NOTES ON VARIATIONAL INFERENCE

He Li, NYU Stern

Last update: November 6, 2018

Contents

1 Background

Consider a joint distribution $p(x, z)$ where z is the latent variables. In the scenario of Bayesian statistics, z usually represents the model parameters. Given the prior on z: $p(z)$ and the likelihood function we obtain from data, $p(x|z)$, we can construct the posterior distribution on z:

$$p(z|x) \propto p(z)p(x|z) \tag{1.1}$$

where $z \in \mathbb{R}^m, x \in \mathbb{R}^n$. However, it is sometimes very hard to sample or do calculation from this posterior distribution. Therefore, we may need some technique to approximate the posterior. One way is Markov chain Monte Carlo (MCMC), it is statistically convincing but computationally expensive. Another way is using variational inference, a method that is much faster.

In variational inference, our goal is to find the best alternative distribution within a family of densities $\mathcal{Z}$ that is closed to the posterior under KL divergence.

$$q*(z) = \underset{q(z)\in\mathcal{Z}}{\operatorname{argmin}} \operatorname{KL}\left(q(z)\|p(z|x)\right) \tag{1.2}$$

2 The Evidence Lower Bound

The minimization task (1.2) in last section is not computable, since it contains the term $\log p(x)$. Let's see it.

$$\begin{aligned}
\operatorname{KL}\left(q(z)\|p(z|x)\right) &= \mathbb{E}_q[\log q(z)] - \mathbb{E}_q[\log p(z|x)] \\
&= \mathbb{E}_q[\log q(z)] - \mathbb{E}_q[\log p(z,x)] + \mathbb{E}_q[\log p(x)] \\
&= \mathbb{E}_q[\log q(z)] - \mathbb{E}_q[\log p(z,x)] + \log p(x)
\end{aligned} \tag{2.1}$$

We call this term $\log p(x) = \log \int p(z, x)\,\mathrm{d}z$ the *evidence* and in some cases this term needs exponential time to compute (the integral).

Instead of minimizing the (1.2), we define the *evidence lower bound (ELBO)* as

$$\begin{aligned}
\operatorname{ELBO}(q) :&= \mathbb{E}_q[\log p(z,x)] - \mathbb{E}_q[\log q(z)] \\
&= \mathbb{E}_q \log p(z) + \mathbb{E}_q p(x|z) - \mathbb{E}_q \log q(z) \\
&= \mathbb{E}_q p(x|z) - \operatorname{KL}\left(q(z)\|p(z)\right)
\end{aligned} \tag{2.2}$$

Notice that

$$\operatorname{ELBO}(q) = -\operatorname{KL}\left(q(z)\|p(z|x)\right) + \log p(x) \tag{2.3}$$

Obviously, maximizing ELBO is equivalent to minimizing the KL divergence. Given the fact that KL divergence is non-negative, we observe that

$$\operatorname{ELBO}(q) \leq \log p(x) \tag{2.4}$$

which indicates the name of ELBO. This can also be derived from below,

$$\begin{aligned}
\log p(x) &= \log \int p(x,z)\,\mathrm{d}z \\
&= \log \int \frac{p(x|z)p(z)q(z)}{q(z)}\,\mathrm{d}z \\
&= \log \mathbb{E}_q\left[\frac{p(x|z)p(z)}{q(z)}\right] \\
&\geq \mathbb{E}_q \log\left[\frac{p(x|z)p(z)}{q(z)}\right] \quad \text{Jensen's Inequality} \\
&= \mathbb{E}_q \log p(x|z) + \mathbb{E}_q \log p(z) - \mathbb{E}_q \log q(z) \\
&= \mathrm{ELBO}(q)
\end{aligned} \tag{2.5}$$

3 The Mean-field Variational Family and CAVI

We now focus on the mean-field variational family, where the latent variables are mutually independent and each governed by a distinct factor in the variational density, e.g.

$$q(\mathbf{z}) = \prod_{j=1}^{m} q_j(z_j) \tag{3.1}$$

Regarding this mean-field variational family, we now introduce the most commonly used alogrithm to solve this optimization problem: Coordinate ascent variational inference (CAVI). The CAVI is based on the following observation: consider a member $q_j(z_j)$, fix all the other variational factors, the optimal $q_j(z_j)$ is given by

$$q_j^*(z_j) \propto \exp\left\{\mathbb{E}_{-j}[\log p(z_j|\mathbf{z}_{-j}, x)]\right\} \tag{3.2}$$

To see this result, according to chain rule in the probability,

$$p(\mathbf{z}, \mathbf{x}) = p(\mathbf{x}) \prod_{k=1}^{m} p(z_k|\mathbf{z}_{1:k-1}, \mathbf{x}) \tag{3.3}$$

For any given z_j, consider it as the last member in the product, therefore the terms that are related to $q_j(z_j)$ in ELBO is

$$\begin{aligned}
&\mathbb{E}_q[\log p(z_j|\mathbf{z}_{-j}, \mathbf{x})] - \mathbb{E}_q[\log q_j(z_j)] \\
=& \int q_j(z_j)\mathbb{E}_{q_{-j}}\left[\log p(z_j|\mathbf{z}_{-j}, \mathbf{x})\right]\,\mathrm{d}z_j - \int q_j(z_j)\log q_j(z_j)\,\mathrm{d}z_j
\end{aligned} \tag{3.4}$$

To calculate the derivative of $q_j(z_j)$ on ELBO, we first need to define the functional derivative.

Definition 3.1. Given a manifold M representing (continuous/smooth) functions ρ (with certain boundary conditions etc.), and a functional F defined as

$$F : M \to \mathbb{R}$$

the functional derivative of $F[\rho]$, denoted $\frac{\delta F}{\delta \rho}$, is defined by

$$\int \frac{\delta F}{\delta \rho} \phi(x)\, \mathrm{d}x = \lim_{\epsilon \to 0} \frac{F[\rho + \epsilon\phi] - F(\rho)}{\epsilon} = \left[\frac{\mathrm{d}}{\mathrm{d}\epsilon} F[\rho + \epsilon\phi] \right]_{\epsilon=0} \tag{3.5}$$

where ϕ is any nice function and $\phi = 0$ on the boundary of the region of integration. The quantity $\epsilon\phi$ is called the variation of ρ. In other words,

$$\phi \to \left[\frac{\mathrm{d}}{\mathrm{d}\epsilon} F[\rho + \epsilon\phi] \right]_{\epsilon=0} \tag{3.6}$$

is a linear functional, so by the Riesz–Markov–Kakutani representation theorem, this functional is given by integration against some measure. Then $\frac{\delta F}{\delta \rho}$ is defined to be the Radon–Nikodym derivative of this measure.

Given a functional

$$F[\rho] = \int f(x, \rho(x), \nabla\rho(x))\, \mathrm{d}x \tag{3.7}$$

and any function ϕ, the functional derivative of $F[\rho]$ is,

$$\begin{aligned}
\int \frac{\delta F}{\delta \rho} \phi(x)\, \mathrm{d}x &= \left[\frac{\mathrm{d}}{\mathrm{d}\epsilon} \int f(x, \rho(x) + \epsilon\phi(x), \nabla\rho(x) + \epsilon\nabla\phi(x))\, \mathrm{d}x \right]_{\epsilon=0} \\
&= \int \left(\frac{\partial f}{\partial \rho}\phi + \frac{\partial f}{\partial \nabla\rho}\nabla\phi \right) \mathrm{d}x \\
&= \int \left[\frac{\partial f}{\partial \rho}\phi + \nabla \cdot \left(\frac{\partial f}{\partial \nabla\rho}\phi \right) - \left(\nabla \cdot \frac{\partial f}{\partial \nabla\rho} \right) \phi \right] \mathrm{d}x \\
&= \int \left[\frac{\partial f}{\partial \rho}\phi - \left(\nabla \cdot \frac{\partial f}{\partial \nabla\rho} \right) \phi \right] \mathrm{d}x + \oint_S \left(\frac{\partial f}{\partial \nabla\rho}\phi \right) \mathrm{d}x, \quad \text{Gauss Theorem} \\
&= \int \left[\frac{\partial f}{\partial \rho}\phi - \left(\nabla \cdot \frac{\partial f}{\partial \nabla\rho} \right) \phi \right] \mathrm{d}x, \quad \phi = 0 \text{ on } S \\
&= \int \left[\frac{\partial f}{\partial \rho} - \left(\nabla \cdot \frac{\partial f}{\partial \nabla\rho} \right) \right] \phi\, \mathrm{d}x
\end{aligned} \tag{3.8}$$

Therefore, we have

$$\int \left[\frac{\delta F}{\delta \rho} - \frac{\partial f}{\partial \rho} + \left(\nabla \cdot \frac{\partial f}{\partial \nabla\rho} \right) \right] \phi\, \mathrm{d}x = 0 \tag{3.9}$$

According to the fundamental lemma of calculus of variations below, we have

$$\frac{\delta F}{\delta \rho} = \frac{\partial f}{\partial \rho} - \nabla \cdot \frac{\partial f}{\partial \nabla\rho} \tag{3.10}$$

which is called *Euler–Lagrange equation*. More properties of functional derivatives can be found here.

Theorem 3.2. Fundamental lemma of calculus of variations. If a continuous multivariable function f on an open set $\Omega \subset \mathbb{R}^d$ satisfies the equality

$$\int_\Omega f(x)h(x)\,\mathrm{d}x = 0 \tag{3.11}$$

for all compactly supported smooth functions h on Ω, then $f \equiv 0$ on Ω.

Similarly, one may consider a continuous function f on the closure of Ω, assuming that h vanishes on the boundary of Ω (rather than compactly supported).

Also, for discontinuous multivariable functions, Let $\Omega \subset \mathbb{R}^d$ be an open set, and $f \in \mathrm{L}^2(\Omega)$ satisfy the equality

$$\int_\Omega f(x)h(x)\,\mathrm{d}x = 0 \tag{3.12}$$

for all compactly supported smooth functions h on Ω. Then $f \equiv 0$.

Now let's head back to (3.2), denote $\mathrm{ELBO}(q) = \mathcal{L}$, we have

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial q_j(z_j)} &= \frac{\partial}{\partial q_j(z_j)}\left[\int q_j(z_j)\mathbb{E}_{q_{-j}}\left[\log p(z_j|\mathbf{z}_{-j},\mathbf{x})\right]\,\mathrm{d}z_j - \int q_j(z_j)\log q_j(z_j)\,\mathrm{d}z_j\right] \\ &= \mathbb{E}_{q_{-j}}\left[\log p(z_j|\mathbf{z}_{-j},\mathbf{x})\right] - \log q_j(z_j) - 1\end{aligned} \tag{3.13}$$

where we use the Euler–Lagrange equation (3.10) (note that this functional does not have $\nabla\rho$ term).

Therefore, the optimal $q_j(z_j)$ is,

$$q_j^*(z_j) \propto \exp\left\{\mathbb{E}_{-j}[\log p(z_j|\mathbf{z}_{-j},x)]\right\} \tag{3.14}$$

which is equivalent to

$$q_j^*(z_j) \propto \exp\left\{\mathbb{E}_{-j}[\log p(z_j,\mathbf{z}_{-j},x)]\right\} \tag{3.15}$$

Algorithm 1: Coordinate Ascent Variational Inference (CAVI)

Data: A model $p(\mathbf{x},\mathbf{z})$, a data set $\mathbf{x}$
Result: A variational density $q(\mathbf{z}) = \prod_{j=1}^m q_j(z_j)$
Initialize: Variational factors $q_j(z_j)$;
while *the ELBO has not converged* **do**
  for $j \in 1,\cdots,m$ **do**
    Set $q_j^*(z_j) \propto \exp\left\{\mathbb{E}_{-j}[\log p(z_j|\mathbf{z}_{-j},x)]\right\}$;
  end
end

4 Black Box Variational Inference

Stochastic Optimization

Under Mean-field assumptions, consider the ELBO

$$\mathcal{L}(\lambda) \triangleq \mathbb{E}_{q_\lambda(z)}[\log p(x,z) - \log q(z)] \tag{4.1}$$

Let's consider **stochastic optimization** method. Define $f(x)$ to be the target function we are going to maximize, and $h(t)$ be the realization of a random variable whose expectation is the gradient of $f(x)$. Then, we can optimize the target iterally with

$$x_{t+1} \leftarrow x_t + \rho_t h_t(x) \tag{4.2}$$

where ρ_t is the learning rate. This converges to a maximum of $f(x)$ when the learning rate schedule follows the Robbins-Monro conditions

$$\sum_{t=1}^{\infty} \rho_t = \infty, \ \sum_{t=1}^{\infty} \rho_t^2 < \infty \tag{4.3}$$

As derived in the appendix of [4], the gradient of ELBO can be written as

$$\nabla_\lambda \mathcal{L} = \mathbb{E}_q[\nabla_\lambda \log q(z|\lambda)(\log p(x,z) - \log q(z|\lambda))] \tag{4.4}$$

Therefore, we can use Monte-Carlo to do the stochastic optimization with

$$\nabla_\lambda \mathcal{L} \approx \frac{1}{S} \sum_{s=1}^{S} \nabla_\lambda \log q(z_s|\lambda)(\log p(x,z_s) - \log q(z_s|\lambda)) \tag{4.5}$$

where $z_s \sim q(z|\lambda)$.

Rao-Blackwellization

In practice, people usually use the **Rao-Blackwellization** method to reduce the variance of Monte-Carlo sampling. If we want to compute the expectation of a function $\mathbb{E}f(x,y)$, Rao-Blackwellization can be simplified as using conditional expectation $\hat{f}(x,y) = \mathbb{E}(f(x,y)|x)$ as an estimator, given the fact that

$$\mathbb{E}\left[\mathbb{E}(f(x,y)|x)\right] = \mathbb{E}(f(x,y)) \tag{4.6}$$

and

$$\mathrm{Var}(f(x,y)) = \mathbb{E}(\mathrm{Var}(f(x,y)|x)) + \mathrm{Var}(\mathbb{E}(f(x,y)|x)) \geq \mathrm{Var}(\mathbb{E}(f(x,y)|x)) \tag{4.7}$$

This means that $\hat{f}(x,y)$ is a lower variance estimator than $f(x,y)$. Due to the mean-field assumption,

$$\begin{aligned}\mathbb{E}(f(x,Y)|x) &= \int \frac{f(x,Y)p(x,Y)}{p(x)}\,\mathrm{d}Y \\ &= \int \frac{f(x,Y)p(x)p(Y)}{p(x)}\,\mathrm{d}Y \\ &= \int f(x,Y)p(Y)\,\mathrm{d}Y \\ &= \mathbb{E}_Y f(x,Y) \end{aligned} \tag{4.8}$$

Therefore, for each component of the gradient, we should compute expectations with respect to the other factors. Derived from the supplement material of [4], we conclude that

$$\nabla_{\lambda_i}\mathcal{L} = \mathbb{E}_{q(i)}[\nabla_{\lambda_i}\log q(z_i|\lambda_i)(\log p_i(x, z_{(i)}) - \log q(z_i|\lambda_i))] \tag{4.9}$$

where $q(i)$ be the distribution of variables in the model that depend on the ith variable, i.e., the Markov blanket of z_i; and $p_i(x, z_{(i)})$ be the terms in the joint that depend on those variables.

Control Variables

A control variate is a family of functions with equivalent expectation. Consider a function h, which has a finite first moment, and a scalar a. Define $\hat{f}$ to be

$$\hat{f}(z) \triangleq f(z) - a\left[h(z) - \mathbb{E}(h(z))\right] \tag{4.10}$$

Therefore, we can choose a to minimize te variance of $\hat{f}$:

$$\mathrm{Var}(\hat{f}) = \mathrm{Var}(f) + a^2\mathrm{Var}(h) - 2a\mathrm{Cov}(f,h) \tag{4.11}$$

where the minimal is obtained at $a^* = \frac{\mathrm{Cov}(f,h)}{\mathrm{Var}(h)}$. Notice that when $a = 0$, $\hat{f} = f$, therefore the minimal variance is no greater than f's variance.

Back to our algorithm, we can define

$$\begin{aligned} f_i(z) &= \nabla_{\lambda_i}\log q(z_i|\lambda_i)(\log p_i(x,z) - \log q_i(z|\lambda_i)) \\ h_i(z) &= \nabla_{\lambda_i}\log q(z_i|\lambda_i) \end{aligned} \tag{4.12}$$

Note that

$$\begin{aligned}\mathbb{E}(\nabla_{\lambda_i}\log q(z_i|\lambda_i)) &= \mathbb{E}\left(\frac{\nabla_{\lambda_i} q(z_i|\lambda_i)}{q(z_i|\lambda_i)}\right) \\ &= \int \nabla_{\lambda_i} q(z_i|\lambda_i)\,\mathrm{d}z_i \\ &= \nabla_{\lambda_i}\int q(z_i|\lambda_i)\,\mathrm{d}z_i \\ &= \nabla_{\lambda_i} 1 = 0 \end{aligned} \tag{4.13}$$

Therefore $\mathbb{E}(\nabla_{\lambda_i} \log q(z|\lambda_i)) = 0$. Moreover, we can use sampling value to estimate $\hat{a}_i^* = \frac{\hat{\text{Cov}}(f_i, h_i)}{\hat{\text{Var}}(h_i)}$. Finally, we can use the estimated gradient:

$$\nabla_{\lambda_i} \mathcal{L} \approx \frac{1}{S} \sum_{s=1}^{S} \nabla_{\lambda_i} \log q_i(z_s|\lambda_i)(\log p_i(x, z_s) - \log q_i(z_s|\lambda_i) - \hat{a}_i^*), \quad z_s \sim q_{(i)}(z|\lambda) \tag{4.14}$$

The logic of black box variational inference is as follows:

find a maximum value $\Longrightarrow$ *use stochastic optimization* $\Longrightarrow$ *monto-carlo sampling to estimate gradient* $\Longrightarrow$ *use Rao-Blackwellization and control variables to lower the sampling variance.*

5 Reparameterization

We reparameterize the latent variable in terms of a known base distribution and a differentiable transformation (such as a location-scale transformation or cumulative distribution function). For example, if $q_\phi(z)$ is a Gaussian distribution $\mathcal{N}(z|\mu, \sigma^2)$, with $\phi = \{\mu, \sigma^2\}$, then the location-scale transformation using the standard Normal as a base distribution allows us to reparameterize z as:

$$z \sim \mathcal{N}(z|\mu, \sigma^2) \Leftrightarrow z = \mu + \sigma\epsilon, \quad \epsilon \sim \mathcal{N}(0, 1) \tag{5.1}$$

Therefore, in the backpropogation step, we can exchange the following derivative on LHS to the easier-to-compute one on the RHS.

$$\nabla_\lambda \mathbb{E}_{q_\lambda(z)}[f(\theta, \lambda, x, z)] \Leftrightarrow \nabla_\lambda \mathbb{E}_{\epsilon \sim \mathcal{N}(0,1)}[f(\theta, \lambda, x, \mu + \sigma\epsilon)] \tag{5.2}$$

Notice that none of the expectations are with respect to distributions that depend on our model parameters, so we can safely move a gradient symbol into them while maintaning equality. That is, given a fixed X and ϵ, this function is deterministic and continuous in the parameters of p and q, meaning backpropagation can compute a gradient that will work for stochastic gradient descent.

$$\nabla_\lambda \mathbb{E}_{\epsilon \sim \mathcal{N}(0,1)}[f(\theta, \lambda, x, \mu + \sigma\epsilon)] = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,1)}[\nabla_\lambda f(\theta, \lambda, x, \mu + \sigma\epsilon)] \tag{5.3}$$

6 Variational Inference with Normalizing flow

Finite Flows

A normalizing flow describes the transformation of a probability density through a sequence of invertible mappings. By repeatedly applying the rule for change of variables, the initial density "flows" through the sequence of invertible mappings. At the end of this sequence we obtain a valid probability distribution and hence this type of flow is referred to as a normalizing flow.

The basic rule for transformation of densities considers an invertible, smooth mapping $f : \mathbb{R}^d \to \mathbb{R}^d$ with inverse $f^1 = g$, i.e. the composition $g \circ f(\mathbf{z}) = \mathbf{z}$. If we use this mapping to transform a random variable $\mathbf{z}$ with distribution $q(\mathbf{z})$, the resulting random variable $\mathbf{z}' = f(\mathbf{z})$ has a distribution

$$q(\mathbf{z}') = q(\mathbf{z}) \left| \det \frac{\partial f^{-1}}{\partial \mathbf{z}'} \right| = q(\mathbf{z}) \left| \det \frac{\partial f}{\partial \mathbf{z}'} \right|^{-1} \tag{6.1}$$

where the last equality can be seen by applying the chain rule (inverse function theorem) and is a property of Jacobians of invertible functions. We can construct arbitrarily complex densities by composing several simple maps and successively applying (6.1). The density $q_K(z)$ obtained by successively transforming a random variable z_0 with distribution q_0 through a chain of K transformations f_k is:

$$\mathbf{z}_K = f_K \circ \cdots \circ f_2 \circ f_1(\mathbf{z}_0) \tag{6.2}$$

$$\log q_k(\mathbf{z}_K) = \log q_0(\mathbf{z}_0) - \sum_{k=1}^{K} \log \left| \det \frac{\partial f}{\partial \mathbf{z}_{k-1}} \right| \tag{6.3}$$

The path traversed by the random variables $\mathbf{z}_k = f_k(\mathbf{z}_{k1})$ with initial distribution $q_0(\mathbf{z}_0)$ is called the flow and the path formed by the successive distributions q_k is a normalizing flow.

A property of such transformations, often referred to as the law of the unconscious statistician (LOTUS), is that expectations w.r.t. the transformed density q_K can be computed without explicitly knowing q_K . Any expectation $\mathbb{E}_{q_K}[h(\mathbf{z})]$ can be written as an expectation under q_0 as:

$$\mathbb{E}_{q_K}[h(\mathbf{z})] = \mathbb{E}_{q_0}[h(f_K \circ \cdots \circ f_2 \circ f_1(\mathbf{z}_0)] \tag{6.4}$$

which does not require computation of the the logdet-Jacobian terms when $h(\mathbf{z})$ does not depend on q_K.

Infinitesimal Flows

Please refer to [5] for some introduction on infinitesimal flows like *Langevin Flow* as well as *Hamiltonian Flow*.

Invertible Linear-time Transformations

We introduce two class of invertible linear-time transformations to compute the determinant efficient.

We consider a family of transformations of the form:

$$f(\mathbf{z}) = \mathbf{z} + \mathbf{u}h(\mathbf{w}^T\mathbf{z} + b) \tag{6.5}$$

where $h(\cdot)$ is a smooth element-wise non-linear function with first order derivative $h'(\cdot)$.

Lemma 6.1. Suppose A is an invertible square matrix and u, v are column vectors. Then the *matrix determinant lemma* states that

$$\det(\mathbf{A} + \mathbf{u}\mathbf{v}^T) = \det(\mathbf{A})(1 + \mathbf{v}^T\mathbf{A}^{-1}\mathbf{u}) \tag{6.6}$$

we can compute the logdet-Jacobian term in $O(D)$ time (D is the dimension of $\mathbf{z}$):

$$\left| \det \frac{\partial f}{\partial \mathbf{z}} \right| = |\det(\mathbf{I} + \mathbf{u}\psi(\mathbf{z})^T)| = |1 + \mathbf{u}^T\psi(\mathbf{z})| \tag{6.7}$$

where $\psi(\mathbf{z}) = h'(\mathbf{w}^T\mathbf{z} + b)\mathbf{w}$.

Therefore,

$$\ln q_k(\mathbf{z}_K) = \log q_0(\mathbf{z}_0) - \sum_{k=1}^{K} \log |1 + \mathbf{u}_k^T \psi_k(\mathbf{z_{k-1}})| \tag{6.8}$$

and the ELBO defined in (4.1) in this case can be written as

$$\begin{aligned}
\mathcal{L}(\lambda) &= \mathbb{E}_{q_\lambda(\mathbf{z})}[\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z})] \\
&= \mathbb{E}_{q_0(\mathbf{z})}[\log q_K(\mathbf{z}_K) - \log p(\mathbf{x}, \mathbf{z}_K)] \\
&= \mathbb{E}_{q_0(\mathbf{z})}[\log q_K(\mathbf{z}_K)] - \mathbb{E}_{q_0(\mathbf{z})}[\log p(\mathbf{x}, \mathbf{z}_K)] - \mathbb{E}_{q_0(\mathbf{z})}\left[\sum_{k=1}^{K} \log[1 + \mathbf{u}_k^T \psi_k(\mathbf{z_{k-1}})]\right]
\end{aligned} \tag{6.9}$$

we can therefore construct an inference model using a deep neural network to build a mapping from the observations x to the parameters of the initial density $q_0 = \mathcal{N}(\mu, \sigma^2)$.

The flow defined by the transformation (6.5) modifies the initial density q_0 by applying a series of contractions and expansions in the direction perpendicular to the hyperplane $\mathbf{w}^T\mathbf{z}+b = 0$, hence we refer to these maps as planar flows.

As an alternative, we can consider a family of transformations that modify an initial density q_0 around a reference point $\mathbf{z}_0$. The transformation family is:

$$f(\mathbf{z}) = \mathbf{z} + \beta h(\alpha, r)(\mathbf{z} - \mathbf{z}_0), \quad r = |\mathbf{z} - \mathbf{z}_0|, \; h(\alpha, r) = \frac{1}{\alpha + r} \tag{6.10}$$

and similiarly,

$$\left|\det \frac{\partial f}{\partial \mathbf{z}}\right| = [1 + \beta h(\alpha, r)]^{d-1}[1 + \beta h(\alpha, r) + \beta h'(\alpha, r)r] \tag{6.11}$$

References

[1] David Blei, Alp Kucukelbir, and Jon McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[2] Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

[3] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[4] Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 814–822, 2014.

[5] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.