.

Examples: **variable elimination algorithms** that rewrite N-ary sum-product expressions to sequences of binary sum-product expressions.

Observation: these interpreters are completely **generic in the sum and product operations**.

We can apply them to evaluate sum-product expressions with any commutative semiring, like max-sum, or even multiple semirings.

$\boxed{\text{Exact}}$

$$
\prod_{v/c} e \Rightarrow \textsc{MarkovProduct}(e, v, c) \qquad \text{if } v \text{ is a bounded integer variable and } c \subseteq (\text{fv}(e) - v) \times (\text{fv}(e) - v)
$$

$\boxed{\text{Optimize}}$

$$
\sum_V e \Rightarrow e \qquad \text{if } V \subseteq \text{fv}(e) \text{ is empty}
$$

$$
\sum_V e_1 \times \cdots \times e_n \Rightarrow \sum_{V'} \Big(\sum_{V_1 \cap V} e_1 \Big) \times \cdots \times \Big(\sum_{V_n \cap V} e_n \Big)
$$

where $V_k = \text{fv}(e_k) - \bigcup_{j \neq k} \text{fv}(e_j)$
and where $V' = V - \bigcup_k V_k$
if $V \subseteq \text{fv}(e_1 \times \cdots \times e_n)$
and if any $v \in V$ appear in only one e_k

$$
\sum_V e_1 \times \cdots \times e_n \Rightarrow \sum_{V - V'} \Big(\sum_{V' \cap V} e_{\sigma_1} \times e_{\sigma_2} \Big) \times e_{\sigma_3} \times \cdots \times e_{\sigma_n}
$$

where $\sigma_1, \ldots, \sigma_n$ is a min-cost path
where $V' = \text{fv}(e_{\sigma_1} \times e_{\sigma_2}) - \bigcup_{k>2} \text{fv}(e_{\sigma_k})$
if $V \subseteq \text{fv}(e_1 \times \cdots \times e_n)$
and if all $v \in V$ appear in ≥ 2 e_k

Operational semantics: closure under approximation

Some expressions have no equivalent form under exact semantics. The term language was carefully chosen to be closed under popular approximations which are interpreters that **preserve types but not semantics**. This allows us to evaluate all expressions semi-numerically (rather than symbolically).

`MomentMatching`: rewrite Gaussian mixture terms to single moment-matched multivariate Gaussian term.

Only specify one additional pattern on top of Exact.

$\boxed{\text{MomentMatching}}$

$$
\sum_v \text{Tensor}(\Gamma_1, w) \times \text{Gaussian}(\Gamma_2, i, \Lambda) \Rightarrow \text{Gaussian}(\Gamma_2 - (v{:}\tau), i', \Lambda') \times \sum_v \text{Tensor}(\Gamma_1, w') \qquad \text{if } v \text{ is a bounded integer variable}
$$

`MonteCarlo`: rewrite Gaussian and Tensor terms to Delta distributions over samples from those terms.

Infinitely differentiable by construction (via DiCE).

$\boxed{\text{MonteCarlo}}$

$$
\begin{array}{rcll}
\sum_v \text{Tensor}(\Gamma, w) \times e & \Rightarrow & \text{Tensor}(\Gamma - (v{:}\tau), w_N \times w_D) \times \sum_v \text{Delta}(\mathrm{v}, \mathrm{e_s}) \times e & \text{if } v \text{ is a bounded integer variable} \\
\sum_v \text{Gaussian}(\Gamma_d \cup (v{:}\tau), i, \Lambda) \times e & \Rightarrow & \text{Tensor}(\Gamma_d, w_N) \times \sum_v \text{Delta}(\mathrm{v}, \mathrm{e_s}) \times e & \text{if } v \text{ is a real array variable}
\end{array}
$$