Stein Variational Importance Sampling

Sky

1 Algorithm

According to section 5 of the paper, the authors use the RBF kernel $k(x, x') = \exp(-||x-x'||^2/h)$ with h being the kernel bandwidth. h is defined as $\text{med}^2/(2\log(|A|+1))$ where med is the median of the pairwise distances between the leader particles and $|A|$ is the number of leader particles. (Note: The authors took the median of the squared pairwise distances between the leader particles.)

1.1 Construct mapping for leaders

We first construct the map using the leader particles x_A^ℓ with ℓ representing the ℓth iteration. The pairwise distances $||x - x'||$ are already computed when defining the kernel.

Compute the kernel by expanding the numerator that is exponentiated:

$$||x_A - x'_A||^2 = x_A^T x_A - 2x_A^T x'_A + x'^T_A x'_A \tag{1}$$

Find the median pairwise distance by taking the square root of $||x_A - x'_A||^2$ and checking if it is odd or even. If it is odd, pick the median else take the mean of the two middle values. Square the median and divide by $2\log(|A|+1)$. With the above, we have the kernel of the leader particles.

Assume that we have $\nabla \log p(x_A^\ell)$ obtained using automatic differentiation. Automatic differentiation in Tensorflow is implemented such that it accumulates the variable we are differentiating by. Since the operation $\nabla \log p(x_A^\ell)$ only accumulates each particle once, it is safe to employ automatic differentiation here.

However, computing the Jacobian of the kernel $\nabla_{x_A} k(x_A, x'_A)$ leads to accumulation of each particle $|A|$ times so we have to do it by hand:

$$\nabla_{x_A} k(x_A, x'_A) = \nabla_{x_A} \exp(-(x_A^T x_A - 2x_A^T x'_A + x'^T_A x'_A)/h) \text{ with } h \text{ being constant} \tag{2}$$

$$= -\frac{2(x_A - x'_A)}{h} \exp(-||x_A - x'_A||^2/h) \tag{3}$$

Since the update $\phi_A^{\ell+1}$ consist of two terms being added together and we are performing a sum over them, the operations are commutative so we can split it up into adding the sum of one term to the sum of the other. Hence $\phi_A^{\ell+1}(\cdot)$ is equivalent to $\frac{1}{|A|}\{k(x_A, \cdot)^T \nabla \log p(x_A) + \sum_j -\frac{2(x_{A_j} - \cdot)}{h} \exp(-||x_{A_j} - \cdot||^2/h)\}$.

1.2 Construct mapping for followers

We do the same for the followers:

- $||x_A - x_B||^2 = x_A^T x_A - 2x_A^T x_B + x_B^T x_B$
- $\nabla_{x_A} k(x_A, x_B) = -\frac{2(x_A - x_B)}{h} \exp(-||x_A - x_B||^2/h)$
- $\phi_B^{\ell+1}(\cdot) = \frac{1}{|A|}\{k(x_A, \cdot)^T \nabla \log p(x_A) + \sum_j -\frac{2(x_{A_j} - \cdot)}{h} \exp(-||x_{A_j} - \cdot||^2/h)\}$

1.3 Update leaders and followers

Update ϕ_A and ϕ_B to both leader and follower particle by adding $\epsilon * \phi$ to them.

1.4 Calculate density values of followers

We now compute $\nabla_{x_{B_i}} \phi_B(x_{B_i}) = \frac{1}{A} \sum_A [\nabla_{x_A} \log p(x_{B_i})^T \nabla_{x_{B_i}} k(x_A, x_{B_i}) + \nabla_{x_A s} \nabla_{x_{B_i}} k(x_A, x_{B_i})]$. Like before but with a slight tweak, $\nabla_{x_{B_i}} k(x_A, x_{B_i})$ is:

$$\nabla_{x_{B_i}} \exp(-(x_A^T x_A - 2x_A^T x_{B_i} + x_{B_i}^T x_{B_i})/h) = \frac{2(x_A - x_{B_i})}{h} \exp(-||x_A - x_{B_i}||^2/h) \tag{4}$$

Shifting the focus to the $\nabla_{x_A} \nabla_{x_{B_i}} k(x_A, x_{B_i})$ term, we begin with the above result

$$\begin{aligned}
&\nabla_{x_A} \nabla_{x_{B_i}} k(x_A, x_{B_i}) && (5)\\
&= \nabla_{x_A} \frac{2(x_A - x_{B_i})}{h} \exp(-||x_A - x_{B_i}||^2/h) && (6)\\
&= \frac{2}{h}[\nabla_{x_A} x_A \exp(-||x_A - x_{B_i}||^2/h) + \nabla_{x_A} x_{B_i} \exp(-||x_A - x_{B_i}||^2/h)] && (7)\\
&= \frac{2}{h}[I - \frac{2(x_A - x_{B_i})}{h} x_A^T - \frac{2(x_A - x_{B_i})}{h} x_{B_i}^T] \exp(-||x_A - x_{B_i}||^2/h) && (8)\\
&= \frac{2}{h}[I - \frac{2}{h}(x_A - x_{B_i})(x_A - x_{B_i})^T] \exp(-||x_A - x_{B_i}||^2/h) && (9)
\end{aligned}$$

Hence, we have

$$\begin{aligned}
\nabla_{x_{B_i}} \phi_B(x_{B_i}) &= \frac{1}{|A|} \sum_{j \in A} [\nabla_{x_{A_j}} \log p(x_{A_j})^T \frac{2(x_{A_j} - x_{B_i})}{h} \exp(-||x_{A_j} - x_{B_i}||^2/h)) \\
&\quad + \frac{2}{h}(I - \frac{2}{h}(x_{A_j} - x_{B_i})(x_{A_j} - x_{B_i})^T) \exp(-||x_{A_j} - x_{B_i}||^2/h)] \qquad (10)\\
&= \frac{1}{|A|}[\nabla_{x_A} \log p(x_A)^T \cdot \mathrm{diag}(\exp(-||x_A - x_{B_i}||^2/h)) \cdot \frac{2(x_A - x_{B_i})}{h} \\
&\quad + \left(\frac{2}{h} \cdot \sum_{j \in A} I \cdot \exp(-||x_{A_j} - x_{B_i}||^2/h)\right) \\
&\quad - \frac{2(x_A - x_{B_i})}{h}^T \cdot \mathrm{diag}(\exp(-||x_A - x_{B_i}||^2/h)) \cdot \frac{2(x_A - x_{B_i})}{h}] \qquad (11)
\end{aligned}$$

1.5 Update density values of followers

With $\nabla_{x_{B_i}} \phi_B(x_{B_i})$ in hand, we update the density values using the last equation of the Stein IS algorithm by using Tensorflow to compute the absolute determinant of $I + \epsilon \nabla_{x_{B_i}} \phi_B(x_{B_i})$ and multiplying the current density values by the inverse of the result.