Chapter 6

Exponential families and maximum entropy

In this set of notes, we give a very brief introduction to exponential family models, which are a broad class of distributions that have been extensively studied in the statistics literature [4, 1, 2, 7]. There are deep connections between exponential families, convex analysis [7], and information geometry and the geometry of probability measures [1], and we will only touch briefly on a few of those here.

6.1 Review or introduction to exponential family models

We begin by defining exponential family distributions, giving several examples to illustrate a few of their properties. To define an exponential family distribution, we always assume there is some base measure μ on a space $\mathcal{X}$, and there exists a *sufficient statistic* $\phi : \mathcal{X} \to \mathbb{R}^d$, where $d \in \mathbb{N}$ is some fixed integer. For a given sufficient statistic function ϕ, let $\theta \in \mathbb{R}^d$ be an associated vector of *canonical* parameters. Then with this notation, we have the following.

Definition 6.1. *The* exponential family *associated with the function ϕ and base measure μ is defined as the set of distributions with densities p_θ with respect to μ, where*

$$p_\theta(x) = \exp\left(\langle \theta, \phi(x) \rangle - A(\theta)\right), \tag{6.1.1}$$

and the function A is the log-partition-function *(or cumulant function) defined by*

$$A(\theta) := \log \int_{\mathcal{X}} \exp\left(\langle \theta, \phi(x) \rangle\right) d\mu(x), \tag{6.1.2}$$

whenever A is finite.

In some settings, it is convenient to define a base function $h : \mathcal{X} \to \mathbb{R}_+$ and define

$$p_\theta(x) = h(x) \exp(\langle \theta, \phi(x) \rangle - A(\theta)),$$

though we can always simply include h in the base measure μ. In some scenarios, it may be convient to re-parameterize the problem in terms of some function $\eta(\theta)$ instead of θ itself; we will not worry about such issues and simply use the formulae that are most convenient.

We now give a few examples of exponential family models.

Example 6.1 (Bernoulli distribution)**:** In this case, we have $X \in \{0, 1\}$ and $P(X = 1) = p$ for some $p \in [0, 1]$ in the classical version of a Bernoulli. Thus we take μ to be the counting measure on $\{0, 1\}$, and by setting $\theta = \log \frac{p}{1-p}$ to obtain a canonical representation, we have

$$P(X = x) = p(x) = p^x(1-p)^{1-x} = \exp(x \log p - x \log(1-p)) \\ = \exp\left(x \log \frac{p}{1-p} + \log(1-p)\right) = \exp\left(x\theta - \log(1+e^\theta)\right).$$

The Bernoulli family thus has log-partition function $A(\theta) = \log(1 + e^\theta)$. ♣

Example 6.2 (Poisson distribution)**:** The Poisson distribution (for count data) is usually parameterized by some $\lambda > 0$, and for $x \in \mathbb{N}$ has distribution $P_\lambda(X = x) = (1/x!)\lambda^x e^-\lambda$. Thus by taking μ to be counting (discrete) measure on $\{0, 1, \ldots\}$ and setting $\theta = \log \lambda$, we find the density (probability mass function in this case)

$$p(x) = \frac{1}{x!}\lambda^x e^{-\lambda} = \exp(x \log \lambda - \lambda)\frac{1}{x!} = \exp(x\theta - e^\theta)\frac{1}{x!}.$$

Notably, taking $h(x) = (x!)^{-1}$ and log-partition $A(\theta) = e^\theta$, we have probability mass function $p_\theta(x) = h(x) \exp(\theta x - A(\theta))$. ♣

Example 6.3 (Normal distribution)**:** For the normal distribution, we take μ to be Lebesgue measure on $(-\infty, \infty)$. Then $\mathsf{N}(\mu, \Sigma)$ can be re-parameterized as as $\Theta = \Sigma^{-1}$ and $\theta = \Sigma^{-1}\mu$, and we have density

$$p_{\theta,\Theta}(x) \propto \exp\left(\langle \theta, x\rangle + \frac{1}{2}\left\langle xx^\top, \Theta\right\rangle\right),$$

where $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product. ♣

6.1.1 Why exponential families?

There are many reasons for us to study exponential families. As we see presently, they arise as the solutions to several natural optimization problems on the space of probability distributions. They also enjoy certain robustness properties related to optimal Bayes' procedures (more to come on this topic). Moreover, they are analytically very tractable, and have been the objects of substantial study for nearly the past hundred years. As one example, the following result is well-known (see, e.g., Wainwright and Jordan [7, Proposition 3.1] or Brown [4]):

Proposition 6.4. *The log-partition function $\theta \mapsto A(\theta)$ is infinitely differentiable on its open domain $\Theta := \{\theta \in \mathbb{R}^d : A(\theta) < \infty\}$. Moreover, A is convex.*

Proof We show convexity; the proof of the infinite differentiability follows from an argument using the dominated convergence theorem that allows passing the derivative through the integral defining A. For convexity, let let $\theta_\lambda = \lambda\theta_1 + (1-\lambda)\theta_2$, where $\theta_1, \theta_2 \in \Theta$. Then $1/\lambda \geq 1$ and $1/(1-\lambda) \geq 1$, and Hölder's inequality implies

$$\log \int \exp(\langle \theta_\lambda, \phi(x)\rangle) d\mu(x) = \log \int \exp(\langle \theta_1, \phi(x)\rangle)^\lambda \exp(\langle \theta_2, \phi(x)\rangle)^{1-\lambda} d\mu(x) \\ \leq \log \left(\int \exp(\langle \theta_1, \phi(x)\rangle)^{\frac{\lambda}{\lambda}} d\mu(x)\right)^\lambda \left(\int \exp(\langle \theta_2, \phi(x)\rangle)^{\frac{1-\lambda}{1-\lambda}} d\mu(x)\right)^{1-\lambda} \\ = \lambda \log \int \exp(\langle \theta_1, \phi(x)\rangle) d\mu(x) + (1-\lambda) \log \int \exp(\langle \theta_2, \phi(x)\rangle) d\mu(x),$$

as desired. □

As a final remark, we note that this convexity makes estimation in exponential families substantially easier. Indeed, given a sample $X_1, \ldots, X_n$, assume that we estimate θ by maximizing the likelihood (equivalently, minimizing the log-loss):

$$\minimize_\theta \sum_{i=1}^n \log \frac{1}{p_\theta(X_i)} = \sum_{i=1}^n \left[- \langle \theta, \phi(X_i) \rangle + A(\theta) \right],$$

which is thus convex in θ. This means there are no local minima, and tractable algorithms exist for solving maximum likelihood. Later we will explore some properties of these types of minimization and log-loss problems.

6.2 Shannon entropy

We now explore a generalized version of entropy known as Shannon entropy, which allows us to define an entropy functional for essentially arbitrary distributions. This comes with a caveat, however: to define this entropy, we must fix a base measure μ ahead of time against which we integrate. In this case, we have

Definition 6.2. *Let μ be a base measure on $\mathcal{X}$ and assume P has density p with respect to μ. Then the* Shannon entropy *of P is*

$$H(P) = - \int p(x) \log p(x) d\mu(x).$$

Notably, if $\mathcal{X}$ is a discrete set and μ is counting measure, then $H(P) = -\sum_x p(x) \log p(x)$ is simply the standard entropy. However, for other base measures the calculation is different. For example, if we take μ to be Lebesgue measure, meaning that $d\mu(x) = dx$ and giving rise to the usual integral on $\mathbb{R}$ (or $\mathbb{R}^d$), then we obtain *differential entropy* [5, Chapter 8].

Example 6.5: Let P be the uniform distribution on $[0, a]$. Then the differential entropy $H(P) = -\log(1/a) = \log a$. ♣

Example 6.6: Let P be the normal distribution $\mathsf{N}(\mu, \Sigma)$ and μ be Lebesgue measure. Then

$$\begin{aligned} H(P) &= - \int p(x) \left[\log \frac{1}{\sqrt{2\pi \det(\Sigma)}} - \frac{1}{2} (x - \mu)^\top \Sigma^{-1} (x - \mu) \right] dx \\ &= \frac{1}{2} \log(2\pi \det(\Sigma)) + \frac{1}{2} \mathbb{E}[(X - \mu)^\top \Sigma^{-1} (X - \mu)] \\ &= \frac{1}{2} \log(2\pi \det(\Sigma)) + \frac{d}{2}. \end{aligned}$$

♣

6.3 Maximizing Entropy

The maximum entropy principle, proposed by Jaynes in the 1950s (see Jaynes [6]), originated in statistical mechanics, where Jaynes showed that (in a sense) entropy in statistical mechanics and information theory were equivalent. The maximum entropy principle is this: given some constraints (prior information) about a distribution P, we consider all probability distributions satisfying said constraints. Then to encode our prior information while being as "objective" or "agnostic" as possible (essentially being as uncertain as possible), we should choose the distribution P satisfying the constraints to maximize the Shannon entropy.

While there are many arguments for and against the maximum entropy principle, we shall not dwell on them here, instead showing how maximizing entropy naturally gives rise to exponential family models. We will later see connections to Bayesian and minimax procedures. The one thing that we must consider, about which we will be quite explicit, is that the base measure μ is *essential* to all our derivations: it radically effects the distributions P we consider.

6.3.1 The maximum entropy problem

We begin by considering linear (mean-value) constraints on our distributions. In this case, we are given a function $\phi : \mathcal{X} \to \mathbb{R}^d$ and vector $\alpha \in \mathbb{R}^d$, we wish to solve

$$\text{maximize } H(P) \quad \text{subject to} \quad \mathbb{E}_P[\phi(X)] = \alpha \tag{6.3.1}$$

over all distributions P having densities with respect to the base measure μ, that is, we have the (equivalent) absolute continuity condition $P \ll \mu$. Rewriting problem (6.3.1), we see that it is equivalent to

$$\begin{aligned} \text{maximize} \quad & -\int p(x) \log p(x) d\mu(x) \\ \text{subject to} \quad & \int p(x)\phi_i(x)d\mu(x) = \alpha_i, \quad p(x) \geq 0 \text{ for } x \in \mathcal{X}, \quad \int p(x)d\mu(x) = 1. \end{aligned}$$

Let

$$\mathcal{P}_\alpha^{\text{lin}} := \{P \ll \mu : \mathbb{E}_P[\phi(X)] = \alpha\}$$

be distributions with densities w.r.t. μ satisfying the expectation (linear) constraint $\mathbb{E}[\phi(X)] = \alpha$. We then obtain the following theorem.

Theorem 6.7. *For $\theta \in \mathbb{R}^d$, let P_θ have density*

$$p_\theta(x) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)), \quad A(\theta) = \log \int \exp(\langle \theta, \phi(x) \rangle) d\mu(x),$$

with respect to the measure μ. If $\mathbb{E}_{P_\theta}[\phi(X)] = \alpha$, then P_θ maximizes $H(P)$ over $\mathcal{P}_\alpha^{\text{lin}}$; moreover, the distribution P_θ is unique.

Proof We first give a heuristic derivation—which is not completely rigorous—and then check to verify that our result is exact. First, we write a Lagrangian for the problem (6.3.1). Introducing Lagrange multipliers $\lambda(x) \geq 0$ for the constraint $p(x) \geq 0$, $\theta_0 \in \mathbb{R}$ for the normalization constraint

that $P(\mathcal{X}) = 1$, and θ_i for the constraints that $\mathbb{E}_P[\phi_i(X)] = \alpha_i$, we obtain the following Lagrangian:

$$\mathcal{L}(p, \theta, \theta_0, \lambda) = \int p(x) \log p(x) d\mu(x) + \sum_{i=1}^{d} \theta_i \left(\alpha_i - \int p(x) \phi_i(x) d\mu(x) \right) + \theta_0 \left(\int p(x) d\mu(x) - 1 \right) - \int \lambda(x) p(x) d\mu(x).$$

Now, heuristically treating the density $p = [p(x)]_{x \in \mathcal{X}}$ as a finite-dimensional vector (in the case that $\mathcal{X}$ is finite, this is completely rigorous), we take derivatives and obtain

$$\frac{\partial}{\partial p(x)} \mathcal{L}(p, \theta, \theta_0, \lambda) = 1 + \log p(x) - \sum_{i=1}^{d} \theta_i \phi_i(x) + \theta_0 - \lambda(x) = 1 + \log p(x) - \langle \theta, \phi(x) \rangle + \theta_0 - \lambda(x).$$

To find the minimizing p for the Lagrangian (the function is convex in p), we set this equal to zero to find that

$$p(x) = \exp\left(\langle \theta, \phi(x) \rangle - 1 - \theta_0 - \lambda(x) \right).$$

Now, we note that with this setting, we always have $p(x) > 0$, so that the constraint $p(x) \geq 0$ is unnecessary and (by complementary slackness) we have $\lambda(x) = 0$. In particular, by taking $\theta_0 = -1 + A(\theta) = -1 + \log \int \exp(\langle \theta, \phi(x) \rangle) d\mu(x)$, we have that (according to our heuristic derivation) the optimal density p should have the form

$$p_\theta(x) = \exp\left(\langle \theta, \phi(x) \rangle - A(\theta) \right).$$

So we see the form of distribution we would like to have.

Let us now consider any other distribution $P \in \mathcal{P}_\alpha^{\text{lin}}$, and assume that we have some θ satisfying $\mathbb{E}_{P_\theta}[\phi(X)] = \alpha$. In this case, we may expand the entropy $H(P)$ as

$$\begin{aligned} H(P) &= -\int p \log p d\mu = -\int p \log \frac{p}{p_\theta} d\mu - \int p \log p_\theta d\mu \\ &= -D_{\rm kl}\left(P \| P_\theta\right) - \int p(x) [\langle \theta, \phi(x) \rangle - A(\theta)] d\mu(x) \\ &\stackrel{(\star)}{=} -D_{\rm kl}\left(P \| P_\theta\right) - \int p_\theta(x) [\langle \theta, \phi(x) \rangle - A(\theta)] d\mu(x) \\ &= -D_{\rm kl}\left(P \| P_\theta\right) - H(P_\theta), \end{aligned}$$

where in the step $(\star)$ we have used the fact that $\int p(x) \phi(x) d\mu(x) = \int p_\theta(x) \phi(x) d\mu(x) = \alpha$. As $D_{\rm kl}\left(P \| P_\theta\right) > 0$ unless $P = P_\theta$, we have shown that P_θ is the unique distribution maximizing the entropy, as desired. □

6.3.2 Examples of maximum entropy

We now give three examples of maximum entropy, showing how the choice of the base measure μ strongly effects the resulting maximum entropy distribution. For all three, we assume that the space $\mathcal{X} = \mathbb{R}$ is the real line. We consider maximizing the entropy over all distributions P satisfying

$$\mathbb{E}_P[X^2] = 1.$$

Example 6.8: Assume that the base measure μ is counting measure on the support $\{-1, 1\}$, so that $\mu(\{-1\}) = \mu(\{1\}) = 1$. Then the maximum entropy distribution is given by $P(X = x) = \frac{1}{2}$ for $x \in \{-1, 1\}$. ♣

Example 6.9: Assume that the base measure μ is Lebesgue measure on $\mathcal{X} = \mathbb{R}$, so that $\mu([a, b]) = b - a$ for $b \geq a$. Then by Theorem 6.7, we have that the maximum entropy distribution has the form $p_\theta(x) \propto \exp(-\theta x^2)$; recognizing the normal, we see that the optimal distribution is simply $\mathsf{N}(0, 1)$. ♣

Example 6.10: Assume that the base measure μ is counting measure on the integers $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, \ldots\}$. Then Theorem 6.7 shows that the optimal distribution is a discrete version of the normal: we have $p_\theta(x) \propto \exp(-\theta x^2)$ for $x \in \mathbb{Z}$. That is, we choose $\theta > 0$ so that the distribution $p_\theta(x) = \exp(-\theta x^2)/\sum_{j=-\infty}^{\infty} \exp(-\theta j^2)$ has variance 1. ♣

6.3.3 Generalization to inequality constraints

It is possible to generalize Theorem 6.7 in a variety of ways. In this section, we show how to generalize the theorem to general (finite-dimensional) convex cone constraints (cf. Boyd and Vandenberghe [3, Chapter 5]). To remind the reader, we say a set $\mathcal{C}$ is a *convex cone* if for any two points $x, y \in \mathcal{C}$, we have $\lambda x + (1 - \lambda)y \in \mathcal{C}$ for all $\lambda \in [0, 1]$, and $\mathcal{C}$ is closed under positive scaling: $x \in \mathcal{C}$ implies that $tx \in \mathcal{C}$ for all $t \geq 0$. While this level of generality may seem a bit extreme, it does give some nice results. In most cases, we will always use one of the following two standard examples of cones (the positive orthant and the semi-definite cone):

i. *The orthant.* Take $\mathcal{C} = \mathbb{R}^d_+ = \{x \in \mathbb{R}^d : x_j \geq 0, j = 1, \ldots, d\}$. Then clearly $\mathcal{C}$ is convex and closed under positive scaling.

ii. *The semidefinite cone.* Take $\mathcal{C} = \{X \in \mathbb{R}^{d \times d} : X = X^\top, X \succeq 0\}$, where a matrix $X \succeq 0$ means that $a^\top X a \geq 0$ for all vectors a. Then we have that $\mathcal{C}$ is convex and closed under positive scaling as well.

Given a convex cone $\mathcal{C}$, we associate a cone ordering $\succeq$ with the cone and say that for two elements $x, y \in \mathcal{C}$, we have $x \succeq y$ if $x - y \succeq 0$, that is, $x - y \in \mathcal{C}$. In the orthant case, this simply means that x is component-wise larger than y. For a given inner product $\langle \cdot, \cdot \rangle$, we define the dual cone

$$\mathcal{C}^* := \{y : \langle y, x \rangle \geq 0 \text{ for all } x \in \mathcal{C}\}.$$

For the standard (Euclidean) inner product, the positive orthant is thus self-dual, and similarly the semidefinite cone is also self-dual. For a vector y, we write $y \succeq_* 0$ if $y \in \mathcal{C}^*$ is in the dual cone.

With this generality in mind, we may consider the following linearly constrained maximum entropy problem, which is predicated on a particular cone $\mathcal{C}$ with associated cone ordering $\preceq$ and a function ψ mapping into the ambient space in which $\mathcal{C}$ lies:

$$\text{maximize } H(P) \quad \text{subject to} \quad \mathbb{E}_P[\phi(X)] = \alpha, \quad \mathbb{E}_P[\psi(X)] \preceq \beta, \tag{6.3.2}$$

where the base measure μ is implicit. We denote the family of distributions (with densities w.r.t. μ) satisfying the two above constraints by $\mathcal{P}^{\rm lin}_{\alpha,\beta}$. Equivalently, we wish to solve

$$\begin{aligned} \text{maximize} \quad & -\int p(x)\log p(x)d\mu(x) \\ \text{subject to} \quad & \int p(x)\phi(x)d\mu(x) = \alpha, \quad \int p(x)\psi(x)d\mu(x) \preceq \beta, \\ & p(x) \geq 0 \text{ for } x \in \mathcal{X}, \quad \int p(x)d\mu(x) = 1. \end{aligned}$$

We then obtain the following theorem:

Theorem 6.11. *For $\theta \in \mathbb{R}^d$ and $K \in \mathcal{C}^*$, the dual cone to $\mathcal{C}$, let $P_{\theta,K}$ have density*

$$p_{\theta,K}(x) = \exp\left(\langle\theta,\phi(x)\rangle - \langle K,\psi(x)\rangle - A(\theta,K)\right), \quad A(\theta,K) = \log\int \exp(\langle\theta,\phi(x)\rangle - \langle K,\psi(x)\rangle)d\mu(x),$$

with respect to the measure μ. If

$$\mathbb{E}_{P_{\theta,K}}[\phi(X)] = \alpha \quad \textit{and} \quad \mathbb{E}_{P_{\theta,K}}[\psi(X)] = \beta,$$

then $P_{\theta,K}$ maximizes $H(P)$ over $\mathcal{P}^{\rm lin}_{\alpha,\beta}$. Moreover, the distribution $P_{\theta,K}$ is unique.

We make a few remarks in passing before proving the theorem. First, we note that we must assume both equalities are attained for the theorem to hold. We may also present an example.

Example 6.12 (Normal distributions maximize entropy subject to covariance constraints)**:** Suppose that the cone $\mathcal{C}$ is the positive semidefinite cone in $\mathbb{R}^{d\times d}$, that $\alpha = 0$, that we use the Lebesgue measure as our base measure, and that $\psi(x) = xx^\top \in \mathbb{R}^{d\times d}$. Let us fix $\beta = \Sigma$ for some positive definite matrix Σ. This gives us the problem

$$\text{maximize} \quad -\int p(x)\log p(x)dx \quad \text{subject to } \mathbb{E}_P[XX^\top] \preceq \Sigma$$

Then we have by Theorem 6.11 that if we can find a density $p_K(x) \propto \exp(-\left\langle K, xx^\top\right\rangle) = \exp(-x^\top Kx)$ satisfying $\mathbb{E}[XX^\top] = \Sigma$, this distribution maximizes the entropy. But this is not hard: simply take the normal distribution $\mathsf{N}(0,\Sigma)$, which gives $K = \frac{1}{2}\Sigma^{-1}$. ♣

Now we provide the proof of Theorem 6.11.

Proof We can provide a heuristic derivation of the form of $p_{\theta,K}$ identically as in the proof of Theorem 6.7, where we also introduce the dual variable $K \in \mathcal{C}^*$ for the constraint $\int p(x)\psi(x)d\mu(x) \preceq \beta$. Rather than going through this, however, we simply show that the distribution $P_{\theta,K}$ maximizes $H(P)$. Indeed, we have for any $P \in \mathcal{P}^{\rm lin}_{\alpha,\beta}$ that

$$\begin{aligned} H(P) &= -\int p(x)\log p(x)d\mu(x) = -\int p(x)\log\frac{p(x)}{p_{\theta,K}(x)}d\mu(x) - \int p(x)\log p_{\theta,K}(x)d\mu(x) \\ &= -D_{\rm kl}\left(P\|P_{\theta,K}\right) - \int p(x)\left[\langle\theta,\phi(x)\rangle - \langle K,\psi(x)\rangle - A(\theta,K)\right]d\mu(x) \\ &\leq -D_{\rm kl}\left(P\|P_{\theta,K}\right) - \left[\langle\theta,\alpha\rangle - \langle K,\beta\rangle - A(\theta,K)\right], \end{aligned}$$

where the inequality follows because $K \succeq_* 0$ so that if $\mathbb{E}[\psi(X)] \preceq \beta$, we have

$$\langle K, \mathbb{E}[\psi(X) - \beta] \rangle \le \langle K, 0 \rangle = 0 \quad \text{or} \quad \langle K, \mathbb{E}[\psi(X)] \rangle \le \langle K, \beta \rangle .$$

Now, we note that $\int p_{\theta,K}(x)\phi(x)d\mu(x) = \alpha$ and $\int p_{\theta,K}(x)\psi(x)d\mu(x) = \beta$ by assumption. Then we have

$$\begin{aligned}
H(P) &\le -D_{\rm kl}\left(P \| P_{\theta,K}\right) - \left[\langle \theta, \alpha \rangle - \langle K, \beta \rangle - A(\theta, K)\right] \\
&= -D_{\rm kl}\left(P \| P_{\theta,K}\right) - \int p_{\theta,K}(x) \left[\langle \theta, \phi(x) \rangle - \langle K, \psi(x) \rangle - A(\theta, K)\right] d\mu(x) \\
&= -D_{\rm kl}\left(P \| P_{\theta,K}\right) - \int p_{\theta,K}(x) \log p_{\theta,K}(x) d\mu(x) = -D_{\rm kl}\left(P \| P_{\theta,K}\right) + H(P_{\theta,K}).
\end{aligned}$$

As $D_{\rm kl}\left(P \| P_{\theta,K}\right) > 0$ unless $P = P_{\theta,K}$, this gives the result. □

Bibliography

[1] S. Amari and H. Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. American Mathematical Society, 2000.

[2] O. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. Wiley, 1978.

[3] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[4] L. D. Brown. *Fundamentals of Statistical Exponential Families*. Institute of Mathematical Statistics, 1986.

[5] T. M. Cover and J. A. Thomas. *Elements of Information Theory, Second Edition*. Wiley, 2006.

[6] E. T. Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9): 939–952, Sept. 1982.

[7] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.