INNF+ Workshop 2020

Phillip Urquijo, *Justin Tan*

Anomaly Detection w/ Importance Sampling

July 16, 2020

THE UNIVERSITY OF MELBOURNE

Density Estimation & Latent Variables

- Want to model $p_\theta(\mathbf{x}) \approx p^*(\mathbf{x})$.
  - $p^*(\mathbf{x})$: true data density.
  - $p_\theta(\mathbf{x})$: density under (maybe generative) model parameterized by θ
  - $\min_\theta D_{\mathrm{KL}}\left(p^* \| p_\theta\right) \Leftrightarrow \max_\theta \log p_\theta(\mathbf{x})$.
- Introduce 'hidden' variables $\mathbf{z}$, joint density $p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})$.
- Log-density via marginalization:

$$\log p_\theta(\mathbf{x}) = \log \int_{\mathbf{z}} d\mathbf{z}\, p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z}). \tag{1}$$

- Problem: quadrature takes exponential time in $\dim(\mathbf{z})$, likely non-analytic.

Density Estimation & Latent Variables

Introduce proposal/variational, distribution $q(\mathbf{z};\mathbf{x})$ with efficient sampling + evaluation. Get lower bound on the log-model evidence:

$$\begin{aligned}
\log p_\theta(\mathbf{x}) &= \log \int_{\mathbf{z}} d\mathbf{z}\, p(\mathbf{x}|\mathbf{z})p(\mathbf{z})\frac{q(\mathbf{z};\mathbf{x})}{q(\mathbf{z};\mathbf{x})} \\
&= \log \mathbb{E}_{q(\mathbf{z};\mathbf{x})}\left[\frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{q(\mathbf{z};\mathbf{x})}\right] \\
&\geq \mathbb{E}_{q(\mathbf{z};\mathbf{x})}\left[\log \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{q(\mathbf{z};\mathbf{x})}\right] \triangleq \mathcal{L}(\psi)
\end{aligned}$$

$\mathcal{L}(\psi)$ can be efficiently optimized as a proxy for $p_\theta(\mathbf{x})$ - standard ELBO objective, but we can do better ...

Density Estimation & Latent Variables

Let $\mathbf{E}_q\left[\psi(\mathbf{z})\right] = p_\theta(\mathbf{x})$. Then:

$$\left|\log p_\theta(\mathbf{x}) - \mathbf{E}_q\left[\log\psi\right]\right| \approx \frac{1}{2p_\theta(x)^2}\mathbb{V}\left[\psi\right] \tag{2}$$

(Intuition: Taylor expand $\log\psi$ in small quantity $\Delta = p_\theta(\mathbf{x}) - \psi(\mathbf{x})$).

- Tightness of bound scales with estimator variance.
- Simple way to reduce variance is to consider the sample mean of ψ:

$$\hat{\psi}_K = \frac{1}{K}\sum_k \psi_k = \frac{1}{K}\sum_k \frac{p(\mathbf{x}|\mathbf{z}_k)p(\mathbf{z}_k)}{q(\mathbf{z}_k;\mathbf{x})}; \quad \mathbf{z}_k \sim q(\mathbf{z};\mathbf{x}) \tag{3}$$

Density Estimation & Latent Variables

- Tighten bound by considering importance-sampling estimate (Burda, 2016). $K = 1$ is standard ELBO.
- Tradeoff compute for tightness.

$$\log p(\mathbf{x}) \geq \mathbf{E}_{q(\mathbf{z};\mathbf{x})}\left[\log \frac{1}{K}\sum_{k=1}^{K}\frac{p(\mathbf{x}|\mathbf{z}_k)p(\mathbf{z}_k)}{q(\mathbf{z}_k|\mathbf{x})}\right] \triangleq \mathcal{L}_K(q) \tag{4}$$

$$\geq \mathbf{E}_{q(\mathbf{z};\mathbf{x})}\left[\log \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{q(\mathbf{z};\mathbf{x})}\right] \tag{5}$$

Use of $\log p_\theta(\mathbf{x})$ as anomaly detection score extensively researched (e.g. Nalisnick, 2018).

Density Estimation & Latent Variables

- Use RD LHC Olympics Dataset. Same experimental setup as in (Nachman/Shih, 2020).
  - 1/2 train-test split.
  - 1000 injected signal events.
  - Apply smoothing transformation $g : [0,1]^D \to \mathbb{R}^D, g(\mathbf{x}) = \mathrm{Logit}\left(\epsilon + (1-\epsilon)\mathbf{x}\right)$ (Dinh, 2017).
- Features:
  - M_{J_1}: The invariant mass of leading jet.
  - $\Delta M_{J_1,J_2}$: The difference in invariant mass between the leading, subleading jets.
  - $\tau_{21}^{J_1}, \tau_{21}^{J_2}$: The $2,1$-subjettiness ratio for the leading, subleading jets, respectively.

log $p_\theta(\mathbf{x})$ scoring

Model	BPD	Signal BPD	Bkg. BPD	$\lvert\Delta_{\mathrm{BPD}}\rvert$	AUC
$\mathcal{L}_{16}(q)$	-0.739 ± 0.001	-0.076 ± 0.007	-0.743 ± 0.001	0.667 ± 0.007	0.814 ± 0.002
SUMO	-0.748 ± 0.001	-0.149 ± 0.006	-0.751 ± 0.001	0.602 ± 0.006	0.818 ± 0.004
$\mathcal{L}_{1024}(q)$	-0.742 ± 0.001	-0.067 ± 0.007	-0.746 ± 0.001	$\mathbf{0.679} \pm \mathbf{0.007}$	0.815 ± 0.002
FFJORD	-0.744 ± 0.001	-0.060 ± 0.018	-0.747 ± 0.001	$\mathbf{0.687} \pm \mathbf{0.017}$	0.817 ± 0.003
Real-NVP	-0.747 ± 0.000	-0.136 ± 0.001	-0.750 ± 0.000	0.614 ± 0.001	$\mathbf{0.826} \pm \mathbf{0.001}$
MAF	-$\mathbf{0.758} \pm \mathbf{0.001}$	-0.215 ± 0.001	-0.761 ± 0.000	0.547 ± 0.000	0.807 ± 0.001

Model performance using log $p(\mathbf{x})$ as a scoring function, reported in the average Bits per Dimension for all events (BPD), for signal and background events (Signal BPD and Bkg. BPD, respectively) over the test dataset, and the area under the curve (AUC) obtained by thresholding log $p(\mathbf{x})$. For the purposes of density modelling, lower BPD is better.

$\lambda(\mathbf{x}) = \frac{p_\theta(\mathbf{x}|\text{bkg-only})}{p_\theta(\mathbf{x}|\text{data})}$ scoring

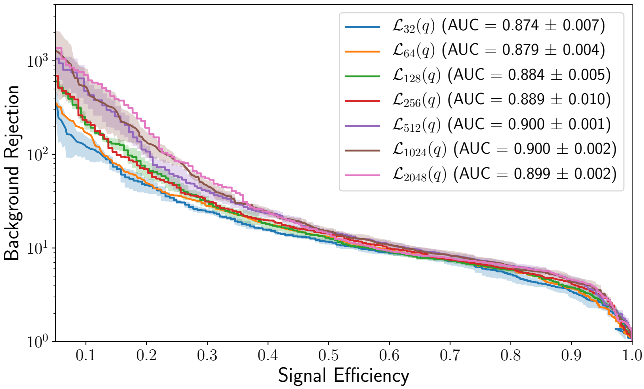


Figure 1: Background rejection versus signal efficiency curves over the signal region in the test set for the $\mathcal{L}_K(q)$ estimator with different number of importance samples K. Results are given as the mean and standard deviation over 3 trials with different random seeds. The uncertainty bands show the 1σ deviation from the mean.

Studentizing

- Well-known that importance-sampling density should have heavy-tails (Mackay, 2003).
- Instead of having $q(\mathbf{z}; \mathbf{x})$ Gaussian-distributed, consider Student's t-distribution.
- Degrees of Freedom parameter ν controls tail heaviness.

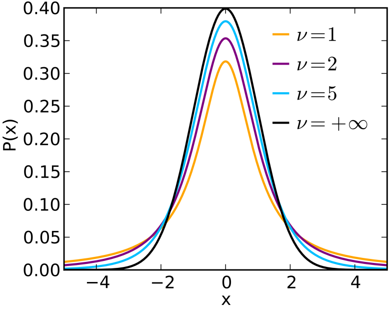


Figure 2: (Wikipedia, 2020)

Studentizing

Differentiable reparameterization:

- Gaussian reparameterization:

$$\mathbf{z} \sim \mathcal{N}(\mu, \Sigma) \leftarrow \quad \mathbf{z} = \mu + L^{*}\epsilon, \quad \epsilon \sim \mathcal{N}(0, 1)$$

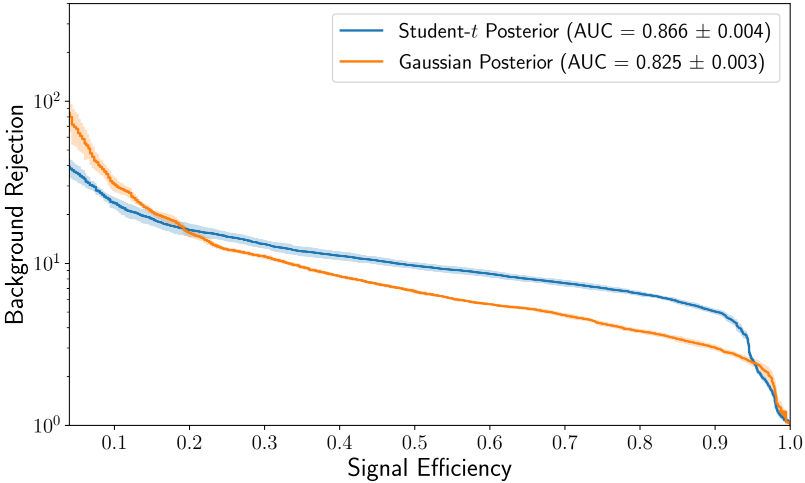
.

- Student's t reparameterization:

$$\mathbf{z} \sim \mathrm{St}(\mu, \Sigma, \nu) \leftarrow \quad \mathbf{z} = \mu + \frac{L^{*}\epsilon}{\sqrt{u/\nu}}, \quad \epsilon \sim \mathcal{N}(0, 1), u \sim \mathrm{Gamma}(\nu/2, 1/2)$$

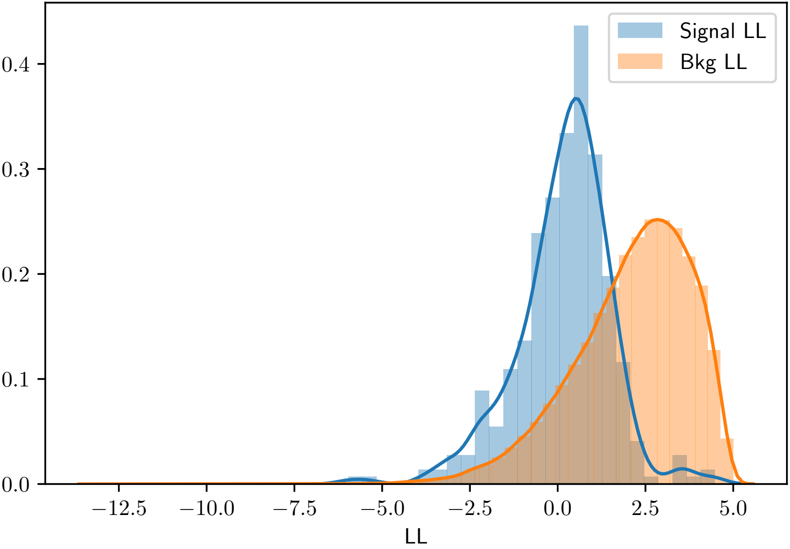
- Output distributional parameters (μ, ν, Σ) vs. (μ, Σ).
- Restrict $\nu > 1$, otherwise random samples are extremely large.

Studentizing



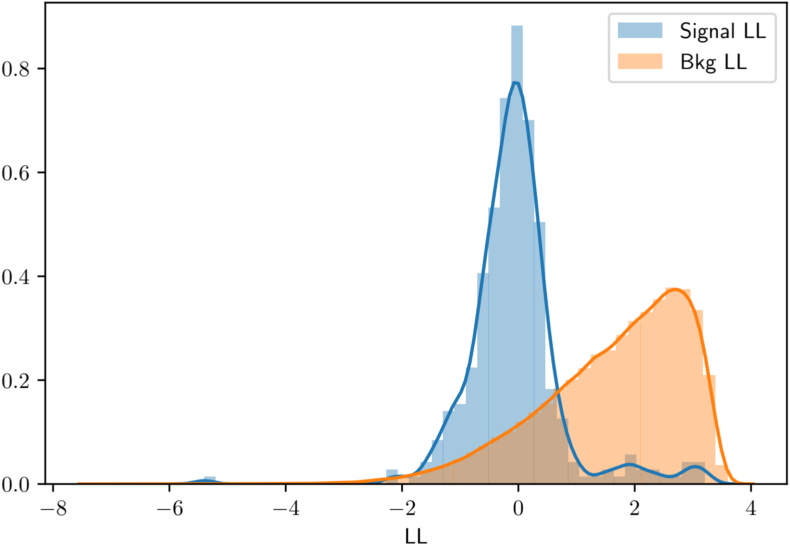
Studentizing

Gaussian-VAE.



Studentizing

Student's-t-VAE.



Future Work

- See if superior density modelling translates to better anomaly detection (preliminary results supportive.)
- Look at more expressive posterior distributions given by normalizing flows.
- Investigate scaling to higher-dimensional feature set (preliminary results non-supportive.)

Bayes' Theorem

- Conditioning on the known value of data x yields Bayes' theorem:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

  - The likelihood $p(y|\theta)$ is the conditional probability of the data y given fixed θ.
  - The prior $p(\theta)$ represents information we have that is not part of the collected data y.
  - The evidence $p(y)$ is the average over all possible values of θ.

$$p(y) = \sum_{\theta} p(y|\theta)p(\theta)$$

- $p(\theta|y)$ is the posterior distribution, which represents our updated beliefs under our prior $p(\theta)$ now we have observed the data y.

Bayes' Theorem

- Alternatively, observe the effect of some unknown cause. Wish to determine the cause:

$$p(\text{Cause}|\text{Effect}) = \frac{p(\text{Effect}|\text{Cause})p(\text{Cause})}{p(\text{Effect})}$$

  - The likelihood $p(\text{Effect}|\text{Cause})$ describes the relationship in the causal direction.
  - Computing the posterior $p(\text{Cause}|\text{Effect})$ allows us to *diagnose potential causes*.

Bayes' Theorem

- We have a set of given hypotheses $\{H_1, \dots H_n\}$, corresponding to different values of θ.
- Want to find most likely hypothesis given the data we have collected so far.
- Look at ratio of posterior density at different points θ, θ' corresponding to hypotheses H, H':

$$\frac{p(\theta|y)}{p(\theta'|y)} = \frac{p(\theta)p(y|\theta)}{p(\theta')p(y|\theta')}$$

- This allows us to bypass calculation of the (potentially intractable) evidence $p(y)$.

Naive Bayes

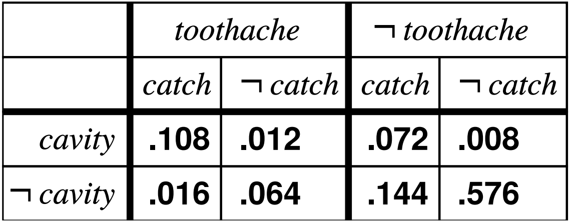
- For n possible boolean evidence variables there are 2^n possible combinations of conditional probabilities we need to know.
- Conditional independence of two variables X, Y given a third Z allows us to use only a reasonable number of combinations.

$$P(X, Y|Z) = P(X|Z)P(Y|Z)$$

- For n effects that are all conditionally independent given the cause, the representation is $\mathcal{O}(n)$ instead of $\mathcal{O}(2^n)$.
- If a single cause is the direct cause of a number of effects, all of which are conditionally independent, then the full joint distribution is:

$$P(\text{Cause}, \text{Effect}_1, \ldots \text{Effect}_n) = P(\text{Cause}) \prod_i^n P(\text{Effect}_i|\text{Cause})$$

Marginalization



	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ cavity	**.016**	**.064**	**.144**	**.576**

- For any proposition ϕ:

$$P(\phi) = \sum_{\{\omega : \phi(\omega) = \text{True}\}} P(\omega)$$

- More generally, find the distribution of ϕ by averaging all possible values of $P(\phi|x)$:

$$P(\phi) = \sum_x P(\phi|x)P(x)$$

Bayes' Rule

- We know the test reports positive, want to find the posterior probability of actual Leckieitis with this knowledge.

Bayes' Rule

- We know the test reports positive, want to find the posterior probability of actual Leckieitis with this knowledge.
- Probability of contracting Leckieitis: $p(L) = 10^{-4}$.

Bayes' Rule

- We know the test reports positive, want to find the posterior probability of actual Leckieitis with this knowledge.
- Probability of contracting Leckieitis: $p(L) = 10^{-4}$.
- Probability that the test is positive, given patient has Leckieitis: $p(\text{Test} = +|L) = 0.99$

Bayes' Rule

- We know the test reports positive, want to find the posterior probability of actual Leckieitis with this knowledge.
- Probability of contracting Leckieitis: $p(L) = 10^{-4}$.
- Probability that the test is positive, given patient has Leckieitis: $p(\text{Test} = +|L) = 0.99$
- Probability that the test is positive:

$$
\begin{aligned}
p(\text{Test} = +) &= p(\text{Test} = +|L)p(L) + p(\text{Test} = +|\neg L)p(\neg L) \\
&= 0.99 \times 10^{-4} + 0.01 \times (1 - 10^{-4}) \\
&= 0.0098
\end{aligned}
$$

Bayes' Rule

- We know the test reports positive, want to find the posterior probability of actual Leckieitis with this knowledge.
- Probability of contracting Leckieitis: $p(L) = 10^{-4}$.
- Probability that the test is positive, given patient has Leckieitis: $p(\text{Test} = +|L) = 0.99$
- Probability that the test is positive:

$$\begin{aligned} p(\text{Test} = +) &= p(\text{Test} = +|L)p(L) + p(\text{Test} = +|\neg L)p(\neg L) \\ &= 0.99 \times 10^{-4} + 0.01 \times (1 - 10^{-4}) \\ &= 0.0098 \end{aligned}$$

- Probability of having Leckieitis given the test is positive:

$$\begin{aligned} p(L|\text{Test} = +) &= \frac{p(\text{Test} = +|L)p(L)}{p(\text{Test} = +)} \\ &= 9.8 \times 10^{-3} \end{aligned}$$

Bayes' Rule

- Discrimination between blue/green taxis is 75% reliable, and you observed a blue taxi.
- Probability that the actual color is blue, given you observed blue:

$$p(\text{Actual} = \textcolor{blue}{\bullet} | \text{Observed} = \textcolor{blue}{\bullet}) = \frac{p(\text{Observed} = \textcolor{blue}{\bullet} | \text{Actual} = \textcolor{blue}{\bullet}) p(\text{Actual} = \textcolor{blue}{\bullet})}{p(\text{Observed} = \textcolor{blue}{\bullet})}$$

- Probability that the actual color is green, given you observed blue:

$$p(\text{Actual} = \textcolor{green}{\bullet} | \text{Observed} = \textcolor{blue}{\bullet}) = \frac{p(\text{Observed} = \textcolor{blue}{\bullet} | \text{Actual} = \textcolor{green}{\bullet}) p(\text{Actual} = \textcolor{green}{\bullet})}{p(\text{Observed} = \textcolor{blue}{\bullet})}$$

Bayes' Rule

- Want to find the ratio of both posteriors, or odds-ratio:

$$O = \frac{p(\text{Observed} = \text{🔵} | \text{Actual} = \text{🔵}) p(\text{Actual} = \text{🔵})}{p(\text{Observed} = \text{🔵} | \text{Actual} = \text{🟢}) p(\text{Actual} = \text{🟢})}$$

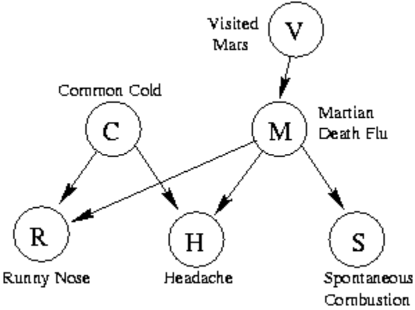
- Using the fact that $p(\text{Actual} = \text{🔵} | \text{Observed} = \text{🔵}) = 0.75$:

$$O = \frac{3p(\text{Actual} = \text{🔵})}{p(\text{Actual} = \text{🟢})}$$

- If we know the *prior probabilities*: $p(\text{Actual} = \text{🟢}) = 9p(\text{Actual} = \text{🔵})$, then we can incorporate this into our calculation of the posterior ratio to find that, while you swear that the taxi is blue, being struck by a green taxi is still 3 times more likely.

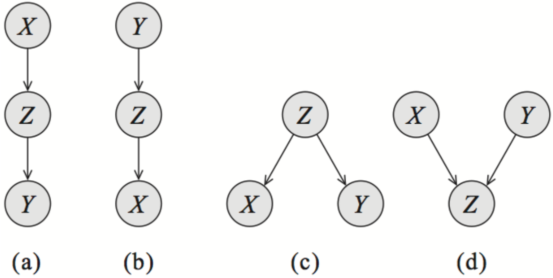
Bayesian Networks

- Full joint probability distribution specifies probability of each assignment of values to random variables. For n variables there are 2^n entries.
- Conditional independence between effect variables, given a cause variable, allows factorization of the full joint distribution into smaller conditional distributions.
- Bayesian Networks are a compact representation of the full joint distribution that shows dependencies between variables graphically.



Bayesian Networks

- Vertices correspond to random variables.
- Edges between vertices, e.g. $X \rightarrow Y$ indicates X has a direct influence on Y. Causes should be parents of effects.
- Each vertex has a conditional probability distribution summarizing effects of parents on the random variable $P(X_i|\text{Parents}(X_i))$.



Bayesian Networks

- Chain rule allows decomposition of joint into conditionals:

$$P(x_1, x_2, \ldots, x_n) = p(x_1)p(x_2|x_1)\ldots p(x_n|x_{n-1}, \ldots, x_2, x_1)$$

- Via conditional independence, each random variable x_i only directly depends on a small number of variables: Parents(x_i).

$$P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|\text{Parents}(x_i))$$

- If each variable has d possible values and at most k parents, then the joint distribution has $\mathcal{O}(nd^k)$ entries (versus $\mathcal{O}(d^n)$).
- e.g. 20 random variables, each with 5 parents, then the Bayesian network approach uses 640 random variables versus over 10^6 for the full joint.

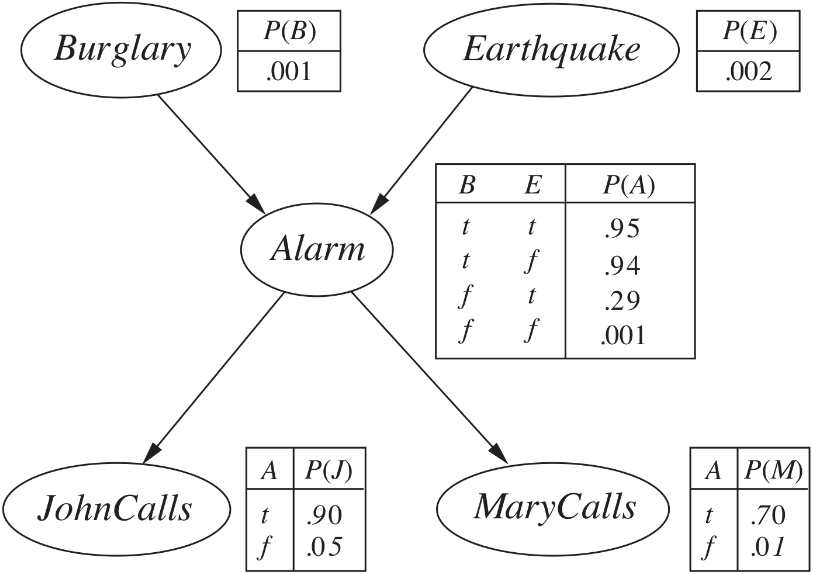
Bayesian Inference

- In the context of Bayesian networks, compute posterior $P(X|\mathbf{e})$ for a query X (some assignment of random variables) given observed event $\mathbf{e}$ (assignment to a set of evidence variables).
- 'Find probability of X, given we know $\mathbf{e}$ has occurred.'
  - Let $\mathbf{H}$ denote all variables outside X, $\mathbf{e}$ (call $\mathbf{H}$ hidden variables), let Z be the evidence (i.e. normalizing constant). From Bayes' Theorem:

$$\begin{aligned} P(X|\mathbf{e}) &= \frac{1}{Z} P(X, \mathbf{e}) \\ &= \frac{1}{Z} \sum_{\mathbf{H}} P(X, \mathbf{e}|\mathbf{H}) p(\mathbf{H}) \\ &= \frac{1}{Z} \sum_{\mathbf{H}} P(X, \mathbf{e}, \mathbf{H}) \end{aligned}$$

- To solve a Bayesian inference problem:
  - Identify query, evidence and hidden variables.
  - Decompose joint distribution using Bayesian network structure into product of simpler conditional distributions.
  - Fix evidence variables to observed values.
  - Sum (marginalize) over remaining hidden variables $\mathbf{H}$.

Bayesian Inference



$P(B)$
.001

$P(E)$
.002

B	E	$P(A)$
t	t	.95
t	f	.94
f	t	.29
f	f	.001

A	$P(J)$
t	.90
f	.05

A	$P(M)$
t	.70
f	.01

Bayesian Inference

- If A observed, B and E are no longer independent! Knowledge of A couples the parent variables. This is an example of a V-structure.
- Parents are independent if child is unobserved, but coupled when child is observed.
- Simpler example: suppose your lawn is wet in the morning (C). A (rain) and B (sprinkler) are two possible causes for it being wet. If we know C is true and A is false, then B must be true. i.e. A and B are not conditionally independent given C.

Sequential Bayesian Updates

- Bayes' Theorem allows us to update our uncertainty as new information is acquired.
- Hypothesis H; observe a series of independent measurements $\{x_1, x_2, \dots x_T\}$.
- How does our uncertainty about H evolve given these observations?

Sequential Bayesian Updates

- Given sequential measurements $\{x_1, x_2, \dots x_T\}$, our likelihood at time t summarizes the probability of the data given the hypothesis H:

  $$p(x_1, \dots x_t | H) = p(x_1|H)p(x_2|x_1, H) \dots p(x_t|\mathbf{x}_{t-1}, H)$$

  Where we let $\mathbf{x}_n = (x_1, x_2, \dots, x_n)$.

- Bayes' Theorem:

  $$p(H|\mathbf{x}_t) \propto p(\mathbf{x}_t|H)p(H)$$

- At time $t+1$, the posterior is:

  $$p(H|\mathbf{x}_{t+1}) \propto p(\mathbf{x}_{t+1}|H)p(H)$$

- How to get from $P(H|\mathbf{x}_t)$ to $P(H|\mathbf{x}_{t+1})$?

Sequential Bayesian Updates

- Use the chain rule:

$$
\begin{aligned}
p(H|\mathbf{x}_{t+1}) &\propto p(\mathbf{x}_{t+1}|H)p(H) \\
&= p(x_{t+1}, \mathbf{x}_t|H)p(H) \\
&= p(x_{t+1}|\mathbf{x}_t, H)p(\mathbf{x}_t|H)p(H) \\
&\propto p(x_{t+1}|H)p(H|\mathbf{x}_t)
\end{aligned}
$$

$$
\text{New posterior} = \text{Likelihood of new measurement} \times \text{Current posterior} \quad (6)
$$

- How does our uncertainty about H evolve given these observations?
  - Answer: Reuse the current posterior distribution as the prior distribution in the next time step, and normalize appropriately.

Sequential Bayesian Updates

- Let $\pi_t(H)$ be the posterior at time t, then the recursive update reads:

$$\pi_{t+1}(H) \propto p(x_{t+1}|H)\pi_t(H)$$

- $\pi_t(H)$ summarizes entire history of the sequence.
    - Normalization factor Z is average over all possible values of H: $Z = \sum_{h'} p(x_{t+1}|H = h')\pi_t(H = h')$
- In summary, Bayesian inference provides an efficient way of sequentially updating our belief about a state that only depends on the current measurement and posterior.

Sequential Bayesian Updates in Robotics

- In robotics, your hypothesis can be e.g., your position or state θ, which evolves in time.
- Assume your dynamics are Markov. i.e. the state θ_{t+1} only depends on the current state θ_t:

$$p(\theta_0) = \pi(\theta_0), \quad p(\theta_{t+1}|\theta_0, \theta_1, \dots, \theta_t) = p(\theta_{t+1}|\theta_t)$$

- State θ is hidden - only measurements x_t observed. To understand θ, look at joint density of all states $\boldsymbol{\theta}_t$ and measurements $\mathbf{x}_t$:

$$p(\boldsymbol{\theta}_t, \mathbf{x}_t) = \pi(\theta_0) \prod_{i=0}^{t} p(\theta_i|\theta_{i-1}) p(x_i|\theta_i)$$

The Bayes Filter

- This Markovian + Bayesian model (HMM) is widely used in:
    - Speech recognition.
    - Robotics.
    - Particle physics.
    - GPS / target tracking.
    - Brain imaging.
- In robotics, use sensor data gathered to recursively update 'belief' of position/velocity estimate.
- Remember that the evolving state θ_t is unknown, typically we want to:
    - Filter: Compute $p(\theta_t|\mathbf{x}_t)$ to estimate current state.
    - Predict: Compute $p(\theta_{t+k}|\mathbf{x}_t)$ to predict future states.
    - Reconstruct: Compute $p(\theta_{t-k}|\mathbf{x}_t)$ to identify pre vious states.

The Bayes Filter

- Want posterior at $t+1$ given observations $\mathbf{x}_{t+1}$:

$$\begin{aligned} p(\theta_{t+1}|\mathbf{x}_{t+1}) &= p(\theta_{t+1}|\mathbf{x}_t, x_{t+1}) && (7) \\ &\propto p(x_{t+1}|\theta_{t+1}, \mathbf{x}_t)p(\theta_{t+1}|\mathbf{x}_t) && (8) \\ &= p(x_{t+1}|\theta_{t+1})p(\theta_{t+1}|\mathbf{x}_t) && (9) \end{aligned}$$

- Compute $p(\theta_{t+1}|\mathbf{x}_t)$ by averaging over current state θ_t:

$$p(\theta_{t+1}|\mathbf{x}_t) = \sum_{\theta_t} p(\theta_{t+1}|\theta_t)p(\theta_t|\mathbf{x}_t)$$

The Bayes Filter

- We perform the prediction step by averaging over all possible values of the current state θ_t:

$$p(\theta_{t+1}|\mathbf{x}_t) = \int_{\theta_t} d\theta_t \; p(\theta_{t+1}|\theta_t)p(\theta_t|\mathbf{x}_t)$$

- Then perform the filter step by the New $\propto$ Current $\times$ Likelihood rule, combining the predictive distribution with the likelihood of the next measurement.

$$p(\theta_{t+1}|\mathbf{x}_{t+1}) \propto p(\theta_{t+1}|\mathbf{x}_t)p(\mathbf{x}_{t+1}|\theta_{t+1})$$

- So the overall process is:

Predict-Observe-Filter-Predict-Observe-Filter-. . .