On the Reparameterisation Gradient for Non-Differentiable but Continuous Models

UNIVERSITY OF OXFORD

Dominik Wagner and Luke Ong

Bayesian Inference

1. Exact Inference
2. Sampling-Based Methods: MCMC, HMC etc.
3. Variational Inference
   - frame posterior inference as (deterministic) *optimisation* problem

Variational Inference

Take: *variational family* $\{q_\theta \mid \theta \in \Theta\}$ of "simpler" *guide* distributions

Approach: find $\theta^* \in \Theta$ s.t. $q_\theta(\mathbf{z})$ is *"closest"* to $p(\mathbf{z} \mid \mathbf{x})$

KL divergence

or *maximise*

$$\mathrm{ELBO}(\theta) := \mathbb{E}_{\mathbf{z}\sim q_\theta}[\underbrace{\log p(\mathbf{z},\mathbf{x})}_{\text{model}} - \underbrace{\log q_\theta(\mathbf{z})}_{\text{guide}}]$$

*solve optimisation via **Stochastic Gradient Descent***

key ingredient!

Gradient Estimation (of Expectation)

- **Score Estimator:**
  widely applicable but *high variance*
- **Reparametrisation Estimator:**
  better in practice but *may be biased!* [Lee et al., NeurIPS 2018]

$$\nabla_\theta \mathbb{E}_{s\sim\mathcal{N}(0,1)}[[\theta + s \geq 0]] \neq \mathbb{E}_{s\sim\mathcal{N}(0,1)}\Big[\underbrace{\nabla_\theta [\theta + s \geq 0]}_{=0 \text{ a.e.}}\Big]$$

Contributions: *Study reparametrisation gradient estimator for continuous but possibly non-differentiable programs*

- categorical models
- prove *unbiasedness* in continuous setting
- *establish continuity* in languages with conditionals compositionally

Idealised Programming Language:

$$\begin{aligned} M ::= {} & x \mid \underline{r} \mid \underline{f} \mid \lambda x.\, M \mid M\, M \\ & \mid \mathbf{sample}\ \mathcal{D}(M,\dots,M) \mid \mathbf{observe}\ M\ \mathbf{from}\ \mathcal{D}(M,\dots,M) \\ & \mid \mathbf{if}\ M<0\ \mathbf{then}\ M\ \mathbf{else}\ M \mid \mathbf{sif}\ M<0\ \mathbf{then}\ M\ \mathbf{else}\ M \end{aligned}$$

smoothed conditional
[KOW23]

($r \in \mathbb{R}$, $f : \mathbb{R}^\ell \to \mathbb{R}$ and $\mathcal{D}$ is a continuous probability distribution)

Example (Temperature Regulation)

- Without intervention, the temperature fluctuates randomly.
- If the temperature drops below a threshold of 18°C the heating is engaged and the power is proportional to the deviation from the threshold.
- Time is discretised and after one time unit we measure a temperature of 21°C.
- We are interested in the distribution of the original temperature.

```
let t0 = sample normal(20,σ0)
    mu = t0 + if t0 < 18 then c * (18−t0)
                          else 0
    observe 21 from normal(mu, σ)
in t0
```

($\sigma_0, \sigma, c > 0$ are constants.)
NB The joint density is not differentiable (yet continuous) at $t_0 = 18$.

Denotational Weight Semantics

Example. $[\![\mathbf{sample}\,\mathcal{N}(\underline{20},\underline{\sigma_0}) + x \cdot (\mathbf{observe}\ 2\ \mathbf{from}\ \mathcal{N}(0,1))]\!]\,(x,[s])$
$= \mathrm{pdf}_{\mathcal{N}}(s \mid 20, \sigma_0) \cdot \mathrm{pdf}_{\mathcal{N}}(2 \mid 0,1)$

- denotational version of weight semantics
- beyond measuarability: capture piecewise definition and continuity
- complication: smoothed conditionals at higher-order [KOW23]

Solution: *Generalise Frölicher spaces, replacing smoothness with functions $\mathbb{R} \to \mathbb{R}$ with mild closure properties, enriched over* **Vect**.

1. Piecewise Analytic Functions under Analytic Partitions (*PAP*):
$$f(x) = \sum_{i=1}^{\ell} [x \in U_i] \cdot f_i(x)$$
($U_1,\dots,U_\ell \subseteq \mathbb{R}^n$ is a partition of analytic sets, each f_i is analytic) [LYRY20]
2. Continuous PAP (*CPAP*):
piecewise definitions agree on boundaries (for each $x \in \overline{U_i} \cap \overline{U_j}$, $f_i(x) = f_j(x)$)

Theorem (Unbiasedness)

If $f \circ \phi_\theta$ is a continuous PAP with partial derivatives which are uniformly dominated by an integrable function then
$$\nabla_\theta \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[f(\phi_\theta(\mathbf{z}))] = \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[\nabla_\theta f(\phi_\theta(\mathbf{z}))]$$
differentiable a.e.

Guarantee pre-conditions: densities are Schwartz functions, partial derivatives of ϕ_θ and primitives are bounded by polynomials

Interpret terms without conditionals (possibly with smoothed conditionals) in **VectCPAP** *and obtain unbiasedness.*

Example. Rephrase conditional via non-differentiable primitive:
$$c \cdot (\underline{\mathrm{ReLU}}\,(18 - t_0))$$

Continuity for Terms with Conditionals

$$\frac{\Gamma \vdash_{\text{cont}} L : R \quad \Gamma \vdash_{\text{cont}} M : \tau \quad \Gamma \vdash_{\text{cont}} N : \tau}{\Gamma \vdash_{\text{cont}} \mathbf{if}\ L<0\ \mathbf{then}\ M\ \mathbf{else}\ N : \tau} \quad \begin{array}{l}\forall \gamma \in [\![\Gamma]\!] .\, [\![L]\!]\,(\gamma) = 0 \\ \to [\![M]\!]\,(\gamma) = [\![N]\!]\,(\gamma)\end{array}$$

Problem: *generally not tractable!*
Solution: *restrict guards to affine terms*

- efficiently sample from boundary
- efficiently check guard's consistency (linear arithmetic solvers)

Example. For analytic primitives f and g,

$$\begin{array}{lll} & \mathbf{if}\ x-y<0\ \mathbf{then}\ \underline{f}\ x\ y\ \mathbf{else}\ \underline{g}\ x\ y & \text{is continuous} \\ \iff & (f-g)_{|U} = 0 & \text{where } U := \{(x,y) \mid x = y\} \\ \iff^{*} & f(x,y) = g(x,y) & (x,y) \sim U \end{array}$$

* *with probability 1*

Efficient Continuity Check

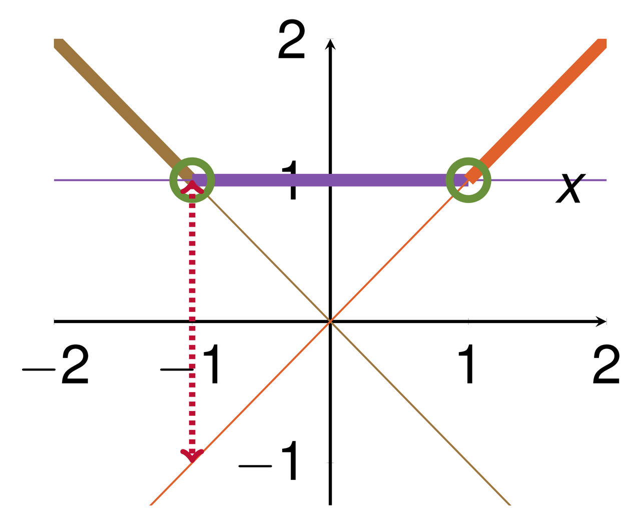
Suppose: $\mathbf{if}\ (\underline{\mathbf{a}}^T\mathbf{x} + \underline{c}) < 0\ \mathbf{then}\ F\ \mathbf{else}\ G$ w.l.o.g. $a_1 \neq 0$

Check: For all *consistent* branches S in F and T in G,
$$[\![S]\!]\,(x_1,\dots,x_n) = [\![T]\!]\,(x_1,\dots,x_n)$$
where $x_2,\dots,x_n \sim \mathcal{D}$ and $x_1 := -\frac{a_2x_2+\cdots+a_nx_n+c}{a_1}$.
(e.g. $\mathcal{N}(0,1)$)

Example (Consistency)

Implement $\max\{|x|,1\}$:

$\mathbf{if}\ x+1<0\ \mathbf{then}$
$\quad -x$
$\mathbf{else}$
$\quad (\mathbf{if}\ x-1<0\ \mathbf{then}\ 1\ \mathbf{else}\ x)$



Sufficient to check:

$[\![-x]\!]\,(-1) \stackrel{?}{=} [\![1]\!]\,(-1)$ ✓ (outer conditional)

$[\![1]\!]\,(1) \stackrel{?}{=} [\![x]\!]\,(1)$ ✓ (inner conditional)

$(x+1 \leq 0) \wedge (x-1 \geq 0)$ is inconsistent, so *no* need to check:

$[\![-x]\!]\,(-1) \stackrel{?}{=} [\![x]\!]\,(-1)$ ✗ (outer conditional)

References

[LYY18] Wonyeol Lee, Hangyeol Yu, and Hongseok Yang: Reparameterization gradient for non-differentiable models. NeurIPS 2018.

[LYRY20] Wonyeol Lee, Hangyeol Yu, Xavier Rival, and Hongseok Yang. On correctness of automatic differentiation for non-differentiable functions. NeurIPS 2020.

[KOW23] Basim Khajwal, C. H. Luke Ong, and Dominik Wagner. Fast and correct gradient-based optimisation for probabilistic programming via smoothing, ESOP 2023 *to appear*.

dominik.wagner@cs.ox.ac.uk