Stein Variational Gradient Descent

Boshu, Saeed and Clemens

March 2, 2021

Table of Content

Prerequisites

Gradient Descent

- **First Glance:** start with an initial condition, measure and iterate with the measurement
- **Objective:** minimize cost function $J(\theta)$

Pseudo code:

- Choose an initial parameters of weight θ and learning rate η
- Repeat until an approximate minimum of cost function is obtained:

$$\theta = \theta - \eta \nabla_{\theta} J$$

Motivation

What is the probability of a hypothesis given the observation? A Bayesian model presents the method to approximate the probability or even, the probability of probability.

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \tag{1}$$

Challenge

However, it quickly involves intractable computations, once the feature dimensions or latent variables rise (depending on problem):

$$p(y_1, ..., y_N|x) = \frac{p(x|y_1, ..., y_N)p(y_1, ..., y_N)}{p(x)} \tag{2}$$

$$p(y_i|x) = \int\int ... \int dy_1 ... dy_{i-1} dy_{i+1} ... dy_n p(y_1, ..., y_N|x) \tag{3}$$

Variational Inference

Variational Inference

- **Idea:** Choose closest approximate distribution through optimization given a statistical distance.
- **Pros:** Low variance; Suitable for big data; Fast.
- **Cons:** Accuracy depends on posterior assumptions (potential bias).

Hilbert space

Definition

A Hilbert space $\mathcal{H}$ is an inner product space that is also a complete metric space (contains Cauchy sequence limits).

Kernel

Definition

Let $\mathcal{X}$ be a non-empty set. $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a **kernel** if there exist an $\mathcal{H}$-Hilbert space and a map $\phi : \mathcal{X} \to \mathcal{H}$ such that $\forall x, x' \in \mathcal{X}$,

$$k(x, x') := \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}. \tag{4}$$

Reproducing Kernel

Definition

$k(\cdot,\cdot)$ is called a reproducing kernel of a Hilbert space $\mathcal{H}$, if $\forall f \in \mathcal{H},\ \langle k(x,\cdot), f(\cdot)\rangle = f(x)$.

Stein's identity

Definition

Let $p(x)$ be a smooth density on $\mathcal{X} \subseteq \mathbb{R}^d$, and $\phi(x) = [\phi_1(x), ..., \phi_d(x)]^T$ a smooth and sufficiently regular vector function, then we can write:

$$\mathcal{A}_p \boldsymbol{\phi}(x) := \boldsymbol{\phi}(x) \nabla_x log\ p(x)^T + \nabla_x \boldsymbol{\phi}(x) \tag{5}$$

where $\mathcal{A}_p$ is called the **Stein operator**, which acts on function ϕ and yields the following identity, known as **Stein's identity**:

$$\mathbb{E}_{x \sim p}[\mathcal{A}_p \boldsymbol{\phi}(x)] = 0 \tag{6}$$

Stein discrepancy

Definition

Let $q(x)$ be a different (from $p(x)$) smooth density $\mathcal{X}$. Given Stein operator $\mathcal{A}_p$ (defined for $p(x)$, we do not expect $\mathbb{E}_{x\sim q}[\mathcal{A}_p\boldsymbol{\phi}(x)]$ to be zero, but we can show that its value relates to the difference between p and q. Hence **Stein Discrepancy**, $\mathbb{S}$, given $\boldsymbol{\phi}$ in some proper function set $\mathcal{F}$ between two smooth densities p and q, is defined as maximum violation of Stein's identity:

$$\mathbb{S}(q,p) = \max_{\boldsymbol{\phi}\in\boldsymbol{\mathcal{F}}}\{[\mathbb{E}_{x\sim q} trace(\mathcal{A}_p\boldsymbol{\phi}(x))]^2\} \tag{7}$$

Kernelized Stein discrepancy

- Bounded Lipschitz norms: discriminative but intractable
- Kernelized Stein discrepancy: discriminative and tractable

Variational Inference Using Smooth Transformations

Key concept of Variational Inference

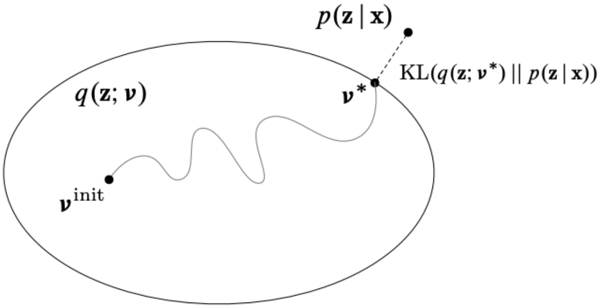

Figure: VI turns inference into optimization: Define **variational family** of distributions. Fit variational parameters to be **close (in KL) to true posterior**. [Blei et al. 2016]

Kullback-Leibler Divergence, D_{KL} (relative entropy)

Definition

Given two distributions $p(x)$ and $q(x)$ defined on the same probability space $\mathcal{X}$, $D_{KL}(q \parallel p)$ is:

$$D_{KL}(q \parallel p) := \mathbb{E}_q[log\frac{q(x)}{p(x)}] = \int_{x\in\mathcal{X}} q(x)\ log\frac{q(x)}{p(x)}dx \tag{8}$$

Motivation

D_{KL} is the mean extra information necessary to encode observations from source/target $p(x)$, using a code optimized for $q(x)$.

KL Divergence as objective function

The optimisation is insensitive to multiplicative coefficient, which is why we can use the unnormalised distribution for approximation:

$$\begin{aligned} KL(q \parallel Zp) &= \mathbb{E}_q[log \frac{q(x)}{Zp(x)}] \\ &= \mathbb{E}_q[log\ q(x)] - \mathbb{E}_q[log\ Zp(x)] \\ &= \mathbb{E}_q[log\ q(x)] - \mathbb{E}_q[log\ p(x)] - logZ \end{aligned} \tag{9}$$

$$q^* = \arg\min_{q \in Q} KL(q \parallel p) = \arg\min_{q \in Q} KL(q \parallel Zp) \tag{10}$$

The ideal function set

Figure: Traditional VI: Approximation using families. The ideal set is accurate (=approximates well), tractable (=consisting of simple distributions) and efficiently solvable by minimisation of KL. [Rocca 2019]

Particle Flows I

- Generate samples (particles) from tractable reference distribution $x_i \sim q_0(\cdot)$
- Deterministic particle flow over time

$$\frac{dz_i(t)}{dt} := \phi_t(z_i(t))$$

- Find mapping $\phi_t(\cdot)$ such that for $t \to \infty$ density of particles $\boldsymbol{z}(t) \sim q_\infty(\cdot)$ is close (in KL) to true posterior.

Particle Flows II

- Let $\boldsymbol{z}(t) = \boldsymbol{T}(\boldsymbol{x})$ be the transformed particles at time t where $\boldsymbol{T} : \mathcal{X} \rightarrow \mathcal{X}$ is a smooth one-to-one transformation.
- Step-wise construction of overall transformation $\boldsymbol{T}$: $\boldsymbol{T}_{t+1}(\boldsymbol{x}) = \boldsymbol{z}(t+1) = \boldsymbol{T}_t + \epsilon \phi_t(\boldsymbol{x})$
- Approximate density $q_{[\boldsymbol{T}]}(z)$ then given by change of variables:

$$q_{[\boldsymbol{T}]}(z) = q(\boldsymbol{T}^{-1}(z)) \cdot |det(\nabla_z \boldsymbol{T}^{-1}(z))|$$

Stein Variational Gradient Descent

Objective

Objective function: To minimize $D_{KL}(q \parallel p)$; Where $p(x)$ is the target and $q(x)$ is an approximate distribution.

Theorem (3.1.)

When $x \sim q(x)$, and $z = \boldsymbol{T}(x)$ so $q_{[\boldsymbol{T}]}(z)$ the density under smooth transformation $\boldsymbol{T}$; If $\boldsymbol{T}(x) = x + \epsilon\boldsymbol{\phi}(x)$ we have:

$$\nabla_\epsilon KL(q_{[\boldsymbol{T}]} \parallel p)|_{\epsilon=0} = -\mathbb{E}_{x\sim q}[trace(\mathcal{A}_p\boldsymbol{\phi}(x))], \tag{11}$$

where $\mathcal{A}_p\boldsymbol{\phi}(x) := \nabla_x log\ p(x)\boldsymbol{\phi}(x)^T + \nabla_x\boldsymbol{\phi}(x)$ is the Stein operator.

Steepest descent

Lemma (3.2)

Considering all the perturbation directions $\phi \in \boldsymbol{\mathcal{H}}^d$*, the direction of steepest descent,* $\phi^*_{q,p}$ *that minimizes* $\mathbb{S}(q,p) = -\nabla_\epsilon KL(q_{[\boldsymbol{T}]} \parallel p)|_{\epsilon=0}$*, is:*

$$\phi^*_{q,p}(\cdot) = \mathbb{E}_{x\sim q}[k(x,\cdot)\nabla_x log\ p(x) + \nabla_x k(x,\cdot)] \tag{12}$$

Optimal Transformation and Stepping in RKHS

Theorem (3.3)

Let $\boldsymbol{T}(x) = x + \boldsymbol{f}(x)$, where $\boldsymbol{f} \in \mathcal{H}^d$, and $q_{[\boldsymbol{T}]}(z)$ the density of $z = \boldsymbol{T}(x)$ when $x \sim q$,

$$\nabla_{\boldsymbol{f}} KL(q_{[\boldsymbol{T}]} \parallel p)|_{\boldsymbol{f}=0} = -\boldsymbol{\phi}^*_{q,p}(x), \tag{13}$$

whose squared RKHS norm is $\parallel \boldsymbol{\phi}^*_{q,p}(x) \parallel^2_{\mathcal{H}^d} = \mathbb{S}(q,p)$.

Another important result of Theorem 3.3

Complexity and Efficient Implementation

Theorem 3.3. suggests that $\boldsymbol{T}^*(x) = x + \epsilon \cdot \boldsymbol{\phi}^*_{q,p}$ is equivalent to step of gradient descent in RKHS. Therefore in every iteration of our process, we only need to evaluate the functional gradient for $\epsilon \cdot \boldsymbol{\phi}^*_{q,p}$ leading to identity map $\boldsymbol{T}(x) = x$; Hence saving us the expensive inverse Jacobian ($[\nabla_x \boldsymbol{T}(x)]^{-1}$) computation.

Ingredients

- sufficiently small ϵ
- a strictly positive definite kernel with decaying property
- initial distribution q_0 (insignificant in the process)
- initial particles $\{x_i^0\}_{i=1}^n$ (drawn from q_0 or arbitrary)
- target density function $p(x)$

Algorithm

Algorithm 1: Stein Variational Gradient Descent

Inputs : set of initial particles $\{x_i^0\}_{i=1}^n$, target distribution $p(x)$
Output: set of particles $\{x_i\}_{i=1}^n$ approximation $p(x)$
for *iteration* l **do**

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

$$x_i^{l+1} \leftarrow x_j^l + \epsilon_l \hat{\phi}^*(x_i^l)$$

end

Postface

Thank you for your attention

Questions? Thoughts? Advice?

References

- Liu, Qiang; Wang, Dilin (2016) "Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm", NIPS 2016
https://arxiv.org/pdf/1608.04471.pdf
- Blei, David; Ranganath, Rajesh; Mohamed, Shakir (2016) "Variational Inference: Foundations and Modern Methods", NIPS 2016 Tutorial
https://media.nips.cc/Conferences/2016/Slides/6199-Slides.pdf
- Rocca, Joseph (2019) "Bayesian inference problem, MCMC and variational inference", Towards Data Science Blog
https://towardsdatascience.com/bayesian-inference-problem-mcmc-and-variational-inference-25a8aa9bce29

MCMC vs. VI

Monte Carlo Markov Chain

- **Idea:** Setup Markov Chain with stationary distribution. Simulate random state sequence, keep some to compute statistics, etc.
- **Pros:** Asymptotically correct; No model assumptions.
- **Cons:** High variance; No strict convergence; Slow.

Variational Inference

- **Idea:** Choose closest approximate distribution through optimization given a statistical distance.
- **Pros:** Low variance; Suitable for big data; Fast.
- **Cons:** Accuracy depends on posterior assumptions (potential bias).

Hilbert Space

Definition

A Hilbert space $\mathcal{H}$ is a inner product space that is also a complete metric space (contains Cauchy sequence limits).

Hilbert Space

Definition

A Hilbert space $\mathcal{H}$ is a inner product space that is also a complete metric space (contains Cauchy sequence limits).

Definition (inner product)

$\langle\cdot,\cdot\rangle_{\mathcal{S}} : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ is an inner product on vector space $\mathcal{S}$ if is:

- bilinear: $\langle ax + by, z\rangle = a\langle x, z\rangle + b\langle y, z\rangle$
- conjugate symmetric: $\langle x, y\rangle = \overline{\langle y, x\rangle}$
- positive semi-definite: $\langle\cdot,\cdot\rangle \geq 0$ and $\langle x, x\rangle = 0 \iff x = 0$

Kernel

Definition

Let $\mathcal{X}$ be a non-empty set. $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a **kernel** if there exist an $\mathcal{H}$-Hilbert space and a map $\phi : \mathcal{X} \to \mathcal{H}$ such that $\forall x, x' \in \mathcal{X}$,

$$k(x, x') := \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}. \tag{14}$$

Theorem

Let $\mathcal{X}$, $\mathcal{H}$ and ϕ be defined as above. Then one can prove $\langle \phi(x), \phi(x') \rangle_{\mathcal{H}} =: k(x, x')$ is positive definite. Which is to say:

$$\forall (a_1, ..., a_n) \in \mathbb{R}^n,\ \forall (x_1, ..., x_n) \in \mathcal{X}^n : \ \sum_{j,i} a_i a_j k(x_i, x_j) \geq 0 \tag{15}$$

Kernel properties

1. $k(x, x) \geq 0$
2. $k(x, x')^2 \leq k(x, x)k(x', x')$
3. $k_1(x, x') + k_2(x, x') = k(x, x')$
4. $ak_1(x, x') = k(x, x')$ *where* $a > 0$
5. $k_1(x, x') \cdot k_2(x, x') = k(x, x')$
6. $f(x) \cdot f(y) = k(x, x')$ for any function f on x
7. Let $\mathcal{X}$ and $\tilde{\mathcal{X}}$ be sets, and define a map $A : \mathcal{X} \to \tilde{\mathcal{X}}$. The kernel k on $\tilde{\mathcal{X}}$, $k(A(x), A(x'))$ is a kernel on $\mathcal{X}$.

Properties of KL Divergence

1. is almost positive definite:
   - $D_{KL}(q \parallel p) \geq 0$ (non-negative)
   - $D_{KL}(q \parallel p) = 0 \iff q(x) = p(x)$ (identity of indispensables)

Properties of KL Divergence

1. is almost positive definite:
   - $D_{KL}(q \parallel p) \geq 0$ (non-negative)
   - $D_{KL}(q \parallel p) = 0 \iff q(x) = p(x)$ (identity of indispensables)
2. is not a proper metric
   - $D_{KL}(q \parallel p) \neq D_{KL}(p \parallel q)$ (asymmetric)
   - does not satisfy triangle inequality

Properties of KL Divergence

1. is almost positive definite:
   - $D_{KL}(q \parallel p) \geq 0$ (non-negative)
   - $D_{KL}(q \parallel p) = 0 \iff q(x) = p(x)$ (identity of indispensables)
2. is not a proper metric
   - $D_{KL}(q \parallel p) \neq D_{KL}(p \parallel q)$ (asymmetric)
   - does not satisfy triangle inequality
3. is convex

Properties of KL Divergence

1. is almost positive definite:
   - $D_{KL}(q \parallel p) \geq 0$ (non-negative)
   - $D_{KL}(q \parallel p) = 0 \iff q(x) = p(x)$ (identity of indispensables)
2. is not a proper metric
   - $D_{KL}(q \parallel p) \neq D_{KL}(p \parallel q)$ (asymmetric)
   - does not satisfy triangle inequality
3. is convex
4. is invariant under parameter transformations

Properties of KL Divergence

1. is almost positive definite:
   - $D_{KL}(q \parallel p) \geq 0$ (non-negative)
   - $D_{KL}(q \parallel p) = 0 \iff q(x) = p(x)$ (identity of indispensables)
2. is not a proper metric
   - $D_{KL}(q \parallel p) \neq D_{KL}(p \parallel q)$ (asymmetric)
   - does not satisfy triangle inequality
3. is convex
4. is invariant under parameter transformations
5. is additive for independent distributions

Connecting KL and Stein operator - Proof

Gradient of $\nabla_\epsilon KL(q_{[T]}||p)$ at $\epsilon = 0$ is equivalent to $\frac{dKL(q_t||p)}{dt}$, which can be rewritten as:

$$\frac{d}{dt}\int q_t(x) log\ q_t(x)dx - \frac{d}{dt}\int q_t(x) log\ p(x)dx =$$
$$\int \nabla \cdot (q_t(x)\phi_t(x)) log\ q_t(x)dx - \int \nabla \cdot (q_t(x)\phi_t(x)) log\ p(x)dx =$$
$$-\int q_t(x)\phi_t(x) \cdot \nabla log\ q_t(x)dx + \int q_t(x)\phi_t(x) \cdot \nabla log\ p(x)dx =$$
$$-\int \nabla q_t(x)\phi_t(x)dx + \int q_t(x)\phi_t(x) \cdot \nabla log\ p(x)dx =$$
$$\mathbb{E}_q[\nabla \cdot \phi_t(x) + \phi_t(x) \cdot \nabla log\ p(x)] \tag{16}$$