Exponential Family

Contents

1 The Exponential Family **2**

1.1 Example: Dirichlet distribution 2

1.2 i.i.d samples from an exponential family 2

2 Exponential Family: Maximum Likelihood Estimation **3**

3 Exponential Family: Expectation Maximization **3**

4 Exponential Family: Conjugate Priors **4**

4.1 Example: Inverse Wishart prior on covariance matrix of multivariate Gaussian with known mean 4

4.2 General formalism 5

5 Exponential Family: Conditional Conjugacy **6**

5.1 Example: Bayesian normal model with conditionally conjugate prior 6

A EF representation of Multivariate Gaussian in message passing **8**

1 The Exponential Family

We are interested in the exponential family primarily because of it makes inference easier. When a problem can be cast within the exponential family framework, inference can be tied to general principles, and parameter updates often have nice interpretations. This is true regardless of whether we're doing maximum likelihood, expectation maximization, variational inference, or Gibbs sampling.

1.1 Definition

We define an *exponential family* of probability distributions as those distributions whose density has the following form

$$p(x \mid \theta) = h(x) \exp\{\eta(\theta)^T t(x) - a(\theta)\} \tag{1.1.1}$$

where we refer to h as the base measure, η as the natural parameter, t as the sufficient statistics, and a as the log normalizer. [1]

Remark 1.1.1. (*Non-uniqueness of natural parameter*) Note from (1.0.1) that natural parameters are not unique since, for example, η could be multiplied by a non-zero constant c if $t(x)$ is divided by c. [2] Thus, we should speak of *a* natural parameter, rather than *the* natural parameter.

1.2 Example: Dirichlet distribution

Example 1.2.1. (Dirichlet Distribution) We can write the density of the Dirichlet distribution in exponential form:

$$\begin{aligned} p(\pi \mid \alpha) &= \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \pi_1^{\alpha_1 - 1} \cdots \pi_K^{\alpha_K - 1} \\ &= \exp\left\{ \sum_{k=1}^{K} (\alpha_k - 1) \log \pi_k - \left[\sum_k \log \Gamma(\alpha_k) - \log \Gamma(\sum \alpha_k) \right] \right\} \end{aligned}$$

with natural parameter $\eta(\alpha) = [\alpha_1 - 1, ..., \alpha_K - 1]^T$, sufficient statistics $t(\pi) = \log \pi = [\log \pi_1, ..., \log \pi_K]^T$, base measure $h(\pi) = 1$, and log normalizer $a(\alpha) = \sum_k \log \Gamma(\alpha_k) - \log \Gamma(\sum_k \alpha_k)$. □

For an example of how the natural parametrization can help provide insight into message passing, see Section A.[3]

TODO: Add multivariate Gaussian example, showing that the natural parameters are the precision Σ^{-1} and precision-weighted mean $\Sigma^{-1}\mu$, as we use this in Section 5.1 combined with the exponential family formalism to derive the updates to the mean for a Bayesian normal model with conditionally conjugate prior.

[1] TODO: It would be helpful to get more solid on integrating against probability measure here, so that I can set this up in a more precise way, as Jordan does. He remarks on this somewhere in his exponential family lecture notes. What also may be helpful is this beautiful excerpt from pp.38 of [1]: " [...] we represent the probability distribution as a density p absolutely continuous with respect to some measure η. This base measure η might be the counting measure on $\{0, 1, ..., r-1\}$, in which case p is a probability mass function; alternatively, for a continuous random vector, the base measure η could be the ordinary Lebesgue measure on $\mathbb{R}$."

[2] Are they unique up to scalar multiplication?

[3] **Remark** The exponential family representation of the Dirichlet, as given in Example 1.1.1, is useful when we want to compute the expectation of a log probability from a Dirichlet distributed probability vector (as happens in the derivation of LDA with variational inference; see my notes on variational inference).

In those notes, we see

$$\begin{aligned} \mathbb{E}[\log \pi_k] = \mathbb{E}[t_k(p)] &\overset{1}{=} \frac{\partial}{\partial \eta_k} a(\eta) \\ &= \Psi(\alpha_k) - \Psi(\sum_k \alpha_k) \end{aligned} \tag{1.2.1}$$

where (1) uses a well-known exponential family property and where $\Psi(\cdot)$ is the first derivative of the $\log \Gamma$ function. It is known as the *digamma function*. □

1.3 i.i.d samples from an exponential family

If $\boldsymbol{x} = (x_1, ..., x_n)$ are n independent samples from the same exponential family, then

$$p(\boldsymbol{x} \mid \theta) = \prod_{i=1}^{n} h(x_i) \exp\left\{\eta(\theta)^T \sum_{i=1}^{n} t(x_i) - n\, a(\eta(\theta))\right\} \tag{1.3.1}$$

2 Exponential Family: Maximum Likelihood Estimation

The goal for maximum likelihood is to determine the parameter

$$\theta_{ML} = \operatorname*{argmax}_{\theta} \log p(\boldsymbol{x} \mid \theta) \tag{2.0.1}$$

Let us assume that $\boldsymbol{x} = (x_1, ..., x_n)$ are i.i.d observations from a fixed exponential family, so that the likelihood has form (1.2.1). Let us compute the gradient with respect to the natural parameter η of $\ell(\eta) := \log p(\boldsymbol{x} \mid \eta)$

$$\nabla_\eta \ell(\eta) = \sum_{i=1}^{n} t(x_i) - n\, \nabla_\eta a(\eta)$$

Setting the gradient to zero, we obtain

$$\nabla_\eta a(\eta) = \frac{1}{n} \sum_{i=1}^{n} t(x_i)$$

But $\nabla_\eta a(\eta) = \mathbb{E}[t(X)]$ [2]. Thus, we should set θ_{ML} such that

$$\mu(\theta_{ML}) = \frac{1}{n} \sum_{i=1}^{n} t(x_i)$$

where $\mu := \mathbb{E}[t(x)]$ refers to the mean parametrization of the likelihood. [4]

3 Exponential Family: Expectation Maximization

Some models have latent variables associated with each observation, and so maximum likelihood is not possible. Let us see how expectation maximization looks when the complete data likelihood is in the exponential family.

The expectation maximization algorithm is

$$\boldsymbol{\theta}^{(t+1)} = \operatorname{argmax}_{\boldsymbol{\theta}} \mathbb{E}_{p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}^{(t)})}\left[\ln p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})\right] \tag{3.0.1}$$

We see how this plays out in the exponential family by following the logic of Section 2. Let us assume that $(\boldsymbol{x}, \boldsymbol{z}) = ((x_1, z_1), ..., (x_n, z_n))$ are n independent samples from the same exponential family, where $\boldsymbol{x}$ is observed data and $\boldsymbol{z}$ is unobserved data. Moreover, let us assume that the complete data likelihood is in the exponential family

$$p(\boldsymbol{x}, \boldsymbol{z} \mid \theta) = \prod_{i=1}^{n} h(x_i, z_i) \exp\left\{\eta(\theta)^T \sum_{i=1}^{n} t(x_i, z_i) - n\, a(\eta(\theta))\right\} \tag{3.0.2}$$

[4] TODO: This switching of parameterization should be handled much more explicitly.

Here we want to find $\boldsymbol{\theta}$ to optimize

$$f(\boldsymbol{\theta}) = \mathbb{E}_{p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}^{(t)})}\Big[\ln p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})\Big]$$

Following the logic of Section 2, we determine that we should select $\theta^{(t+1)}$ such that

$$\mu(\theta^{(t+1)}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}^{(t)})} t(x_i, z_i)$$

where $\mu := \mathbb{E}[t(x_1, z_1)]$ refers to the mean parametrization of the likelihood.

This is why an EM iteration is often described and/or implemented as performing maximum likelihood with the expected sufficient statistics.

TODO: But is EM *always* equivalent to performing ML with ESS's? Or is this ONLY true if I'm working within the exponential family? I need to read up some more on EM theory.

TODO: Check this section, especially with respect to the fact that I am dealing with three parametrizations here - μ, θ, ν; that is, mean, arbitrary, and natural, respectively. Really the core problem is that it's not sufficiently clear in how head how and when reparametrizations affect things.

4 Exponential Family: Conjugate Priors

TODO: State what a conjugate prior is without using the formalism of Section 4.2

4.1 Example: Inverse Wishart prior on covariance matrix of multivariate Gaussian with known mean

Here we will show that the Inverse Wishart is a conjugate prior for the covariance of a multivariate normally distributed random variable with known mean.

This situation comes up

Example 4.1.1. (*Inverse Wishart prior on the covariance of a Multivariate Normal sampling model with known mean*)

Consider the sampling model for $\boldsymbol{y} := (\boldsymbol{y}_1, ..., \boldsymbol{y}_n) \overset{\text{iid}}{\sim} \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$\begin{aligned} p(y \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) &\propto |\boldsymbol{\Sigma}|^{-n/2} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu})\right] \\ &= |\boldsymbol{\Sigma}|^{-n/2} \exp\left[-\frac{1}{2}\text{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}_\mu)\right] \end{aligned} \tag{4.1.1}$$

where $\boldsymbol{S}_\mu := \sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{\mu})(\boldsymbol{y}_i - \boldsymbol{\mu})^T$ is the sum of pairwise deviation products, and where the equality in (4.1.1) is justified in Remark 4.1.2.

Let us take the mean $\boldsymbol{\mu}$ to be known, and let us take the prior on the covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ to be given by $\boldsymbol{\Sigma} \sim \mathcal{W}^{-1}(\boldsymbol{\Psi}, \nu)$, i.e.

$$p(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(\nu+d+1)/2} \exp\left[-\frac{1}{2}\text{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Psi})\right] \tag{4.1.2}$$

where $\boldsymbol{\Sigma} \succ 0$ and $\nu > d - 1$ to have a proper prior. Note that $\mathbb{E}[\boldsymbol{\Sigma}] = \frac{\boldsymbol{\Psi}}{\nu - d - 1}$.

It is easy to see from the forms of the likelihood (4.1.1) and prior (4.1.2) that the Inverse Wishart is a conjugate prior in this context. In particular

$$p(\boldsymbol{\Sigma} \mid \boldsymbol{\mu}, \boldsymbol{y}) \propto |\boldsymbol{\Sigma}|^{-(\nu+n+d+1)/2} \exp\left[-\frac{1}{2}\text{tr}\big(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Psi} + \boldsymbol{S}_\mu)\big)\right] \tag{4.1.3}$$

where S_μ was defined above. Thus, we have

$$\boldsymbol{\Sigma} \mid \boldsymbol{\mu}, \boldsymbol{y} \sim \mathcal{W}^{-1}\left(\boldsymbol{\Psi} + \sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{\mu})(\boldsymbol{y}_i - \boldsymbol{\mu})^T, \nu + n\right)$$

And so the conjugate updates are given by

$$\nu' \leftarrow \nu + n \tag{4.1.4}$$

$$\boldsymbol{\Psi}' \leftarrow \boldsymbol{\Psi} + \sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{\mu})(\boldsymbol{y}_i - \boldsymbol{\mu})^T \tag{4.1.5}$$

Remark 4.1.1. (*Interpreting the hyperparameters of the Inverse Wishart*) Note that the hyperparameters of the Inverse Wishart can be interpreted (as per conjugacy) in the following way: the covariance was estimated from ν observations with a sum of pairwise deviation products $\boldsymbol{\Psi}$. [5]

Remark 4.1.2. (*Expressing the Multivariate Gaussian density in a nice form for the Inverse Wishart prior on the Covariance Matrix*)

Here we justify the equality of (4.1.1).

We will show that $\sum_{i=1}^{n} \boldsymbol{x}_i^T \boldsymbol{A} \boldsymbol{x}_i = \operatorname{tr}(\boldsymbol{A} \sum \boldsymbol{x}_i \boldsymbol{x}_i^T)$ for $\boldsymbol{x} \in \mathbb{R}^d$, and $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ symmetric.

$$\begin{aligned}
\sum_{i=1}^{n} \boldsymbol{x}_i^T \boldsymbol{A} \boldsymbol{x}_i &= \sum_{i=1}^{n} \sum_{j,k=1}^{n} a_{jk} x_{ij} x_{ik} \\
&= \sum_{j,k=1}^{n} \left(\boldsymbol{A} \circ \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T \right)_{jk} \\
&\overset{(*)}{=} \operatorname{tr}(\boldsymbol{A} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T)
\end{aligned}$$

where $\circ$ is the Hadamard, also called the elementwise, operator, and where (*) holds by properties of the tr operator

$$\operatorname{tr}(\boldsymbol{A}\boldsymbol{B}) = \sum_{i,j} (\boldsymbol{A}^T \circ \boldsymbol{B})_{ij} \overset{\boldsymbol{A} \text{ symmetric}}{=} \sum_{i,j} (\boldsymbol{A} \circ \boldsymbol{B})_{ij}$$

4.2 General formalism

Here we provide some notes, following [2], about conjugate priors for exponential family data models.

Writing the exponential family density in canonical form, we have

$$p(x \mid \eta) = h(x) \exp\{\eta^T T(x) - A(\eta)\}$$

where η is the canonical parameter, $T(x)$ are the sufficient statistics, $h(x)$ is the base measure, and $A(\eta)$ is the log normalizer (and so is *not* a degree of freedom).

The natural parameter space is

$$\left\{ \eta : \int h(x) \exp\{\eta^T T(x) - A(\eta)\} < \infty \right\}$$

Given a random sample, $\boldsymbol{x} = (x_1, x_2, .., x_N)$, we obtain:

$$p(\boldsymbol{x} \mid \eta) = \left(\prod_{i=1}^{n} h(x_i) \right) \exp\left\{ \eta^T \sum_{i=1}^{n} T(x_i) - N A(\eta) \right\}$$

[5] This interpretation also makes the formula for $\mathbb{E}[\boldsymbol{\Sigma}]$ more intuitive.

as the likelihood function.

A conjugate prior can be obtained by mimicking the likelihood

$$p(\eta \mid \tau, \eta_0) = H(\tau, \eta_0) \exp\{\tau^T \eta - \eta_o A(\eta)\} \tag{4.2.1}$$

where now $H(\tau, \eta_0)$ is the normalizing factor. (For conditions on normalizability, see [**?**]). Note that τ has the dimensionality of the canonical parameter η and n_0 is a scalar.

To verify conjugacy, we compute the posterior density

$$p(\eta \mid \boldsymbol{x}, \tau, \eta_0) \propto \exp\left\{\left(\tau + \sum_{n=1}^{N} T(x_n)\right)^T \eta - (n_0 + N) A(\eta)\right\}$$

which retains the form of (4.2.1)

Thus, the prior-to-posterior conversion can be summarized with the following update rules

$$\begin{aligned} \tau &\to \tau + \sum_{n=1}^{N} T(x_n) \\ n_0 &\to n_0 + N \end{aligned} \tag{4.2.2}$$

For conjugate Bayesian models, the predictive posterior distribution, $p(x_{\text{new}} \mid x)$ is always tractable, because it has the same form (integrating a likelihood against the parameter distribution) as does the evidence term in Bayes law. For exponential family models, the predictive posterior takes the form of a ratio of normalizing factors

$$p(x_{\text{new}} \mid x) = \frac{H(\tau_{\text{post}}), n_0 + N}{H(\tau_{\text{post}} + T(x_{\text{new}}), n_0 + N + 1)} \tag{4.2.3}$$

TODO: Redo some of the examples using the exponential family conjugate prior formalism. A possibly useful resource in the giant table at `https://en.wikipedia.org/wiki/Exponential_family`.

5 Exponential Family: Conditional Conjugacy

A family of prior distributions for a parameter is called conditionally conjugate if the conditional posterior distribution (often called the *complete conditional*), given the data and all other parameters in the model, is also in that class [3]. The posterior distribution for conditionally conjugate models is easily approximated with Gibbs sampling or Mean Field Variational Inference – the former samples from the complete conditional, whereas the latter takes variational expectations with respect to the natural parameter of the complete conditional.

Below we give perhaps the simplest example.

5.1 Example: Bayesian normal model with conditionally conjugate prior

Consider the following model with a normal sampling distribution and conditionally conjugate prior[6]:

$$\begin{aligned} \boldsymbol{\mu} &\sim \mathcal{N}_d(\boldsymbol{m}_0, \boldsymbol{V}_0) \\ \boldsymbol{\Sigma} &\sim \mathcal{W}^{-1}(\nu_0, \boldsymbol{\Psi}_0) \\ \boldsymbol{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma} &\overset{\text{iid}}{\sim} \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad i = 1, ..., N \end{aligned}$$

We define $\boldsymbol{x} := (\boldsymbol{x}_1, \ldots \boldsymbol{x}_N)$, where each $\boldsymbol{x}_i \in \mathbb{R}^d$.

[6] TODO: Prove that the prior, although conditionally conjugate, is not conjugate. (I *believe* this is true, based on context clues from experience, but I am not currently certain about it.)

The complete conditionals are well-known. In particular

$$\boldsymbol{\mu} \mid \boldsymbol{\Sigma}, \boldsymbol{x} \sim \mathcal{N}_d(\boldsymbol{m}, \boldsymbol{V}) \tag{5.1.1}$$

where

$$\boldsymbol{m} = \left(\boldsymbol{V}_0^{-1} + N\boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\boldsymbol{V}_0^{-1}\boldsymbol{m}_0 + N\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{x}}\right)$$
$$\boldsymbol{V} = \left(\boldsymbol{V}_0^{-1} + N\boldsymbol{\Sigma}^{-1}\right)^{-1}$$

and

$$\boldsymbol{\Sigma} \mid \boldsymbol{\mu}, \boldsymbol{x} \sim \mathcal{W}^{-1}(\nu, \boldsymbol{\Psi}) \tag{5.1.2}$$

where

$$\nu = \nu_0 + N$$
$$\boldsymbol{\Psi} = \Psi_0 + \sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T$$

Indeed, we derived (5.1.2) in Section 4.1. [7]

Note that the model is different than the model fully conjugate (Normal-Inverse-Wishart) prior on the pair $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The conditionally conjugate prior lacks closed-form posterior updating, but is also more expressive. [8]

These conjugate posterior updates have nice interpretations:

- **Hyperparameter updates for** $(\boldsymbol{\mu} \mid \boldsymbol{\Sigma}, \boldsymbol{x})$: On the precision scale, $\boldsymbol{V}$ is the sum of the prior precision matrix $\boldsymbol{V}_0^{-1}$ and N copies of the precision for each observation, $\boldsymbol{\Sigma}^{-1}$. Similarly, $\boldsymbol{m}$ is the precision-weighted convex combination of $\boldsymbol{m}_0$, the prior mean and the empirical average $\bar{\boldsymbol{x}}$.
- **Hyperparameter updates for** $(\boldsymbol{\Sigma} \mid \boldsymbol{\mu}, \boldsymbol{x})$: The covariance was estimated from ν observations with a sum of pairwise deviation products Ψ.

[7] We still need to add a derivation for (5.1.1) TODO , but the birds' eye view for one approach is to use the general formalism for conjugacy updates in the exponential family (4.2.2), noting that the natural parameters for a multivariate Gaussian are its precision and precision-weighted mean.

[8] Is it also more expressive once we move to a variational approximation? i.e., can we get more expressive marginals this way?

A EF representation of Multivariate Gaussian in message passing

In a dissertation on Gaussian Belief Propagation [4], referred to in [5], a multivariate Gaussian is considered as a Markov Random Field.

In particular, consider the Markov Random field

$$p(x) = \frac{1}{Z}\left(\prod_{i=1}^{n}\phi(x_i)\prod_{i,j}\psi(x_i,x_j)\right) \tag{A.0.1}$$

Now note that a multivariate Gaussian has a joint distribution which can be expressed as

$$p(x) \propto \exp\left\{-\frac{1}{2}x^T Ax + b^T x\right\}$$

as this is just the exponential family form of a Gaussian (e.g., see [6]), where the natural parameters are given in terms of the *precision* Σ^{-1}

$$\begin{aligned} A &= \Sigma^{-1} \\ b &= \Sigma^{-1}\mu \end{aligned}$$

Thus, the multivariate Gaussian is a MRF where the potentials in (A.0.1) are given by

$$\begin{aligned} \psi_{ij}(x_i,x_j) &:= \exp\left\{-\frac{1}{2}x_i A_{ij} x_j\right\} \\ \phi_i(x_i) &:= \exp\left\{-\frac{1}{2}A_{ii}x_i^2 + b_i x_i\right\} \end{aligned}$$

This seems to be useful in inference for state space models, where one multiplies multiple "messages" that are different Gaussian densities *over the same variable*. For example, see the equations for μ_t and σ_t^2 in Section 4 of [5], where messages from the past and the future of a time series model are combined to get a posterior distribution on the state z_t. The combined parameters have an expression which may at first be puzzling:

$$\mu_t = \frac{\mu_1\sigma_2^2 + \mu_2\sigma_1^2}{\sigma_1^2+\sigma_2^2}, \quad \sigma_t^2 = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}$$

However, these messages have an intuitive form when considered in terms of the natural parameterizations: the combined mean is a weighted combination of the original means, with the weights given by the precisions. The combined precision (inverse covariance) is given simply by the sum of the original precisions. Very nice!

See Figure 1 for the general expression, which explains the formula in [5]. This is an example of where the natural parametrization provides more insight than the standard parametrization.

Lemma 12. *Let $f_1(x)$ and $f_2(x)$ be the probability density functions of a Gaussian random variable with two possible densities $\mathcal{N}(\mu_1, P_1^{-1})$ and $\mathcal{N}(\mu_2, P_2^{-1})$, respectively. Then their product, $f(x) = f_1(x)f_2(x)$ is, up to a constant factor, the probability density function of a Gaussian random variable with distribution $\mathcal{N}(\mu, P^{-1})$, where*

$$\mu = P^{-1}(P_1\mu_1 + P_2\mu_2), \tag{2.9}$$
$$P^{-1} = (P_1+P_2)^{-1}. \tag{2.10}$$

Figure 1: Lemma 12 of [4]

References

[1] Martin J Wainwright and Michael Irwin Jordan. *Graphical models, exponential families, and variational inference*. Now Publishers Inc, 2008.

[2] Michael Jordan. *The exponential family: Conjugate priors*, (accessed September 11, 2020).

[3] Andrew Gelman et al. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, 1(3):515–534, 2006.

[4] Danny Bickson. Gaussian belief propagation: Theory and aplication. *arXiv preprint arXiv:0811.2518*, 2008.

[5] Rahul G Krishnan, Uri Shalit, and David Sontag. Structured inference networks for nonlinear state space models. *arXiv preprint arXiv:1609.09869*, 2016.

[6] Barbara Englehardt. *Gaussian Models*, (accessed November 22, 2020).