Stein Variational Gradient Descend

Justin Pauckert

Monte Carlo Methods in Machine Learning and Artificial Intelligence
TU Berlin

July 26, 2020

Overview

What...

- ... is Gradient Descend?
- ... are Kernels?
- ... is KL divergence?

What is Gradient Descend?

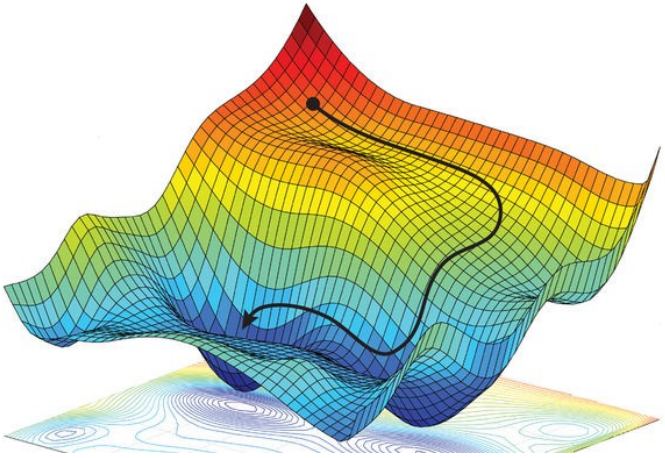
Let F be a real-valued function that is differentiable in a neighborhood of a point a. Ultimately, we want to minimize $F(x)$. To find such a (local) minimum, we choose

$$a_{n+1} = a_n - \epsilon \cdot \nabla F(a_n),$$

where $\epsilon \in \mathbb{R}$ is a small step size. Then, if ϵ is small enough, we have

$$F(a_{n+1}) \leq F(a_n).$$

What is Gradient Descend?



Source: https://www.sciencemag.org/news/2018/05/ai-researchers-allege-machine-learning-alchemy

What are Kernels?

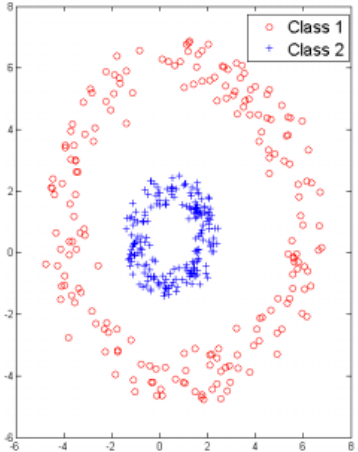
Let $X \subseteq \mathbb{R}^d$ be an input space. A function $k : X \times X \to \mathbb{R}$ is called a **kernel**, if there exists an inner product space $(F, \langle \cdot, \cdot \rangle)$ and a function $\varphi : X \to F$ which satisfies

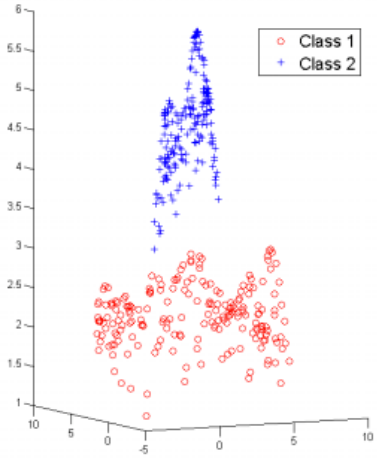
$$k(x, x') = \langle \varphi(x), \varphi(x') \rangle$$

for all $x, x' \in X$. This is used to find linear solutions for Problems in F which would not be linearily solvable in X.

Example: Gaussian RBF Kernel

$k(x, x') = \exp(-\frac{||x-x'||_2^2}{h})$ for some $h \in \mathbb{R}$.





What is KL divergence?

The Kullback-Leibler (KL) divergence is a measure of how different one continuous probability distribution is from another.

$$\mathrm{KL}(q||p) := \int_X p(x) \log \frac{q(x)}{p(x)} dx$$

$$= \int_{x \sim q} \log \frac{q(x)}{p(x)} dx = \mathbb{E}_q[\log\, q(x) - \log\, p(x)]$$

Problem Setup

Let X be a continuous random variable taking values in $\mathcal{X} \subseteq \mathbb{R}^d$, $\{D_k\}$ a set of N i.i.d. observations. Then with a prior distribution $p_0(x)$ we have

$$p(x|\{D_k\}) = \frac{p_o(x) \prod_{k=1}^{N} p(D_k|x)}{\int p_0(x) \prod_{k=1}^{N} p(D_k|x) dx}.$$

Stein's Identity

Let $p(x)$ be a smooth density on $\mathcal{X}$, $\phi(x) = [\phi_1(x), ..., \phi_d(x)]^T$ a smooth vector function. Then if ϕ is sufficiently regular we have

$$\mathbb{E}_p[\mathcal{A}_p\phi(x)] = 0, \tag{1}$$
$$\text{where } \mathcal{A}_p\phi(x) = \phi(x)\nabla_x \log p(x)^T + \nabla_x\phi(x).$$

$\mathcal{A}_p$ is called the **Stein Operator** and we say that ϕ is in the Stein class of p if (1) holds.

Stein Discrepancy

Consider another smooth density $q(x)$. Then we get a discrepancy measure via

$$\mathbb{S}(q,p) = \max_{\substack{\phi \in \mathcal{H}^d \\ ||\phi|| \leq 1}} \{\mathbb{E}_q[\mathsf{tr}(\mathcal{A}_p \phi(x))]^2\},$$

Stein Discrepancy

Consider another smooth density $q(x)$. Then we get a discrepancy measure via

$$\mathbb{S}(q,p) = \max_{\substack{\phi \in \mathcal{H}^d \\ ||\phi|| \leq 1}} \{\mathbb{E}_q[\mathsf{tr}(\mathcal{A}_p \phi(x))]^2\},$$

$\mathcal{H}^d$ being the reproducing kernel Hilbert space (RKHS) for some kernel k. The optimal solution has been shown to be

$$\phi^*_{q,p}(\cdot) = \mathbb{E}_q[\mathcal{A}_p k(x, \cdot)],$$

normalized to $\phi(x) = \phi^*_{q,p}(x)/||\phi^*_{q,p}||$.

The *variational* Part

Given a set of distibutions $\{\mathcal{Q}\}$, we want to approximate the target distribution $p(x)$ by minimizing the KL divergence.

$$q^* = \min_{q \in \mathcal{Q}} \{\mathsf{KL}(q||p)\}$$

The *variational* Part

Given a set of distibutions $\{\mathcal{Q}\}$, we want to approximate the target distribution $p(x)$ by minimizing the KL divergence.

$$q^* = \min_{q \in \mathcal{Q}} \{\mathsf{KL}(q||p)\}$$

Focus on sets $\mathcal{Q}$ consisting of distributions obtained by transforming a reference distribution. If $T : \mathcal{X} \to \mathcal{X}$ is a transformation, consider $x \sim q_0$ and $z = T(x)$ with density

$$q_{[T]}(z) = q(T^{-1}(z)) \cdot |\mathsf{det}(\nabla_x T^{-1}(z))|.$$

Stein Operator and KL Divergence

Let $T(x) = x + \epsilon f(x)$ be a transformation, $x \sim q(x)$ and $z = T(x)$ with density $q_{[T]}(z)$. Then

$$\nabla_\epsilon \mathsf{KL}(q_{[T]}||p)|_{\epsilon=0} = -\mathbb{E}_q[\mathsf{tr}(\mathcal{A}_p \phi(x))].$$

Stein Operator and KL Divergence

Let $T(x) = x + \epsilon f(x)$ be a transformation, $x \sim q(x)$ and $z = T(x)$ with density $q_{[T]}(z)$. Then

$$\nabla_\epsilon \mathsf{KL}(q_{[T]}||p)\big|_{\epsilon=0} = -\mathbb{E}_q[\mathsf{tr}(\mathcal{A}_p\phi(x))].$$

We know from earlier

$$\phi^*_{q,p} = \underset{\substack{\phi\in\mathcal{H}^d \\ ||\phi||\leq 1}}{\mathsf{argmax}}\{\mathbb{E}_q[\mathsf{tr}(\mathcal{A}_p\phi(x))]^2\}.$$

Stein Operator and KL Divergence

Let $T(x) = x + \epsilon f(x)$ be a transformation, $x \sim q(x)$ and $z = T(x)$ with density $q_{[T]}(z)$. Then

$$\nabla_\epsilon \mathsf{KL}(q_{[T]}||p)\big|_{\epsilon=0} = -\mathbb{E}_q[\mathsf{tr}(\mathcal{A}_p\phi(x))].$$

We know from earlier

$$\phi^*_{q,p} = \underset{\substack{\phi\in\mathcal{H}^d \\ ||\phi||\leq 1}}{\mathsf{argmax}}\{\mathbb{E}_q[\mathsf{tr}(\mathcal{A}_p\phi(x))]^2\}.$$

So for $T(x) = x + f(x)$, $f \in \mathcal{H}^d$ we have

$$\nabla_f \mathsf{KL}(q_{[T]}||p)\big|_{f=0} = -\phi^*_{q,p}(x).$$

The Algorithm

$T^*(x) = x + \epsilon \cdot \phi^*_{q,p}(x)$ is equal to a step of functional gradient descend in RKHS. Additionally, we can express $\phi^*_{q,p}$ with respect to our kernel k:

$$\phi^*_{q,p}(\cdot) = \mathbb{E}_q[k(x,\cdot)\nabla_x \log p(x) + \nabla_x k(x,\cdot)]$$

The Algorithm

$T^*(x) = x + \epsilon \cdot \phi^*_{q,p}(x)$ is equal to a step of functional gradient descend in RKHS. Additionally, we can express $\phi^*_{q,p}$ with respect to our kernel k:

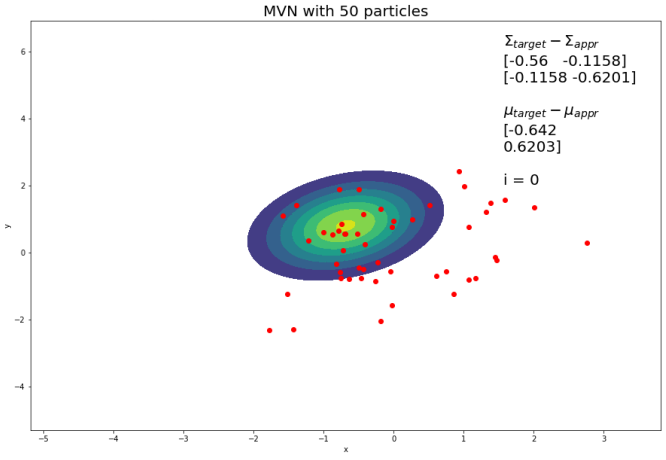
$$\phi^*_{q,p}(\cdot) = \mathbb{E}_q[k(x,\cdot)\nabla_x \log p(x) + \nabla_x k(x,\cdot)]$$

Algorithm Stein Variational Gradient Descend

Input: Target distribution density $p(x)$, set of initial particles $\{x_i\}_{i=1}^n$.
Output: Set of particles $\{x_i\}_{i=1}^n$ that approximate the target distribution.

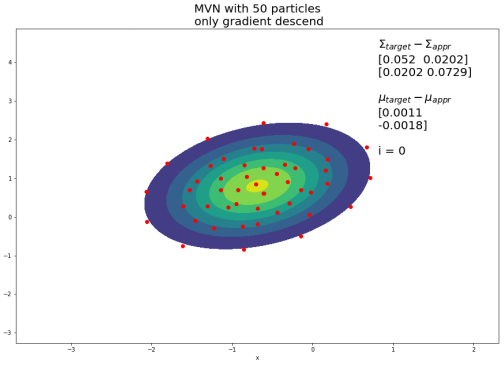
1: **for** $\ell = 1$ to *max_iterations* **do**
2: **for** $i = 1$ to n **do**
3: $x_i^{\ell+1} = x_i^\ell + \epsilon \cdot \hat{\phi}^*(x_i^\ell)$,
where $\hat{\phi}^*(x) = \frac{1}{n}\sum_{j=1}^{n}(k(x_j^\ell, x)\nabla_{x_j^\ell} \log p(x_j^\ell) + \nabla_{x_j^\ell} k(x_j^\ell, x))$
4: **end for**
5: **end for**

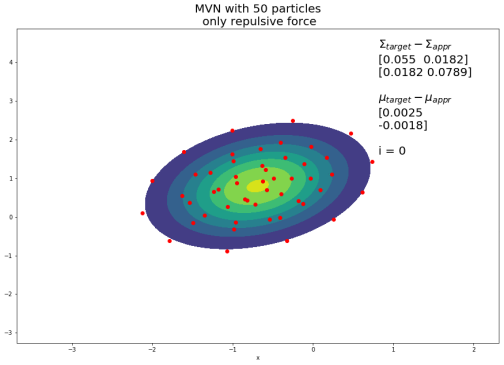
Examples


MVN with 50 particles
$\Sigma_{target} - \Sigma_{appr}$
[-0.56 -0.1158]
[-0.1158 -0.6201]
$\mu_{target} - \mu_{appr}$
[-0.642
0.6203]
i = 0
x
y

Examples

$$\hat{\phi}^*(x) = \frac{1}{n} \sum_{j=1}^{n} \left(k(x_j^\ell, x) \nabla_{x_j^\ell} \log p(x_j^\ell) + \nabla_{x_j^\ell} k(x_j^\ell, x) \right)$$





The Algorithm

Because the update only depends on $\nabla \log p(x)$ and

$$\nabla \log \frac{p(x)}{Z} = \nabla(\log p(x) - \log Z) = \nabla \log p(x),$$

we can completely ignore the normalization constant. Suddenly,

$$p(x|\{D_k\}) = p_o(x) \prod_{k=1}^{N} p(D_k|x)$$

is good enough as an estimator.

References

Liu, Qiang and Wang, Dilin (2016)

Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm

arXiv 1608.04471

github.com/DartML/Stein-Variational-Gradient-Descend