A. Measure-Theoretic Formalization of Stochastic Conditioning

Although stochastic conditioning is defined in terms of the density q of the distribution D, its key idea does not depend on q. In fact, we have already explained informally how stochastic conditioning and our results can be developed even when the density q does not exist, as in the case of Dirac distributions. Also, Proposition 2 assumes this general development. In this section, we spell out this informal explanation, and describe the measure-theoretic formalization of stochastic conditioning.

We start by changing Definitions 1 and 2 such that D is not required to have a density with respect to the Lebesgue measure, and the conditional density $p(y{\sim}D|x)$ is defined for such D.

Definition 3. *A probabilistic model with stochastic conditioning is a tuple $(p(x,y), D)$ where*

- *$p(x,y)$ is the joint probability density of random variable x and observation y, and it is factored into the product of the prior $p(x)$ and the conditional probability $p(y|x)$ (i.e., $p(x,y) = p(x)p(y|x)$);*
- *D is the distribution (i.e., probability measure) from which observation y is sampled.*

Definition 4. *The conditional density $p(y{\sim}D|x)$ of D given x is*

$$p(y{\sim}D|x) = \exp\left(\int_Y (\log p(y|x))\, D(dy)\right) \tag{19}$$

where $D(dy)$ indicates that the integral over Y is taken with respect to the distribution D.

To explain where the term "density" in Definition 4 comes from, we recall the standard setup of the work on random distributions, which studies distributions over distributions.[2] The setup over random distributions on $Y \subseteq \mathbb{R}^m$ is the measurable space $(\mathcal{D}, \Sigma)$ where $\mathcal{D}$ is the set of distributions over Y and Σ is the smallest σ-field generated by the family

$$\Big\{\{D \mid D(A) < r\} \;\Big|\; \text{measurable } A \subseteq Y \text{ and } r \in \mathbb{R}\Big\}.$$

The next theorem generalizes Theorem 1. In a setting that covers both continuous and discrete cases, with or without densities, the theorem describes when $p(y{\sim}D|x)$ has a finite normalization constant.

Theorem 2. *Assume that we are given a distribution D_θ parameterized by $\theta \in \Theta \subseteq \mathbb{R}^p$ such that D is a probability kernel from Θ to Y, and the following μ_x is a well-defined unnormalized distribution (i.e., measure) over Θ: for all measurable subsets B of Θ,*

$$\begin{aligned}\mu_x(B) &= \int_B p(y{\sim}D_\theta|x)\, d\theta \\ &= \int_B \exp\left(\int_Y (\log p(y|x))\, D_\theta(dy)\right) d\theta.\end{aligned}$$

Let ν_x be the push-forward of μ_x along the function $\theta \longmapsto D_\theta$ from Θ to $\mathcal{D}$. The unnormalized distribution ν_x has a finite normalization constant C (i.e., $\nu_x(\mathcal{D}) = C < \infty$) if there exists $C' < \infty$ such that for all measurable subsets A of Y,

$$\int_\Theta D_\theta(A)\, d\theta \le \left(C' \cdot \int_A dy\right). \tag{20}$$

Before proving the theorem, we make two comments. First, when D_θ is defined in terms of a density q_θ, the condition (20) in the theorem is implied by the condition in Theorem 1:

$$\sup_{y\in Y} \int_\Theta q_\theta(y) d\theta \le C'.$$

The implication is shown below:

$$\begin{aligned}\int_\Theta D_\theta(A)\, d\theta &= \int_\Theta \int_A q_\theta(y) dy d\theta = \int_A \int_\Theta q_\theta(y) d\theta dy \\ &\le \int_A \sup_{y'\in Y} \left(\int_\Theta q_\theta(y') d\theta\right) dy \le \left(C' \cdot \int_A dy\right).\end{aligned}$$

Second, when λ_x is the push-forward of the Lebesgue measure along $\theta \longmapsto D_\theta$, our $p(y{\sim}D|x)$ is the density of ν_x with respect to λ_x. This is why we called $p(y{\sim}D|x)$ conditional density.

Proof. The theorem claims that $C = \nu_x(\mathcal{D})$ is finite. But $C = \mu_x(\Theta)$ by the definition of the push-forward measure, and so it suffices to show the finiteness of $\mu_x(\Theta)$. Note

$$\mu_x(\Theta) = \int_\Theta \exp\left(\int_Y (\log p(y|x)) D_\theta(dy)\right) d\theta. \tag{21}$$

We compute a finite bound of $C = \mu_x(\Theta)$ as follows:

$$\begin{aligned}C &\le^1 \int_\Theta \int_Y \Big(\exp(\log p(y|x))\Big)\, D_\theta(dy) d\theta \\ &= \int_\Theta \int_Y p(y|x)\, D_\theta(dy) d\theta \\ &=^2 \left(C' \cdot \int_Y p(y|x)\, dy\right) = C' < \infty\end{aligned} \tag{22}$$

where $\le^1$ is by Jensen's inequality and $=^2$ uses the assumption of the theorem. □

[2] A brief yet good exposition on this topic can be found in Appendix A of Ghosal & van der Vaart (2017).

Besides $\{\mathrm{Normal}(\theta, 1) \mid \theta \in \mathbb{R}\}$ that we discussed already after Theorem 1, the set $\{\mathrm{Dirac}(\theta) \mid \theta \in \mathbb{R}\}$ satisfies the condition (20) in Theorem 2. Thus, $p(y{\sim}D|x)$ can be normalized to a distribution (i.e., a probability measure) in both cases. However, $p(y{\sim}D|x)$ cannot be normalized over the space $\{\beta \cdot \mathrm{Dirac}(0) + (1-\beta) \cdot \mathrm{Dirac}(\theta) \mid \theta \in \mathbb{R}\}$ for $\beta \in (0, 1)$, which consists of the mixtures of two Dirac distributions. The condition (20) in Theorem 2 does not hold. In fact, if $p(0|x) > 0$, the normalization constant of ν_x in the theorem is infinite.

B. Inference algorithms

A simple **bias-adjusted likelihood estimate** $\hat{p}(x, y{\sim}D)$, required for the computation of the weights in importance sampling as well as of the acceptance ratio in pseudo-marginal Markov chain Monte Carlo (Andrieu & Roberts, 2009), can be computed based on (9) as follows (Ceperley & Dewing, 1999; Nicholls et al., 2012; Quiroz et al., 2018). Under the conditions of the central limit theorem, the distribution of

$$\frac{1}{N}\sum_{j=1}^{N} \log p(x, y_j)$$

becomes similar to the normal distribution

$$\mathrm{Normal}\Big(\mu = \mathop{\mathbb{E}}_{y\sim D}[\log p(x,y)], \sigma^2 = \frac{1}{N}\mathop{\mathbb{V}\mathrm{ar}}_{y\sim D}[\log p(x,y)]\Big)$$

as $N \to \infty$. Correspondingly, the distribution of

$$\exp\left(\frac{1}{N}\sum_{j=1}^{N} \log p(x, y_j)\right)$$

and the log-normal distribution with the same parameters become similar under the same asymptotics. But the mean of the log-normal distribution is $\exp(\mu + \frac{\sigma^2}{2})$. Thus, we can construct a bias-adjusted estimate as

$$\begin{aligned} m &= \frac{1}{N}\sum_{j=1}^{N} \log p(x, y_j), \\ s^2 &= \frac{1}{N-1}\sum_{j=1}^{N} (\log p(x, y_j) - m)^2, \\ \hat{p}(x, y{\sim}D) &= \exp(\mu) \qquad (23) \\ &\approx \mathbb{E}_{y_{1:N}\sim D^n}\left[\exp\left(\frac{1}{N}\sum\nolimits_{j=1}^{N} \log p(x, y_j)\right)\right] \\ &\quad \times \exp\left(-\frac{\sigma^2}{2}\right) \\ &\approx \exp\left(m - \frac{s^2}{2N}\right). \end{aligned}$$

In **importance sampling**, x_i's are drawn from a proposal distribution U with probability mass or density $u(x)$ and weighted by the joint probability mass or density of x and observations. In the case of stochastic conditioning, the weight w_i of x_i is approximated as $\hat{w}_i$ using an unbiased estimate $\hat{p}(x_i, D)$ such as (23).

$$\hat{w}_i = \frac{\hat{p}(x_i, D)}{u(x_i)} = \frac{1}{u(x_i)} \exp\left(m_i - \frac{s_i^2}{2N}\right). \qquad (24)$$

Markov chain Monte Carlo algorithms are broadly applied to inference in probabilistic programs, with Lightweight Metropolis-Hastings (Wingate et al., 2011b) as the simplest and universally applicable variant. Many MCMC variants involve proposing a new state x' from a proposal distribution U with probability mass or density $u(x'|x)$ and then either accepting x' or retaining x, with Metropolis-Hastings acceptance ratio α based on the joint probability of x' and observations:

$$\alpha = \min\left\{1,\ \frac{u(x|x')}{u(x'|x)} \times \frac{p(x', y{\sim}D)}{p(x, y{\sim}D)}\right\}. \qquad (25)$$

Just like with importance sampling, $p(x, y{\sim}D)$ cannot be computed exactly for probabilistic programs with stochastic conditioning. However, Andrieu & Roberts (2009) establish that the joint probability can be replaced with an unbiased estimate without affecting the stationary distribution of the Markov chain, resulting in *pseudo-marginal* MCMC. Pseudo-marginal MCMC allows speeding up Monte Carlo inference by subsampling (Bardenet et al., 2017; Quiroz et al., 2019; Dang et al., 2019; Quiroz et al., 2018) and can be applied to stochastic conditioning as well. The main challenge in designing an efficient MCMC algorithm, for both subsampling and stochastic conditiong, is constructing an unbiased low-variance estimate of the joint probability. In a basic case, (23) can be used as a bias-adjusted estimate, resulting in the acceptance ratio $\hat{\alpha}$:

$$\begin{aligned} \hat{\alpha} &= \min\left\{1,\ \frac{u(x|x')}{u(x'|x)} \times \frac{\hat{p}(x', D)}{\hat{p}(x, y{\sim}D)}\right\} \\ &= \min\left\{1,\ \frac{u(x|x')}{u(x'|x)} \times \exp\left(m' - m - \frac{s'^2 - s^2}{2N}\right)\right\}. \end{aligned} \qquad (26)$$

Note that the same samples $y_1, y_2, ..., y_N$ should be used for estimating both m, s^2 and m', s'^2 (Andrieu & Roberts, 2009).

Stochastic gradient Markov chain Monte Carlo (sgMCMC) (Ma et al., 2015) can be used unmodified when the log probability is differentiable with respect to x. sgMCMC uses an unbiased stochastic estimate of the gradient of log probability density. Such estimate is trivially obtained by drawing a single sample y_1 from D and computing the

gradient of the log joint density of x and y:

$$\begin{aligned}\nabla_x \log p(x, y{\sim}D) &= \nabla_x \Big(\log \Big(p(x) \prod_{y \in Y} p(y|x)^{q(y)dy} \Big) \Big) \\ &= \nabla_x \Big(\log \Big(\prod_{y \in Y} p(x,y)^{q(y)dy} \Big) \Big) \\ &= \nabla_x \int_{y \in Y} q(y) \log p(x,y) dy \\ &= \int_{y \in Y} q(y) \Big(\nabla_x \log p(x,y) \Big) dy \\ &\approx \nabla_x \log p(x, y_1).\end{aligned} \tag{27}$$

Stochastic variational inference (Hoffman et al., 2013; Ranganath et al., 2014; Kucukelbir et al., 2017) requires a noisy estimate of the gradient of the evidence lower bound (ELBO) $\mathcal{L}$. The most basic approach is to use the score estimator that is derived from the following equation:

$$\nabla_\lambda \mathcal{L} = \mathbb{E}_{x \sim q(x|\lambda)} \left[(\nabla_\lambda \log q(x|\lambda)) \left(\log \frac{p(x, y{\sim}D)}{q(x|\lambda)} \right) \right]. \tag{28}$$

As in the standard posterior inference setting, maximizing ELBO is equivalent to minimizing the KL divergence from $q(x|\lambda)$ to $p(x|D)$. Substituting (3) into (28), we obtain

$$\begin{aligned}\nabla_\lambda \mathcal{L} &= \mathbb{E}_{x \sim q(x|\lambda)} \Big[\nabla_\lambda \log q(x|\lambda) \Big(\log p(x) + \\ &\qquad \int_{y \in Y} q(y) \log p(y|x) dy - \log q(x|\lambda) \Big) \Big] \\ &= \mathbb{E}_{x \sim q(x|\lambda)} \Big[\int_{y \in Y} \nabla_\lambda \log q(x|\lambda) \Big(\log p(x) + \\ &\qquad \log p(y|x) - \log q(x|\lambda) \Big) q(y) dy \Big] \\ &= \mathbb{E}_{(x,y) \sim q(x|\lambda) \times D} \Big[\nabla_\lambda \log q(x|\lambda) \Big(\log p(x) + \\ &\qquad \log p(y|x) - \log q(x|\lambda) \Big) \Big].\end{aligned} \tag{29}$$

Thus, $\nabla_\lambda \mathcal{L}$ can be estimated using Monte Carlo samples $x_s, y_s \sim q(x|\lambda) \times D$:

$$\begin{aligned}\nabla_\lambda \mathcal{L} \approx \frac{1}{S} \sum_{s=1}^{S} \nabla_\lambda \log q(x_s|\lambda) \Big(\log p(x_s) + \\ \log p(y_s|x_s) - \log q(x_s|\lambda) \Big),\end{aligned} \tag{30}$$

and stochastic variational inference can be directly applied. In fact, van de Meent et al. (2016) use black-box variational inference (Ranganath et al., 2014) for a special case of stochastic conditioning arising in policy search.