Stein Variational Gradient Descent (SVGD)

Un Algoritmo de Propósito General para Inferencia Bayesiana

Manuel Horn, Facundo Alvarez Motta

Noviembre 29, 2024

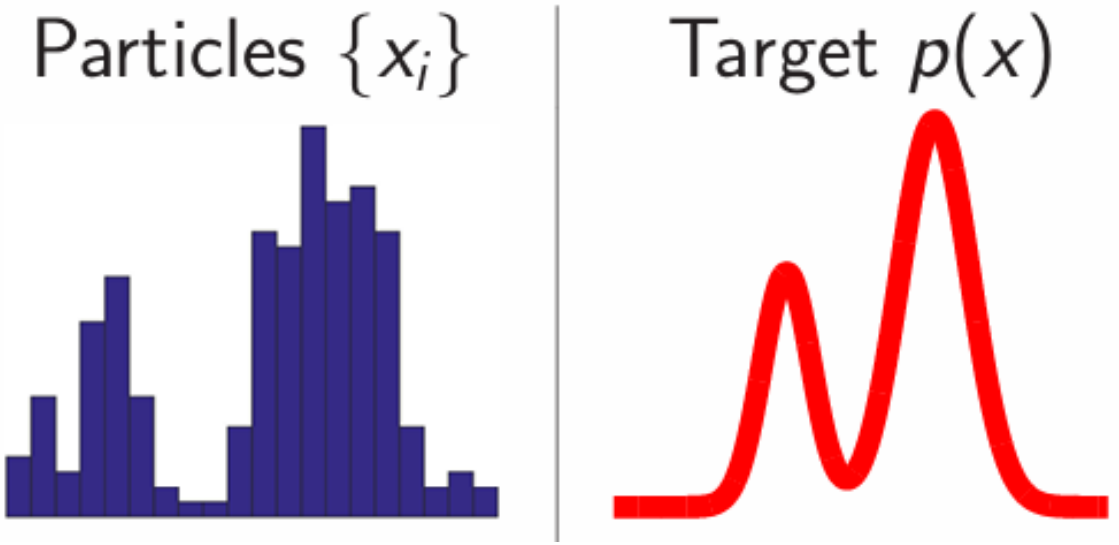
Introducción

Desafíos en la inferencia bayesiana escalable:

- **MCMC:** A menudo lento; difícil determinar la convergencia.
- **Inferencia Variacional:** Depende críticamente del conjunto de distribuciones definido para la aproximación.

Descenso por Gradiente Variacional de Stein (SVGD):

- Minimiza directamente $\mathrm{KL}(\{x_i\} \| p)$.
- No necesita definir una familia explícita de aproximación variacional.
- Aprovecha la información del gradiente.



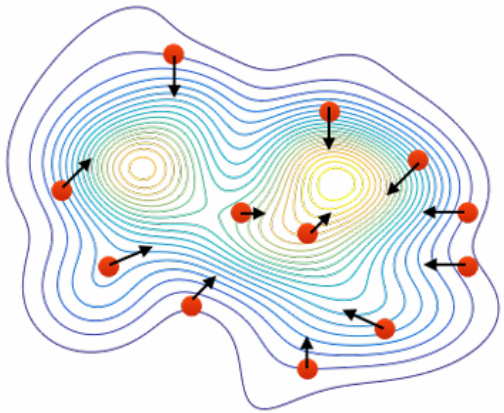
Idea principal

Idea: Mover iterativamente $\{x_i\}_{i=1}^n$ hacia la distribución objetivo $p(x)$ mediante actualizaciones del tipo:

$$x_i' \leftarrow x_i + \epsilon\phi(x_i), \tag{1}$$

donde ϕ es la dirección de perturbación elegida para disminuir al máximo la divergencia KL con $p(x)$:

$$\phi = \arg\max_{\phi \in \mathcal{F}} \left\{ -\frac{\partial}{\partial \epsilon} \mathrm{KL}(q_{[\epsilon\phi]} \| p) \Big|_{\epsilon=0} \right\}, \tag{2}$$



Aquí, $q_{[\epsilon\phi]}$ es la densidad de $x' = x + \epsilon\phi(x)$ y $\mathcal{F}$ es el conjunto de direcciones de perturbación sobre el que optimizamos.

Como encuentro el ϕ óptimo?

Stein Variational Gradient Descent

Resulta que el objetivo en (2) es una funcional lineal simple de ϕ:

$$-\frac{\partial}{\partial \epsilon}\mathrm{KL}(q_{[\epsilon\phi]}\|p)\Big|_{\epsilon=0} = \mathbb{E}_{x\sim q}[A_p\phi(x)]$$

donde

$$A_p\phi(x) \overset{\text{def}}{=} \phi(x)\nabla_x \log p(x)^\top + \nabla_x \cdot \phi(x).$$

Por lo tanto, la optimización en (2) se reduce a:

$$\mathcal{D}(q\|p) \overset{\text{def}}{=} \max_{\phi\in\mathcal{F}} \mathbb{E}_{x\sim q}[A_p\phi(x)]. \tag{3}$$

Identidad de Stein: $\mathbb{E}_{x\sim q}[A_p\phi(x)] = 0$ si y solo si $q = p$.

Stein Variational Gradient Descent (continuación)

Tomemos $\mathcal{F}$ como la bola unidad de un espacio de Hilbert de núcleos reproductores (RKHS) vectorial $\mathcal{H}$. Liu et al. (2016) mostraron que la solución óptima de (3) tiene una forma cerrada simple:

$$\phi^*(x') = \mathbb{E}_{x\sim q}\left[A_p k(x, x')\right] = \mathbb{E}_{x\sim q}\left[\nabla_x \log p(x) k(x, x') + \nabla_x k(x, x')\right]$$

Aproximando $\mathbb{E}_{x\sim q}$ mediante el promedio empírico de las partículas actuales $\{x_i\}_{i=1}^n$, la ecuación (1) se reduce a:

$$x_i \leftarrow x_i + \epsilon \hat{\mathbb{E}}_{x\sim\{x_i\}_{i=1}^n}\left[\nabla_x \log p(x) k(x, x_i) + \nabla_x k(x, x_i)\right]$$

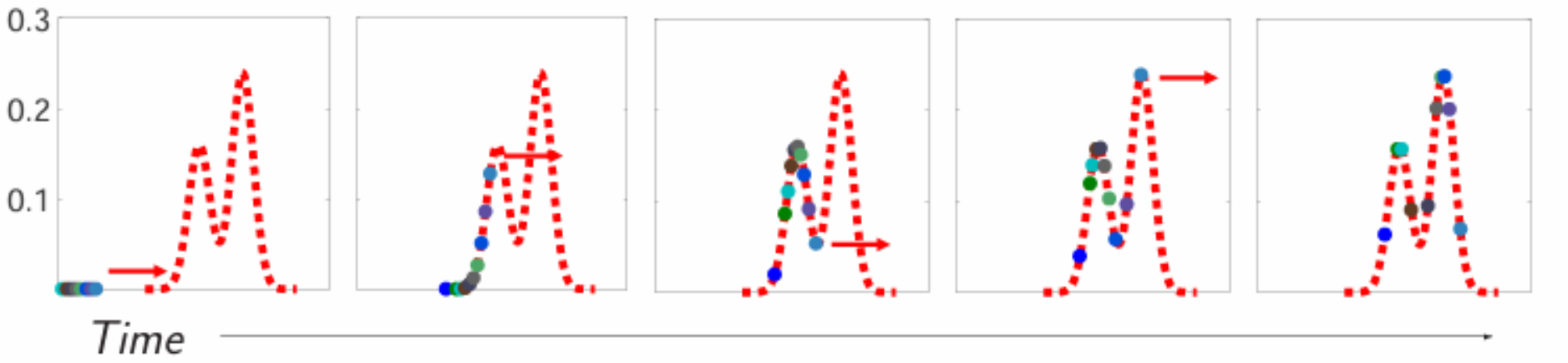
Algoritmo

Entrada: Una distribución objetivo con función de densidad $p(x)$ y un conjunto inicial de partículas $\{x_i^0\}_{i=1}^n$.
Salida: Un conjunto de partículas $\{x_i\}_{i=1}^n$ que aproxima la distribución objetivo.
Repetir:

$$x_i \leftarrow x_i + \epsilon \widehat{\mathbb{E}}_{x\sim\{x_i\}} \left[\underbrace{\nabla_x \log p(x)}_{\text{Gradiente}} k(x, x_i) + \underbrace{\nabla_x k(x, x_i)}_{\text{Fuerza Repulsiva}} \right], \quad \forall i = 1, \ldots, n.$$

- $\nabla_x \log p(x)$: mueve las partículas $\{x_i\}$ hacia regiones de alta probabilidad de $p(x)$.
- $\nabla_x k(x, x')$: impone diversidad en $\{x_i\}$ (evita que todas las partículas colapsen en los modos de $p(x)$).

Complejidad y Implementación Eficiente



En configuraciones de datos grandes, donde $p(x) \propto p_0(x) \prod_{k=1}^{N} p(D_k \mid x)$ con un N muy grande, aproximamos $\nabla_x \log p(x)$ usando mini-batches submuestreados:

$$\nabla_x \log p(x) \approx \nabla_x \log p_0(x) + \frac{N}{|\Omega|} \sum_{k \in \Omega} \nabla_x \log p(D_k \mid x)$$

Ejemplo: Ajuste de Mezcla de Gaussianas

- **Distribución objetivo:** $p(x) = \frac{1}{3}\mathcal{N}(-2, 1) + \frac{2}{3}\mathcal{N}(2, 1)$.
- **Inicialización:** 100 partículas con distribución $\mathcal{N}(-10, 1)$.

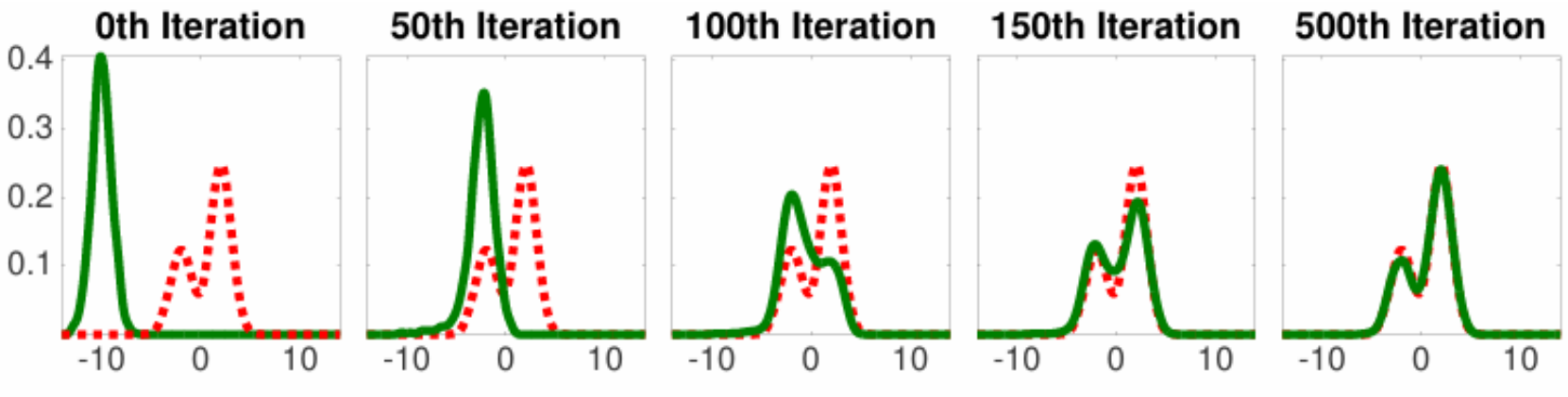


Figure: Evolución de partículas hacia la distribución objetivo.

Ejemplo unimodal

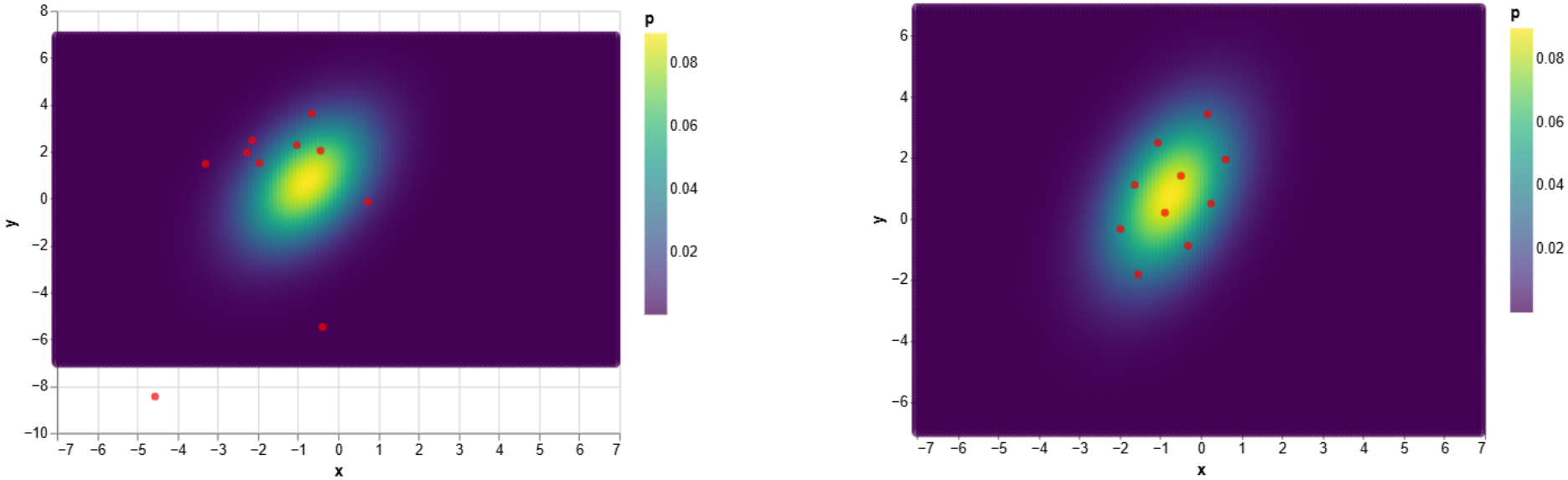


Figure: Simulación unimodal

Ejemplo multimodal

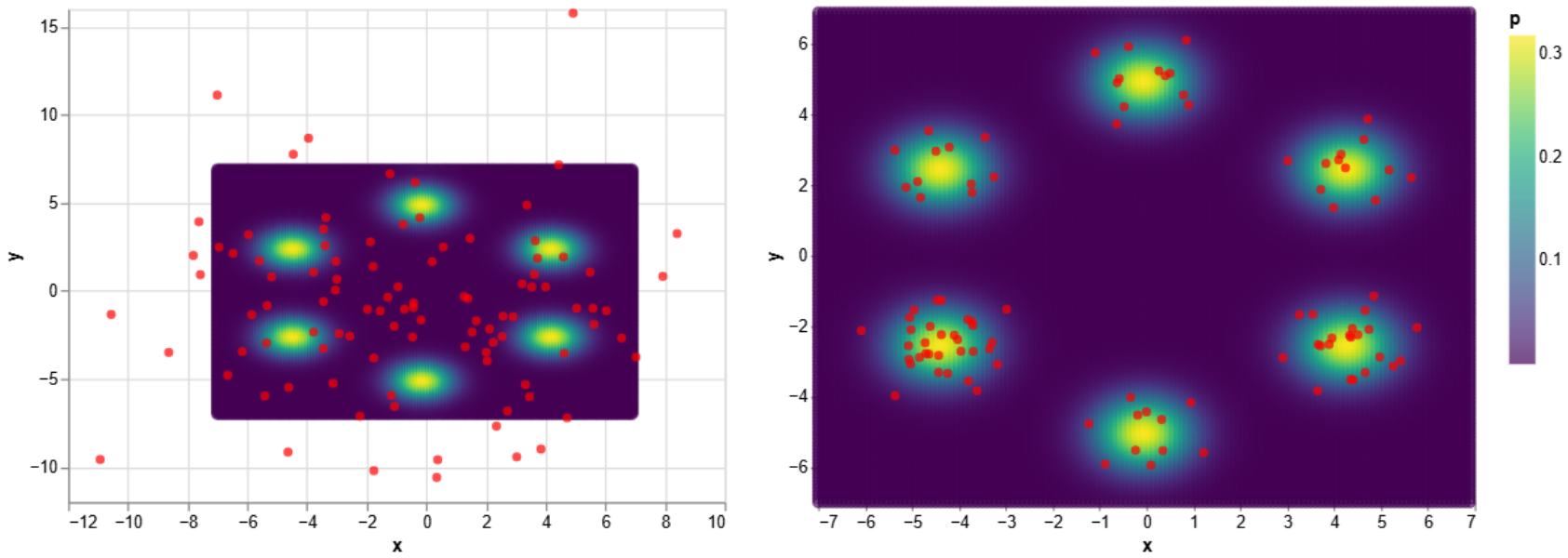


Figure: Simulación multimodal

Cierre y Referencias

- Qiang Liu, Dilin Wang. **Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm.** ICML, 2016. `https://arxiv.org/abs/1608.04471`
- **SVGD Paper Overview (Blog)**: `https://random-walks.org/book/papers/svgd/svgd.html`

¡Gracias! ¿Preguntas?