A Computing conditional expectations in the input space

We here give a derivation for the target density on the constraint manifold in the input space of a differentiable generative model for computing expectations conditional on observations of the output. This is largely a restatement of results in [1, §2.3] and is provided mainly to make those results more easily relatable to the notation of this paper.

For clarity in this section the measure being integrated with respect to will be explicitly denoted. For a variable of integration $\boldsymbol{x}$, $\lambda^D\left\{\mathrm{d}\boldsymbol{x}\right\}$ will denote the D dimensional Lebesgue measure and $\mathcal{H}^D\left\{\mathrm{d}\boldsymbol{x}\right\}$ the D dimensional Hausdorff measure over some space $\mathcal{X}$ which will be specified.

A key result we will use is Federer's *Co-Area Formula* [2, §3.2.12]:

Theorem (Co-Area Formula)**.** *Let $\boldsymbol{M} : \mathcal{X} \subseteq \mathbb{R}^L \to \mathcal{V} \subseteq \mathbb{R}^K$ be Lipschitz with $L > K$ and $h : \mathcal{X} \to \mathbb{R}$ be Lebesgue measurable. Then*

$$\int_{\mathcal{X}} h(\boldsymbol{x}) \left|\frac{\partial \boldsymbol{M}}{\partial \boldsymbol{x}} \frac{\partial \boldsymbol{M}}{\partial \boldsymbol{x}}^{\mathrm{T}}\right|^{\frac{1}{2}} \lambda^L\left\{\mathrm{d}\boldsymbol{x}\right\} = \int_{\mathcal{V}} \int_{\boldsymbol{M}^{-1}(\boldsymbol{v})} h(\boldsymbol{x})\, \mathcal{H}^{L-K}\left\{\mathrm{d}\boldsymbol{x}\right\}\, \lambda^K\left\{\mathrm{d}\boldsymbol{v}\right\} \tag{1}$$

with $\frac{\partial \boldsymbol{M}}{\partial \boldsymbol{x}} = \left[\frac{\partial m_i}{\partial x_j}\right]_{i,j}$ the Jacobian of the map, $\boldsymbol{M}^{-1}(\boldsymbol{v})$ the $L-K$ dimensional sub-manifold embedded in $\mathcal{X}$ with Hausdorff measure $\mathcal{H}^{L-K}\left\{\mathrm{d}\boldsymbol{x}\right\}$, which is the solution set $\left\{\boldsymbol{x} \in \mathcal{X} : \boldsymbol{M}(\boldsymbol{x}) = \boldsymbol{v}\right\}$.

Corollary. *If the Jacobian of $\boldsymbol{M}$ has full row-rank everywhere such that $\left|\frac{\partial \boldsymbol{M}}{\partial \boldsymbol{x}} \frac{\partial \boldsymbol{M}}{\partial \boldsymbol{x}}^{\mathrm{T}}\right| > 0 \;\forall \boldsymbol{x} \in \mathcal{X}$ then for a Lebesgue measurable $h^* : \mathcal{X} \to \mathbb{R}$*

$$\int_{\mathcal{X}} h^*(\boldsymbol{x})\, \lambda^L\left\{\mathrm{d}\boldsymbol{x}\right\} = \int_{\mathcal{V}} \int_{\boldsymbol{M}^{-1}(\boldsymbol{v})} h^*(\boldsymbol{x}) \left|\frac{\partial \boldsymbol{M}}{\partial \boldsymbol{x}} \frac{\partial \boldsymbol{M}}{\partial \boldsymbol{x}}^{\mathrm{T}}\right|^{-\frac{1}{2}} \mathcal{H}^{L-K}\left\{\mathrm{d}\boldsymbol{x}\right\}\, \lambda^K\left\{\mathrm{d}\boldsymbol{v}\right\} \tag{2}$$

which can be easily shown by setting $h(\boldsymbol{x}) = h^*(\boldsymbol{x}) \left|\frac{\partial \boldsymbol{M}}{\partial \boldsymbol{x}} \frac{\partial \boldsymbol{M}}{\partial \boldsymbol{x}}^{\mathrm{T}}\right|^{-\frac{1}{2}}$ in (1).

We will also use what is sometimes termed the *Law of the Unconscious Statistician* (LOTUS) to express expectations of functions of random (vector) variables when an explicit density on the random output of the function is not known.

Theorem (Law of the Unconscious Statistician)**.** *Let* $\mathbf{y}$ *be a random vector on support $\mathcal{X} \subseteq \mathbb{R}^N$ with density $\mathbb{p}_{\mathbf{x}}\left[\boldsymbol{x}\right]$ with respect to the Lebesgue measure $\lambda^N\left\{\mathrm{d}\boldsymbol{x}\right\}$ and $f : \mathcal{Y} \to \mathbb{R}$ be Lebesgue measurable. If we define a new random variable* $\mathrm{v} = f(\mathbf{x})$ *then*

$$\mathbb{E}\left[\mathrm{v}\right] = \mathbb{E}\left[f(\mathbf{x})\right] = \int_{\mathcal{X}} f(\boldsymbol{x})\, \mathbb{p}_{\mathbf{x}}\left[\boldsymbol{x}\right] \lambda^N\left\{\mathrm{d}\boldsymbol{x}\right\}. \tag{3}$$

Corollary. *If* $\mathbf{x}$ *is defined as* $\mathbf{x} = \boldsymbol{G}(\mathbf{u})$ *for some $\boldsymbol{G} : \mathcal{U} \subseteq \mathbb{R}^M \to \mathcal{X}$ then*

$$\mathbb{E}\left[f(\mathbf{x})\right] = \mathbb{E}\left[(f \circ \boldsymbol{G})(\mathbf{u})\right] = \int_{\mathcal{U}} f \circ \boldsymbol{G}(\boldsymbol{u})\, \mathbb{p}_{\mathbf{u}}\left[\boldsymbol{u}\right] \lambda^M\left\{\mathrm{d}\boldsymbol{u}\right\}. \tag{4}$$

This leads us to the main result

Theorem. *Let* $\mathbf{u}$ *be a random vector with density $\mathbb{p}_{\mathbf{u}}\left[\boldsymbol{u}\right] = \rho(\boldsymbol{u})$ with respect to the Lebesgue measure $\lambda^M\left\{\mathrm{d}\boldsymbol{u}\right\}$ on support $\mathcal{U} = \mathbb{R}^M$. Further let $\boldsymbol{G} : \mathcal{U} \to \mathcal{X}$ be a smooth map, with $\mathcal{X} = \mathbb{R}^N$; $N \leq M$ defining a random vector* $\mathbf{x} = \boldsymbol{G}(\mathbf{u})$. *Assume $\frac{\partial \boldsymbol{G}}{\partial \boldsymbol{u}}$ exists and has full row-rank almost everywhere.*

Partition the output space $\mathcal{X} = \mathcal{Y} \times \mathcal{Z}$ with $\mathcal{Y} = \mathbb{R}^{N_y}$ and $\mathcal{Z} = \mathbb{R}^{N_z}$ and $\mathbf{y} = \boldsymbol{G}_y(\mathbf{u})$, $\mathbf{z} = \boldsymbol{G}_z(\mathbf{u})$. *Then the conditional expectation of some function f of* $\mathbf{z}$ *given* $\mathbf{y} = \bar{\boldsymbol{y}}$ *has been observed is*

$$\mathbb{E}\left[f(\mathbf{z}) \,|\, \mathbf{y} = \bar{\boldsymbol{y}}\right] = \frac{1}{\mathbb{p}_{\mathbf{y}}\left[\bar{\boldsymbol{y}}\right]} \int_C f \circ \boldsymbol{G}_z(\boldsymbol{u})\, \rho(\boldsymbol{u}) \left|\frac{\partial \boldsymbol{G}_y}{\partial \boldsymbol{u}} \frac{\partial \boldsymbol{G}_y}{\partial \boldsymbol{u}}^{\mathrm{T}}\right|^{-\frac{1}{2}} \mathcal{H}^{M-N_y}\left\{\mathrm{d}\boldsymbol{u}\right\}. \tag{5}$$

with $\mathbb{p}_{\mathbf{y}}\left[\bar{\boldsymbol{y}}\right]$ the marginal density on $\mathbf{y}$ *with respect to the Lebesgue measure $\lambda^{N_y}\left\{\mathrm{d}\boldsymbol{y}\right\}$ which must be non-zero for the conditional expectation to be well-defined; C is the $M - N_y$ dimensional sub-manifold defined by the solution set $\left\{\boldsymbol{u} \in \mathcal{U} : \boldsymbol{G}_y(\boldsymbol{u}) = \bar{\boldsymbol{y}}\right\}$.*

Proof. By the *Law of Total Expectation* we have that

$$\mathbb{E}\left[f(\mathbf{z})\right] = \int_{\mathcal{Y}} \mathbb{E}\left[f(\mathbf{z}) \,|\, \mathbf{y} = \boldsymbol{y}\right] \, \mathbb{p}_{\mathbf{y}}\left[\boldsymbol{y}\right] \lambda^{N_y}\left\{\mathrm{d}\boldsymbol{y}\right\}. \tag{6}$$

Using LOTUS (3) we get

$$\int_{\mathcal{Y}} \mathbb{E}\left[f(\mathbf{z}) \,|\, \mathbf{y} = \boldsymbol{y}\right] \, \mathbb{p}_{\mathbf{y}}\left[\boldsymbol{y}\right] \lambda^{N_y}\left\{\mathrm{d}\boldsymbol{y}\right\} = \int_{\mathcal{U}} f \circ \boldsymbol{G}_z(\boldsymbol{u})\, \rho(\boldsymbol{u})\, \lambda^{M}\left\{\mathrm{d}\boldsymbol{u}\right\}. \tag{7}$$

Applying the co-area formula corollary (2) to the right-hand side gives

$$\int_{\mathcal{Y}} \mathbb{E}\left[f(\mathbf{z}) \,|\, \mathbf{y} = \boldsymbol{y}\right] \, \mathbb{p}_{\mathbf{y}}\left[\boldsymbol{y}\right] \lambda^{N_y}\left\{\mathrm{d}\boldsymbol{y}\right\} = \tag{8}$$

$$\int_{\mathcal{Y}} \int_{\boldsymbol{G}_y^{-1}(\boldsymbol{y})} f \circ \boldsymbol{G}_z(\boldsymbol{u})\, \rho(\boldsymbol{u}) \left| \frac{\partial \boldsymbol{G}_y}{\partial \boldsymbol{u}} \frac{\partial \boldsymbol{G}_y}{\partial \boldsymbol{u}}^{\mathrm{T}} \right|^{-\frac{1}{2}} \mathcal{H}^{M-N_y}\left\{\mathrm{d}\boldsymbol{u}\right\} \lambda^{N_y}\left\{\mathrm{d}\boldsymbol{y}\right\}. \tag{9}$$

Define $\mathcal{Y}^\star = \left\{\boldsymbol{y} \in \mathcal{Y} : \mathbb{p}_{\mathbf{y}}\left[\boldsymbol{y}\right] > 0\right\}$. Then we have

$$\int_{\mathcal{Y}^\star} \left\{\mathbb{E}\left[f(\mathbf{z}) \,|\, \mathbf{y} = \boldsymbol{y}\right]\right\} \, \mathbb{p}_{\mathbf{y}}\left[\boldsymbol{y}\right] \lambda^{N_y}\left\{\mathrm{d}\boldsymbol{y}\right\} = \tag{10}$$

$$\int_{\mathcal{Y}^\star} \left\{ \frac{1}{\mathbb{p}_{\mathbf{y}}\left[\boldsymbol{y}\right]} \int_{G_y^{-1}(\boldsymbol{y})} f \circ \boldsymbol{G}_z(\boldsymbol{u})\, \rho(\boldsymbol{u}) \left| \frac{\partial \boldsymbol{G}_y}{\partial \boldsymbol{u}} \frac{\partial \boldsymbol{G}_y}{\partial \boldsymbol{u}}^{\mathrm{T}} \right|^{-\frac{1}{2}} \mathcal{H}^{M-N_y}\left\{\mathrm{d}\boldsymbol{u}\right\} \right\} \mathbb{p}_{\mathbf{y}}\left[\boldsymbol{y}\right] \lambda^{N_y}\left\{\mathrm{d}\boldsymbol{y}\right\}. \tag{11}$$

As this holds for arbitrary Lebesgue measurable f, this implies that the terms inside the braces are equal for all $\boldsymbol{y} \in \mathcal{Y}^\star$. As $\bar{\boldsymbol{y}} \in \mathcal{Y}^\star$ by assumption and $C = \boldsymbol{G}_y^{-1}(\bar{\boldsymbol{y}})$ we have

$$\mathbb{E}\left[f(\mathbf{z}) \,|\, \mathbf{y} = \bar{\boldsymbol{y}}\right] = \frac{1}{\mathbb{p}_{\mathbf{y}}\left[\bar{\boldsymbol{y}}\right]} \int_{C} f \circ \boldsymbol{G}_z(\boldsymbol{u})\, \rho(\boldsymbol{u}) \left| \frac{\partial \boldsymbol{G}_y}{\partial \boldsymbol{u}} \frac{\partial \boldsymbol{G}_y}{\partial \boldsymbol{u}}^{\mathrm{T}} \right|^{-\frac{1}{2}} \mathcal{H}^{M-N_y}\left\{\mathrm{d}\boldsymbol{u}\right\}. \tag{12}$$

□

Corollary. *Define a target density with respect to the Hausdorff measure* $\mathcal{H}^{M-N_y}\left\{\mathrm{d}\boldsymbol{u}\right\}$ *on* C

$$\pi(\boldsymbol{u}) \propto \rho(\boldsymbol{u}) \left| \frac{\partial \boldsymbol{G}_y}{\partial \boldsymbol{u}} \frac{\partial \boldsymbol{G}_y}{\partial \boldsymbol{u}}^{\mathrm{T}} \right|^{-\frac{1}{2}}. \tag{13}$$

If we generate a set of MCMC samples $\left\{\boldsymbol{u}^{(s)}\right\}_{s=1}^{S}$ *which leave* $\pi(\boldsymbol{u})$ *invariant with respect to* $\mathcal{H}^{M-N_y}\left\{\mathrm{d}\boldsymbol{u}\right\}$ *on* C*, by the* Law of Large Numbers *we can then form a Monte Carlo estimate for the conditional expectation*

$$\mathbb{E}\left[f(\mathbf{z}) \,|\, \mathbf{y} = \bar{\boldsymbol{y}}\right] = \lim_{S \to \infty} \frac{1}{S} \sum_{s=1}^{S} \left(f \circ \boldsymbol{G}_z(\boldsymbol{u}^{(s)})\right). \tag{14}$$

B Evaluating the target density and its gradient

For the constrained Hamiltonian dynamics we need to be able to evaluate the logarithm of the target density (13) up to an additive constant and its gradient with respect to $\boldsymbol{u}$. We have that

$$\log \pi(\boldsymbol{u}) = \log \rho(\boldsymbol{u}) - \frac{1}{2} \log \left| \frac{\partial \boldsymbol{G}_y}{\partial \boldsymbol{u}} \frac{\partial \boldsymbol{G}_y}{\partial \boldsymbol{u}}^{\mathrm{T}} \right| - \log Z \tag{15}$$

where Z is the normalising constant for the density which is independent of $\boldsymbol{u}$.

In general evaluating the Gram matrix determinant $\log \left| \frac{\partial \boldsymbol{G}_y}{\partial \boldsymbol{u}} \frac{\partial \boldsymbol{G}_y}{\partial \boldsymbol{u}}^{\mathrm{T}} \right|$ has computational cost which scales as $\mathcal{O}(M N_y^2)$. However as part of the constrained dynamics updates the lower-triangular Cholesky decomposition $\boldsymbol{L}$ of the Gram matrix $\frac{\partial \boldsymbol{G}_y}{\partial \boldsymbol{u}} \frac{\partial \boldsymbol{G}_y}{\partial \boldsymbol{u}}^{\mathrm{T}}$ is calculated. Using basic properties of the matrix determinant we have

$$\log \pi(\boldsymbol{u}) = \log \rho(\boldsymbol{u}) - \frac{1}{2} \log \left| \boldsymbol{L}\boldsymbol{L}^{\mathrm{T}} \right| - \log Z \tag{16}$$

$$= \log \rho(\boldsymbol{u}) - \frac{1}{2} \log |\boldsymbol{L}| \left| \boldsymbol{L}^{\mathrm{T}} \right| - \log Z \tag{17}$$

$$= \log \rho(\boldsymbol{u}) - \log |\boldsymbol{L}| - \log Z \tag{18}$$

$$= \log \rho(\boldsymbol{u}) - \sum_{i=1}^{N_y} \log(L_{ii}) - \log Z \tag{19}$$

The base density $\rho(\boldsymbol{u})$ will typically be of a simple form e.g. standard Gaussian, therefore we can evaluate the logarithm of the target density up to an additive constant at a marginal computational cost that scales linearly with dimensionality.

For the gradient we can use reverse-mode automatic differentiation to calculate the gradient of (19) with respect to $\boldsymbol{u}$. This requires propagating partial derivatives through the Cholesky decomposition [3]; efficient implementations for this are present in many automatic differentiation frameworks including Theano.

Alternatively the gradient of (15) can be explicitly derived. The gradient of $\log \rho(\boldsymbol{u})$ will generally be trivial and $\frac{\partial \log Z}{\partial \boldsymbol{u}} = \boldsymbol{0}$. The gradient of the second term can be calculated using

$$\frac{\partial}{\partial u_i} \log \left| \frac{\partial \boldsymbol{G}_y}{\partial \boldsymbol{u}} \frac{\partial \boldsymbol{G}_y}{\partial \boldsymbol{u}}^{\mathrm{T}} \right| = \mathrm{Trace} \left\{ \left[\frac{\partial \boldsymbol{G}_y}{\partial \boldsymbol{u}} \frac{\partial \boldsymbol{G}_y}{\partial \boldsymbol{u}}^{\mathrm{T}} \right]^{-1} \left[\frac{\partial^2 \boldsymbol{G}_y}{\partial u_i \partial \boldsymbol{u}} \frac{\partial \boldsymbol{G}_y}{\partial \boldsymbol{u}}^{\mathrm{T}} + \frac{\partial \boldsymbol{G}_y}{\partial \boldsymbol{u}} \frac{\partial^2 \boldsymbol{G}_y}{\partial u_i \partial \boldsymbol{u}}^{\mathrm{T}} \right] \right\} \tag{20}$$

$$= 2\, \mathrm{Trace} \left\{ \frac{\partial^2 \boldsymbol{G}_y}{\partial u_i \partial \boldsymbol{u}} \left[\boldsymbol{L}^{-\mathrm{T}} \boldsymbol{L}^{-1} \frac{\partial \boldsymbol{G}_y}{\partial \boldsymbol{u}} \right]^{\mathrm{T}} \right\}. \tag{21}$$

The matrix inside the square brackets is independent of i and can be computed once by solving the system of equations by forward and backward substitution. The matrix of second partial derivatives $\frac{\partial^2 \boldsymbol{G}_y}{\partial u_i \partial \boldsymbol{u}}$ can either be manually derived for the specific generator function or calculated using automatic differentiation. The trace of the matrix product is then just the sum over all indices of the element-wise product of the pair.

C Exploiting structure in the generator

Often the generator inputs $\mathbf{u}$ can be split in to two distinct groups — global inputs $\mathbf{v}$ which effect all of the observed outputs (e.g. inputs which map to model parameters) and local 'noise' inputs $\mathbf{n}$, each element of which affect only a subset of the outputs.

In particular systems with a generator function $\boldsymbol{G}_y$ which can be expressed in one of the two forms

$$y_i = g_i(\boldsymbol{v}, n_i) \text{ (element-wise) or } y_i = \tilde{g}_i(\boldsymbol{v}, y_{i-1}, n_i) = g_i(\boldsymbol{v}, \boldsymbol{n}_{\leq i}) \text{ (autoregressive)} \tag{22}$$

have a Jacobian $\frac{\partial \boldsymbol{C}}{\partial \boldsymbol{u}} = \left[\frac{\partial \boldsymbol{C}}{\partial \boldsymbol{v}} \; \frac{\partial \boldsymbol{C}}{\partial \boldsymbol{n}} \right]$ in which $\frac{\partial \boldsymbol{C}}{\partial \boldsymbol{n}}$ is diagonal (element-wise) or triangular (autoregressive).

The decomposition of $\frac{\partial \boldsymbol{C}}{\partial \boldsymbol{u}} \frac{\partial \boldsymbol{C}}{\partial \boldsymbol{u}}^{\mathrm{T}} = \frac{\partial \boldsymbol{C}}{\partial \boldsymbol{n}} \frac{\partial \boldsymbol{C}}{\partial \boldsymbol{n}}^{\mathrm{T}} + \frac{\partial \boldsymbol{C}}{\partial \boldsymbol{v}} \frac{\partial \boldsymbol{C}}{\partial \boldsymbol{v}}^{\mathrm{T}}$ can then be computed by low-rank Cholesky updates of the triangular / diagonal matrix $\frac{\partial \boldsymbol{C}}{\partial \boldsymbol{n}}$ with each of the columns of $\frac{\partial \boldsymbol{C}}{\partial \boldsymbol{v}}$. As $\dim(\boldsymbol{v}) = L$ is often significantly less than, and independent of, the number of outputs conditioned on N_y, the resulting $\mathcal{O}(L N_y^2)$ cost of the Cholesky updates is a significant improvement over the original $\mathcal{O}(N_y^3)$.

Many learnt differentiable generative models have an element-wise noise structure including the Gaussian VAE. The autoregressive noise structure commonly occurs in stochastic dynamical simulations where the outputs are a time sequence of states, with noise being added each time-step, for example the Lotka-Volterra model considered in the experiments in Section 7.

D Lotka-Volterra parameter empirical histogram

Larger version of figure 2b showing empirical histograms for posterior samples of Lotka–Volterra model parameters.

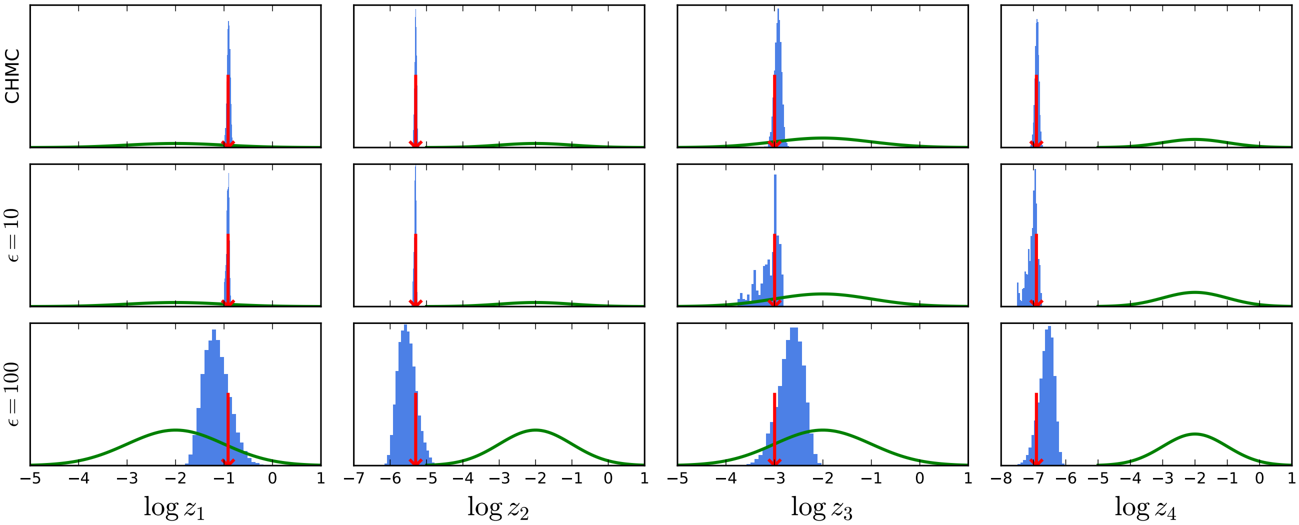

Figure 1: Marginal empirical histograms for the (logarithm of the) four parameters (columns) from constrained HMC samples (top) and ABC samples with $\epsilon = 10$ (middle) and $\epsilon = 100$ (bottom). Horizontal axes shared across columns. Red arrows indicate true parameter values. Green curve - log-normal prior density.

References

[1] P. Diaconis, S. Holmes, and M. Shahshahani. Sampling from a manifold. In *Advances in Modern Statistical Theory and Applications*, pages 102–125. Institute of Mathematical Statistics, 2013.

[2] H. Federer. *Geometric measure theory*. Springer, 2014.

[3] I. Murray. Differentiation of the Cholesky decomposition. *arXiv preprint arXiv:1602.07527*, 2016.