1 Analysis of Local Expectation Gradients (LEG) [1]

We provide an analysis of the LEG estimator including its computational requirements.

Proposition 1.1. (LEG Trick and Estimator) Given a factorised distribution $p(\mathbf{X}) = \prod_{d=1}^{D} p(X_d \mid \mathbf{X}_{<d})$, Local Expectation Gradients are based on the following equalities:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{X}\sim p(\mathbf{X})}\left[f(\mathbf{X})\partial_\lambda \log p(\mathbf{X})\right] = \\
&\quad= \sum_{d=1}^{D} \mathbb{E}_{\mathbf{X}_{\neq d}\sim p(\mathbf{X}_{\neq d})}\left\{\mathbb{E}_{X_d\sim p(X_d|\mathrm{MB}(X_d))}\left[f(\mathbf{X})\partial_\lambda \log p(X_d \mid \mathbf{X}_{<d})\right]\right\} \qquad \text{cf. Eq. 8 in [1]} \\
&\quad= \sum_{d=1}^{D} \underbrace{\mathbb{E}_{\mathbf{X}_{<d}\sim p(\mathbf{X}_{<d})}\mathbb{E}_{X'_d\sim p(X'_d|\mathbf{X}_{<d})}\mathbb{E}_{\mathbf{X}_{>d}\sim p(\mathbf{X}_{>d}|X'_d,\mathbf{X}_{<d})}}_{\text{Used for sampling pivots}}\left\{\underbrace{\mathbb{E}_{X_d\sim p(X_d|\mathbf{X}_{\neq d})}}_{\text{Used for weighted average}}\underbrace{\left[f(\mathbf{X})\partial_\lambda \log p(X_d \mid \mathbf{X}_{<d})\right]}_{\text{Used as score evaluation}}\right\}
\end{aligned} \tag{1.1}$$

where $\mathrm{MB}(X_d)$ is the Markov blanket of variable X_d, the red terms introduce auxiliary variables for each $d \in \{1, \ldots, D\}$ and $p(X_d \mid \mathrm{MB}(X_d)) = p(X_d \mid \mathbf{X}_{\neq d}) = \frac{p(X_d|\mathbf{X}_{<d})\prod_{j>d} p(X_j|\mathbf{X}_{<j})}{\sum_{x_\delta\in\Omega(X_d)} p(x_\delta|\mathbf{X}_{<d})\prod_{j>d} p(X_j|\mathbf{X}_{<j})}$ is a weighting distribution.

The Monte Carlo estimate corresponding to the LEG Trick in Eq. 1.1 (cf. Algorithm 1 in [1]) given N pivot samples is given by the following expression:

$$\mathbb{E}_{\mathbf{X}\sim p(\mathbf{X})}\left[f(\mathbf{X})\partial_\lambda \log p(\mathbf{X})\right] \approx \sum_{d=1}^{D}\frac{1}{N}\sum_{n=1}^{N}\sum_{x_\delta\in\Omega(X_d)} p(x_\delta \mid \mathbf{x}^{(n)}_{\neq d}) f(\mathbf{x}^{(n)}_{\neq d}, x_\delta)\partial_\lambda \log p(x_\delta \mid \mathbf{x}^{(n)}_{<d})$$

Additionally, the computational complexity is $O(N \cdot D^2 \cdot K)$, with $K = \max_d(\mid \Omega(X_d) \mid)$.

Proof. We start by recalling the proof for the first equality in Eq. 1.1.

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}\sim p(\mathbf{X})}\left[f(\mathbf{X})\partial_\lambda \log p(\mathbf{X})\right] &= \mathbb{E}_{\mathbf{X}\sim p(\mathbf{X})}\left[f(\mathbf{X})\partial_\lambda \log \prod_{d=1}^{D} p(X_d \mid \mathbf{X}_{<d})\right] \\
&= \sum_{d=1}^{D}\mathbb{E}_{\mathbf{X}\sim p(\mathbf{X})}\left[f(\mathbf{X})\partial_\lambda \log p(X_d \mid \mathbf{X}_{<d})\right] \\
&= \sum_{d=1}^{D}\mathbb{E}_{\mathbf{X}_{\neq d}\sim p(\mathbf{X}_{\neq d})}\left\{\underbrace{\mathbb{E}_{X_d\sim p(X_d|\mathbf{X}_{\neq d})}\left[f(\mathbf{X})\partial_\lambda \log p(X_d \mid \mathbf{X}_{<d})\right]}_{:=\, g(\mathbf{X}_{\neq d})}\right\}.
\end{aligned} \tag{1.2}$$

By noting that $p(X_d \mid \mathbf{X}_{\neq d}) = p(X_d \mid \mathrm{MB}(X_d))$, we obtain the first result.

Regarding the second equality in Eq. 1.1, we observe that $\mathbb{E}_{\mathbf{X}_{\neq d}\sim p(\mathbf{X}_{\neq d})}\left[g(\mathbf{X}_{\neq d})\right]$ in Eq. 1.2 can be equivalently rewritten in the following way:

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}_{\neq d}\sim p(\mathbf{X}_{\neq d})}\left[g(\mathbf{X}_{\neq d})\right] &= \mathbb{E}_{\mathbf{X}\sim p(\mathbf{X})}\left[g(\mathbf{X}_{\neq d})\right] \\
&= \mathbb{E}_{\mathbf{X}_{<d}\sim p(\mathbf{X}_{<d})}\mathbb{E}_{X'_d\sim p(X'_d|\mathbf{X}_{<d})}\mathbb{E}_{\mathbf{X}_{>d}\sim p(\mathbf{X}_{>d}|X'_d,\mathbf{X}_{<d})}\left[g(\mathbf{X}_{\neq d})\right]
\end{aligned} \tag{1.3}$$

By plugging this result into Eq. 1.2 we obtain the desired result for the second equality in Eq. 1.1.

The Monte Carlo estimate trivially follows by sampling N pivots from the first three expectations in Eq. 1.1 and by taking the exact weighted average of score evaluations.

Regarding computational complexity, we observe that computing the weighting function in the Monte Carlo estimate requires $O(D)$ evaluations. Also, the estimate requires to iterate over three summations, thus having a computational requirement of $O(N \cdot D \cdot K)$. By combining these two results, we obtain the overall complexity of $O(N \cdot D^2 \cdot K)$. □

References

[1] Titsias, Michalis and Lázaro-Gredilla, Miguel. Local Expectation Gradients for Black Box Variational Inference. *NeurIPS*, 2015.