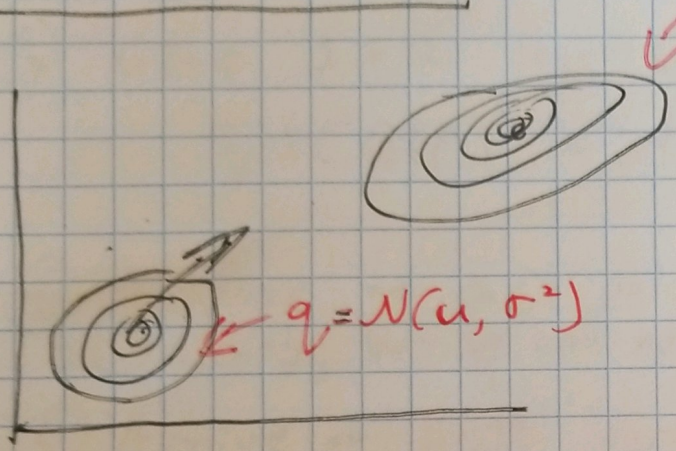
Stein Variational Gradient Descent

Qiang Liu & Dilan Wang, 2016

- Also used:
- "A short introduction to Kernalized Stein Discrepancy"
  - Qiang Liu
- "Functional Gradient Descent"
  - Drew Bagnell

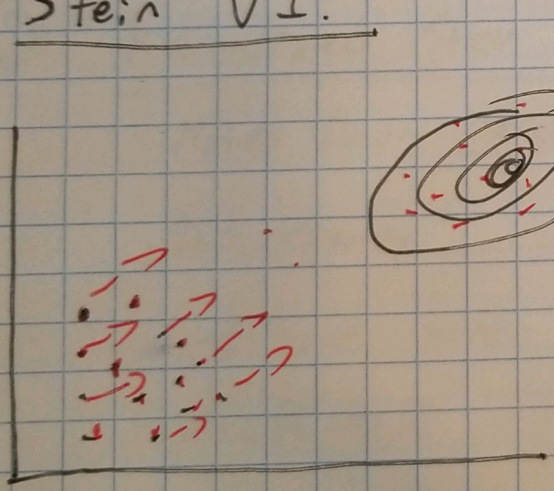
Intuition picture:

Traditional VI



Move $q_0 \to p$ by optimizing $\theta = \{u, \sigma\}$

Stein VI:



Move $\{x_i\}_{i=1}^{N} \to p$, $x_i \sim q_0$ by:

$$x_i^{t+1} \leftarrow x_i^t + \epsilon_t \hat{\phi}^*(x_i^t),$$

where $\hat{\phi}^*(x_i^t)$ comes from the "Kernalized Stein Discrepancy" (Defined Momentarily).

"Gradient Descent on Particles"

Road Map:

1. Introduce "Kernalized Stein Discrepancy" (KSD)
2. Introduce V. I. using "smooth transforms" from Family $\mathcal{Q}$
3. Glue 1 & 2 together by showing that $\nabla_{\mathcal{F}} KL[Q|P] \Rightarrow KSD$
4. Realize we can't solve the stein equations exactly, but we can operate on samples
5. Arrive at "Stein Variational Gradient Descent"

---

1) Kernalized Stein Discrepancy

- Notes from "A short introduction to Kernalized Stein Discrepancy"
- Say we have some smooth density $p(x)$ and a function $f(x): \mathbb{R}^d \to \mathbb{R}$

$$\lim_{\|x\| \to \infty} p(x) f(x) = 0 \quad , \forall x \in \mathbb{R}^d$$

- then we can derive "Stein's Identity" as follows:

$\Rightarrow$

$p(x) f(x) = 0$ $\quad x \in \mathbb{R}^d$, $f: \mathbb{R}^d \to \mathbb{R}$

$$= \int \nabla_x [p(x) f(x)] \, dx = 0$$

$$= \int [f(x) \nabla_x p(x) + p(x) \nabla_x f(x)] \, dx = 0$$

$$= \int [p(x) \nabla_x \log p(x) f(x) + p(x) \nabla_x f(x)] = 0$$

$$= E_{x \sim p(x)} [f(x) \nabla_x \log p(x) + \nabla_x f(x)] = 0$$

$$\therefore E_{x \sim p(x)} [\mathcal{A}_p [f(x)]] = 0$$

$$\mathcal{A}_p[f] = f(x) \nabla_x \log p(x) + \nabla_x f(x)$$

- $\mathcal{A}_p[\cdot]$ is a functional operator, called the "Stein operator"
- Gives an infinite number of identities indexed by choice of function f

"Reverse" of steins identity also holds; ie

Stein Identity

$$E_{x \sim p(x)} [\mathcal{A}_p [f(x)]] = 0$$

"Reverse" stein identity

$$E_{x \sim q(x)} [\mathcal{A}_p [f(x)]] \neq 0$$

(See "A short introduction to KSD" for proof)

→ The reverse stein identity therefore gives a measure of how "different" p & q are.

We can then ask:

> Which f maximizes the violation of Stein's identity?

This leads to the notion of the "Stein Discrepancy":

$$\sqrt{S(q,p)} = \max_{f \in F} E_{x\sim q}[A_p f(x)] \rightarrow S(q,p) = \max_{f \in F}[E_{x\sim q}[A]]$$

Choice of set F is critical;

→ must be computationally tractable & have discriminative power

Previous work (Liu et al) has shown that if you choose F to be the unit ball in the "reproducing kernal hilbert space" H_d the optimization has a closed form solution:

$$S(q,p) = \max_{f \in H^d}\{E_{x\sim q}[A_p f(x)] \quad s.t. \quad \|f\|_H \leq 1\}$$

$$f^*_{q,p}(x) = \frac{\hat{f}(x)}{\|\hat{f}\|_{H^d}}$$

$$\hat{f}_{q,p}(\cdot) = E_{x\sim q}[A_p k(x,\cdot)]$$

$$S(q,p) = \|f^*_{q,p}\|^2_{H_d}$$

"A hilbert space allows us to apply concepts from finite-dimensional linear algebra to infinite-dimensional function spaces" - Internet

"Reproducing" property

A kernal:

x be a non-empty set.

$k(\cdot,\cdot): x \times x \rightarrow \mathbb{R}$

$\phi: x \rightarrow H$

$k(x,x') = \langle \phi(x), \phi(x') \rangle_k$

Reproducing property (1 Defn)

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

$= \langle \phi(x), \phi(x) \rangle$

PART 2. Variational Inference Using Smooth Transforms:

Standard VI approximates $p(x)$ via a simpler distribution from some set $Q = \{q_i(x)\}$ by minimizing KL:

$$q^* = \underset{q \in Q}{\operatorname{argmin}} \{KL(q \| p)\} = E_q \log q(x) - E_q \log \hat{p}(x) + C$$

Consider set Q formed by smooth transforms of reference dist:

Indexed by T: $Q_T = \{z_i = T(x) \mid x \sim q_0(x),\ T: x \to x\}$

Then, by change of variables:

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

Jacobian

How to choose T?

→ Give T a parametric form and opt. parameters
→ "Normalizing Flow"

→ Or, we can iteratively select T's that minimize the KL.

→ Can be thought of as "functional gradient descent"

PART 3

Functional Gradient Descent

Regular gradient descent:

$$L(w) = \sum_{i=1}^{N} (y_i - w^T x_i)^2$$

Goal: $\min_w L(w)$

procedure:
- compute $\nabla_w L$
- select step size α_t

$$w_{t+1} = w_t - \alpha_t \nabla_w L(w)$$

Functional Gradient Descent

$$L[f] = \sum_{i=1}^{N} (y_i - f(x_i))^2$$

Goal: $\min_f L[f]$

procedure
- compute "Functional Gradient" $\nabla_f L[f]$
- select step size α_t
- $f_{t+1} = f_t - \alpha_t \nabla_f L$

Our problem:

$T(x_i) = x_i + f(x_i)$

$$L[f] = KL[\underbrace{q(x_i)}_{T = x + f(x)} \| p(x_i)]$$

Goal: $\min_f L[f]$

↳

Functional Gradients

C.C.F: "Functional Gradient Descent" Note by Drew Bagnell

A gradient can be thought of as:
- A vector of partial derivatives
- The direction of steepest ascent
- ☆ The linear approximation of the function:

$$f(x_0 + \epsilon) = f(x_0) + \epsilon \nabla f(x_0) + O(\epsilon^2)$$

↑ Gradient

A "functional" gradient is similar to the third definition, and is defined implicitly as the linear term of change in a function due to a small perturbation ϵ to its input.

Formal definition:

For any functional $F[f]$ of $f \in H^d$, its (functional) gradient $\nabla_f F[f]$ is a function in H^d s.t.:

$$F[f + \epsilon g] = F[f] + \epsilon \langle \nabla_f F[f], g \rangle_{H^d} + O(\epsilon^2)$$

For any $g \in H^d$ and $\epsilon \in \mathbb{R}$

PART 4

Procedure: (Attempt 1)

- Compute $\nabla_f[L] = \nabla_f KL(q_T \| p)\big|_{f=0} = -\phi^*_{q,p}(x)$ ↳ Lemma 3. in paper
- select step size ϵ_t
- → step

$$f_{t+1}(\cdot) \leftarrow f_t + \epsilon_t \phi^*_{q_t,p}(\cdot) \quad , \quad T_{t+1}(x) = x + \epsilon_t \phi^*_{q_t,p}(\cdot)$$

$$T_{t+1}(x) = x + f_{t+1}$$

$$\phi^*_{q_t,p}(\cdot) = E_{x \sim q_t}\left[k(\cdot, x)\nabla_x \log p(x)^T + \nabla_x k(\cdot, x)\right]$$

The evaluation at 0 is critical. Without it,

$$T_{t+1}(x) = x + \epsilon_t \phi^*_{q_t}(x) + \epsilon_{t-1} \phi^*_{q_{t-1}}(x) + \dots \epsilon_0 \phi^*_{q_0}(x), \quad \text{which}$$

is computationally infeasible.

But this means the step is only valid from $f = 0$, i.e. $T(x) = x + f$,

IE, this isn't allowed:

$$x_0 \sim q_0$$
$$z_1 = T_0(x_0)$$
$$z_2 = T_1(z_1)$$
$$\vdots$$
$$z_t = T_t(z_t)$$

or

$$x_0 \sim q_0$$
$$\Rightarrow (T_t \circ T_{t-1} \circ T_{t-2} \dots T_0)(x_0)$$

PART 5

We can get around this by moving the particles at each step so they are constantly sampled from $q_{[T^t]}$

We then empirically approximate $\hat{\phi}^*(x)$ via:

$$\hat{\phi}^*(\cdot) \approx \frac{1}{n} \sum_{j=1}^{N} [k(\cdot, x_j^t) \nabla_{x_j^t} \log p(x_j^t) + \nabla_{x_j^t} k(\cdot, x_j^t)$$

$$x_j^t \sim q_{[T^t]}(x)$$

This leads to the final algorithm, "Stein Variational (Functional) Gradient Descent":

Input: - Target density function $p(x)$
- Initial particles $\{x_i^0\}_{i=1}^N$

Output: - Particles $\{x_i\}_{i=1}^N$ that approximate $p(x)$

Procedure:

- for iteration t:
  - Make smooth transform function $T^t(\cdot)$:
    - $T^t(x) = x + \hat{\phi}^*(x)$
    - $\hat{\phi}^*(\cdot) = \frac{1}{N} \sum_{j=1}^{N} k(\cdot, x_j^t) \nabla_{x_j^t} \log p(x_j^t) + \nabla_{x_j^t} k(\cdot, x_j^t)$
  - Update Particles:
    $x_i^{t+1} = T^t(x_i^t)$