A Representation Approach for Relative Entropy Minimization with Expectation Constraints

Oluwasanmi Koyejo SANMI.K@UTEXAS.EDU
Department of Electrical and Computer Engineering, University of Texas, Austin, TX 78703

Joydeep Ghosh GHOSH@ECE.UTEXAS.EDU
Department of Electrical and Computer Engineering, University of Texas, Austin, TX 78703

Abstract

We consider the general problem of relative entropy minimization and entropy maximization subject to expectation constraints. We show that the solutions can be represented as members of an exponential family subject to weaker conditions than previously shown, and the representation can be simplified further if an appropriate conjugate prior density is used. As a result, the solutions can be found by optimization with respect to members of the parametric families corresponding to these representations.

1. Introduction

An important conceptual framework for statistical inference is the *principle of maximum entropy* (Jaynes, 1957), which defines a procedure for estimating probability distributions with a minimal set of assumptions other than constraints. A generalization of this concept is the *principle of minimum discrimination information* (Kullback, 1959), where the inference task involves estimating a distribution that satisfies a set of constraints so the estimate is as hard as possible to differentiate from a prior distribution. The constraints are typically specified as restrictions on the expectations of *feature functions*, constrained to match empirical averages computed from observations, or to match more general restrictions that incorporate structure determined by domain knowledge.

It is well known that the maximum entropy and relative minimum entropy distributions subject to equality constraints can be represented as members of the exponential family (Cover & Thomas, 2006) and similar results have been shown for norm ball constraints and other convex constraint sets (Altun & Smola, 2006). Constrained relative entropy minimization and constrained entropy maximization have been studied in several domains [1] including natural language processing (Berger et al., 1996) and ecology (Dudík et al., 2007). Other applications in the machine learning literature include maximum entropy discrimination (MED) and related models where margin constraints are applied for classification (Jaakkola et al., 1999; Zhu et al., 2009), collaborative filtering (Xu et al., 2012) and link prediction (Zhu, 2012).

We consider the general class of problems involving the estimation of the distribution that minimizes the relative entropy to a given prior distribution subject to expectation constraints. Entropy maximization is discussed as a special case of this framework when the prior distribution is uniform. The contributions of this paper are as follows:

- We show that the solution of the relative entropy minimization problem and the solution of the entropy maximization problem subject to expectation constraints can be represented as members of an exponential family. Our results relax the necessary conditions on the constraint set such as convexity.
- We show the existence of conjugate prior distributions for the relative entropy minimization problem. Here, conjugacy implies that the solutions can be represented as members of the conjugate prior distribution family. We propose a paramet-

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28. Copyright 2013 by the author(s).

[1] There are many names in the literature for this class of methods. To avoid confusion, we will not use these names in this paper. For instance, entropy maximization is often referred to as *maxent*. Relative entropy minimization is often referred to as *minxent* or*generalized maxent*.

ric family of conjugate prior distributions.

- We provide sufficient conditions on a subset of distributions so that the solution of the constrained relative entropy minimization and constrained entropy maximization problem can be found by direct optimization with respect to the members of the set.

- We show that these conditions are satisfied by the exponential family corresponding to the representation of the solution, and by the parametric family of the prior distribution if the prior distribution is conjugate. This allows for the application of the large class of efficient optimization tools for parametric optimization (Nocedal & Wright, 2006).

This paper is organized as follows. We begin the with preliminaries (Section 2), background and related work (Section 3). The representation of the minimizer for relative entropy minimization is studied in Section 4 and the representation of the solution that results from the use of a conjugate prior is discussed in Section 4.1. We discuss the optimization approach inspired by these representations in Section 4.2. The special case of entropy maximization is studied in Section 5.

2. Preliminaries

We begin by stating a few basic definitions from convex analysis and probability theory that will be useful in the development of our approach. Let $\mathcal{X}$ be a Banach space and let $\mathcal{X}^*$ be the *dual space* of $\mathcal{X}$. The *Legendre-Fenchel* transformation (convex conjugate) of a function $f : \mathcal{X} \mapsto [-\infty, +\infty]$ is given by $f^* : \mathcal{X}^* \mapsto [-\infty, +\infty]$ as:

$$f^*(x^*) = \sup_{x \in \mathcal{X}} \{\langle x, x^* \rangle - f(x)\}.$$

where $\langle x, x^* \rangle$ denotes the dual pairing. If $\mathcal{B}$ is complete with respect to $\langle \cdot, \cdot \rangle$, then $\mathcal{B}$ is a Hilbert space. See Borwein & Zhu (2005) for further details on Fenchel duality, particularly as applied to variational analysis.

Let $\mathbb{X} \ni x$ denote a sample space and $\mathcal{P} = \{p \,|\, p(x) \geq 0, \int_{\mathbb{X}} p(x)dx = 1\}$ denote the set of probability densities on $\mathbb{X}$. The *relative entropy*, also known as the *Kullback-Leibler divergence* (KL divergence) of the continuous probability measure P with respect to a *background* measure μ is given by:

$$\mathrm{KL}\left(P \| \mu\right) = \int_{\mathbb{X}} \frac{\partial P}{\partial \mu}(x) \log \frac{\partial P}{\partial \mu}(x) \partial \mu(x),$$

where $\frac{\partial P}{\partial \mu}$ is the Radon-Nikodym derivative. In this paper, we assume that P is absolutely continuous with respect to the measure so there exists a density $p \in \mathcal{P}$. We will be primarily interested in two cases. First, when μ is the Lebesgue measure, the relative entropy reduces to the *differential entropy* given by $\mathrm{H}\left(P\right)$. To simplify notation, we denote the differential entropy by:

$$\mathrm{H}\left(p\right) = -\int_{\mathbb{X}} p(x) \log p(x) dx,$$

which reduces to the *entropy* $\mathrm{H}\left(p\right) = -\sum_{\mathbb{X}} p(x) \log p(x)$ for the counting measure on a discrete set.

The other case of interest is when the background measure is also a probability measure Q that induces a probability density $q \in \mathcal{P}$, and P is absolutely continuous with respect to Q. Here, our using simplified notation, the relative entropy is given in terms of the densities as:

$$\mathrm{KL}\left(p \| q\right) = \int_{\mathbb{X}} p(x) \log \frac{p(x)}{q(x)} dx$$

The relative entropy is *strictly convex* with respect to its first argument.

The *expectation* is a linear operator given by[2] $\mathrm{E}_{x \sim P}\left[\, f(x) \,\right] = \int_{\mathbb{X}} p(x) f(x) dx$. With some abuse of notation, we specify the expectation as $\mathrm{E}_p\left[\, f \,\right] = \mathrm{E}_{x \sim P}\left[\, f(x) \,\right]$.

Given a density p in a parametrized family $\mathcal{G} \ni p$ and a function $f : \mathbb{X} \mapsto \mathbb{R}$ (known as the likelihood function in Bayesian statistics), p is the *Bayesian conjugate prior* of f if there exists a density q that satisfies: (i) $q(x) \propto p(x) f(x)$ and, (ii) $q \in \mathcal{G}$. In this paper, we take $a(x) \propto b(x)$ to imply the existence of a constant C which is independent of x (though C may be a function of other parameters) such that $a(x) = Cb(x)$. Further discussion on Bayesian conjugacy may be found in (Raïffa & Schlaifer, 1968).

The *exponential family* are a set of probability distributions whose density functions take the form:

$$p_{\boldsymbol{\theta}}(x) = h(x) e^{\langle \boldsymbol{\eta}(\boldsymbol{\theta}), \mathbf{t}(x) \rangle - A(\boldsymbol{\theta})},$$

$\boldsymbol{\eta}(\boldsymbol{\theta})$ is known as the natural parameter vector, $\mathbf{t}(x)$ are the sufficient statistics, $h(x)$ is known as the base measure, and $A(\boldsymbol{\theta})$ is the the log-partition function. The exponential family is in *canonical* form if $\boldsymbol{\eta}(\boldsymbol{\theta}) = \boldsymbol{\theta}$. The domain of the parameter $\boldsymbol{\theta}$ is a convex set defined as:

$$\boldsymbol{\Theta} = \left\{ \boldsymbol{\theta} \,\middle|\, \int_{\mathbb{X}} h(x) e^{\langle \boldsymbol{\eta}(\boldsymbol{\theta}), \mathbf{t}(x) \rangle} dx < \infty \right\}.$$

[2] We will not distinguish notation between the random variable and the samples when such confusion is unlikely.

The log-partition function is a convex function defined on the domain $\mathbf{\Theta}$. Examples of exponential family distributions in common use include the Gaussian, Bernoulli and Poisson distributions. Further details on exponential family distributions may be found in Brown (1986).

3. Background and Related Work

In the following, we provide an overview of relevant results from Altun & Smola (2006). Let $\boldsymbol{\beta}$ represent *feature functions* that map $\mathbb{X}$ to a feature space with components $\boldsymbol{\beta}(x) = \{\beta_j(x)\}$. We let p denote the *prior* distribution corresponding to the probability distribution on $\mathbb{X}$ without constraints, also known as the *default* distribution. Altun & Smola (2006) considered divergence minimization subject to norm ball constraints given by $\|\mathrm{E}_q[\boldsymbol{\beta}] - \mathbf{b}\|_{\mathcal{B}} \leq \epsilon$ where $\|\cdot\|_{\mathcal{B}}$ is the norm ball on a the Banach space $\mathcal{B}$ centered at $\mathbf{b} \in \mathcal{B}$, and $\epsilon \geq 0$ is the width. The solution was found by an elegant application of Fenchel duality[3] for variational optimization (Borwein & Zhu, 2005) The following Lemma characterizes relative entropy minimization subject to norm ball constraints.

Lemma 1 (Altun & Smola (2006))

$$\min_{q\in\mathcal{P}} \mathrm{KL}\left(q\|p\right) \; s.t. \; \|\mathrm{E}_q\left[\boldsymbol{\beta}\right] - \mathbf{b}\|_{\mathcal{B}} \leq \epsilon \tag{1}$$

$$= \max_{\boldsymbol{\lambda}} \langle \boldsymbol{\lambda}, \mathbf{b}\rangle - \log \int_{\mathbb{X}} p(x) e^{\langle \boldsymbol{\lambda}, \boldsymbol{\beta}(x)\rangle} dx - \epsilon\|\boldsymbol{\lambda}\|_{\mathcal{B}^*} + e^{-1} \tag{2}$$

and the unique solution is given by:

$$q_*(x) = p(x) e^{\langle \boldsymbol{\lambda}_*, \boldsymbol{\beta}(x)\rangle - G(\boldsymbol{\lambda}_*)} \tag{3}$$

where $\boldsymbol{\lambda}_*$ is the solution of the dual optimization (2) and $G(\boldsymbol{\lambda}_*)$ ensures normalization.

There may be several equivalent representations for a given density $q \in \mathcal{P}$. However, Lemma 1 shows that the density that minimizes relative entropy subject to norm ball constraints, if it exists, has a canonical representation a member of the exponential family with base measure p, natural statistics $\boldsymbol{\beta}$ and parameters $\boldsymbol{\lambda}_*$. The conditions for Lemma 1 include constraint qualification, which informally requires the existence of densities that satisfy the set of constraints, and a finite cost function (2) at the solution $\boldsymbol{\lambda}_*$. More details are given in Altun & Smola (2006) and Chapter 4 of Borwein & Zhu (2005). Given M sample observations $\{\tilde{z}_i\}$, the center $\mathbf{b}$ is typically specified as an empirical average $\mathbf{b} = \sum_{m=1}^{M} \boldsymbol{\beta}(\tilde{z}_i)$. Thus, if the empirical center is used, we may interpret Lemma 1 as a mechanism for estimating the closest (in terms of KL divergence) distribution to a prior distribution such that its expectations approximately satisfy the empirical averages.

A important special case of Lemma 1 is when the default distribution is uniform with respect to the background measure. This characterizes the case with no a-priori assumptions. The solution is given by the following Lemma.

Lemma 2 (Altun & Smola (2006))

$$\min_{q\in\mathcal{P}} -\mathrm{H}\left(q\right) \; s.t. \; \|\mathrm{E}_q\left[\boldsymbol{\beta}\right] - \mathbf{b}\|_{\mathcal{B}} \leq \epsilon \tag{4}$$

$$= \max_{\boldsymbol{\lambda}} \langle \boldsymbol{\lambda}, \mathbf{b}\rangle - \log \int_{\mathbb{X}} e^{\langle \boldsymbol{\lambda}, \boldsymbol{\beta}(x)\rangle} dx - \epsilon\|\boldsymbol{\lambda}\|_{\mathcal{B}^*} + e^{-1} \tag{5}$$

and the unique solution is given by:

$$q_*(z) = e^{\langle \boldsymbol{\lambda}_*, \boldsymbol{\beta}(z)\rangle - G(\boldsymbol{\lambda}_*)} \tag{6}$$

where $\boldsymbol{\lambda}_*$ is the solution of the dual optimization (5) and $G(\boldsymbol{\lambda}_*)$ ensures normalization.

It follows that the density that maximizes the differential entropy subject to expectation constraints in the norm ball, if it exists, is a member of the exponential family. When the sample space is discrete, the density (6) is sometimes known as the Gibbs distribution (Dudík et al., 2007).

The following well studied examples illustrate the principle of maximum entropy. Details may be found in (Cover & Thomas, 2006; MacKay, 2003).

Example 1 *Let $\mathbb{X} = \mathbb{R}$ and define the feature functions as the first and second order functions $\boldsymbol{\beta}(x) = \left[\begin{smallmatrix} x \\ x^2 \end{smallmatrix}\right]$ i.e. $\beta_1(x) = x$ and $\beta_2(x) = x^2$. Let the constraint set correspond to equality constraints given by $\mathrm{E}\left[\boldsymbol{\beta}(x)\right] = \mathbb{E}\left[\begin{smallmatrix} x \\ x^2 \end{smallmatrix}\right] = \left[\begin{smallmatrix} \mu \\ \mu^2+\sigma^2 \end{smallmatrix}\right]$, then the constrained maximum entropy solution is a Gaussian distribution with mean μ and variance σ^2. Recall that for a Gaussian random variable $x \sim \mathcal{N}\left(\mu, \sigma^2\right)$, $\mathrm{E}\left[x\right] = \mu$ and $\mathrm{E}\left[x^2\right] = \mu^2 + \sigma^2$. The entropy of the resulting distribution is given by σ^2.*

Example 2 *Let $\mathbb{X} = \mathbb{R}_+$ i.e. the positive reals, and define the feature function as $\beta(x) = x$, and let the constraint set correspond to equality constraints $\mathrm{E}\left[\beta(x)\right] = \frac{1}{c}$, then the constrained maximum entropy solution is a exponential distribution with rate parameter c.*

[3] The Fenchel dual approach applied to infinite dimensional spaces simplifies the analysis required to ensure continuity and differentiability compared to the Lagrange dual approach using variational calculus.

Much of the research on constrained relative entropy minimization and constrained entropy maximization have focused on finite dimensional discrete sample spaces, particularly in the natural language processing domain (Berger et al., 1996). Dudík et al. (2007) studied the constrained relative entropy minimization for finite sample spaces where the constraint set is given by a l_1, l_2 or $l_1 + l_2$ norm ball. The maximum entropy discrimination (MED) framework (Jaakkola et al., 1999) is a special case of relative entropy minimization for classification where margin constraints are imposed on the log likelihood ratios corresponding to each class. Jaakkola et al. (1999) solved the constrained maximum entropy problem by direct optimization of the dual cost function. This direct approach requires complicated high dimensional integrals and is often intractable for complicated constraint set, or with a complicated prior distribution. More recently, researchers have applied constrained relative

4. Exponential Family Representation

We propose an exponential family representation for the distribution that minimizes the relative entropy subject to expectation constraints by extending the known representation results for the equality constrained case. This paper will focus on expectation constraints of the form $\mathrm{E}\left[\boldsymbol{\beta}\right] \in \mathsf{C}$ where $\mathsf{C} \subset \mathcal{B}$ is a constraint set of interest. Relative entropy minimization subject to expectation constraints involves solving the variational optimization problem:

$$\inf_{q \in \mathcal{P}} \mathrm{KL}\left(q \| p\right) \text{ s.t. } \mathrm{E}_q\left[\boldsymbol{\beta}\right] \in \mathsf{C}, \tag{7}$$

Thus, the constrained entropy minimization problem may be regarded as an information projection of the prior density p to the set of distributions that satisfy the constraint $\mathrm{E}\left[\boldsymbol{\beta}\right] \in \mathsf{C}$ (see Fig. 1). Let q_* denote a solution of the constrained minimum relative entropy optimization. By construction, each q_* satisfies the constraints $\mathrm{E}_{q_*}\left[\boldsymbol{\beta}\right] = \mathbf{a}_*$ for an $\mathbf{a}_* \in \mathsf{C}$.

Given a fixed $\mathbf{c} \in \mathsf{C}$, the density that minimizes the relative entropy minimization subject to equality constraints, if it exists, is given by the argument at the solution of:

$$\inf_{q \in \mathcal{P}} \left[\mathrm{KL}\left(q \| p\right) \text{ s.t. } \mathrm{E}_q\left[\boldsymbol{\beta}\right] = \mathbf{c}\right]. \tag{8}$$

For the rest of the paper, we assume that the set of solutions q_* is not empty so the optimization problem is well defined. This implies the existence of at least one density $q \in \mathcal{P}$ that satisfies the constraints $\mathrm{E}_q\left[\boldsymbol{\beta}\right] \in \mathsf{C}$. This also implies that the infimum of the variational optimization problem (7) is achieved at a density q_*. Further, we assume that for each solution q_*, the expectation $\mathrm{E}_{q_*}\left[\boldsymbol{\beta}\right] = \mathbf{a}_*$ is bounded to avoid the degenerate problem of unbounded constraints. Finally, we assume that $\mathsf{C} \subset \mathcal{B}$ is a closed set. This assumption is mostly for convenience and clarity and can easily be relaxed.

Let $\mathcal{S} = \{q_*\}$ represent the set of solutions of the constrained relative entropy minimization problem (7). The following proposition formally states the representation result, and is a direct consequence of Lemma 1 applied to the optimization problem of (8) by setting the width of the ball to $\epsilon = 0$.

Proposition 3 $\exists\ \mathbf{a}_* \in \mathsf{C}$ *such that each density* $q_* \in \mathcal{S}$ *that optimizes the constrained relative entropy:*

$$q_* = \arg\min_{q \in \mathcal{P}} \left[\mathrm{KL}\left(q \| p\right)\ s.t.\ \mathrm{E}_q\left[\boldsymbol{\beta}\right] \in \mathsf{C}\right]$$

can be represented in the parametric form:

$$q_* = p(x) e^{\langle \boldsymbol{\lambda}_{\mathbf{a}_*}, \boldsymbol{\beta}(x) \rangle - G(\boldsymbol{\lambda}_{\mathbf{a}_*})}$$

where $\boldsymbol{\lambda}_{\mathbf{a}_*}$ *is the solution of* (2) *with* $\epsilon = 0$ *and* $G(\boldsymbol{\lambda}_{\mathbf{a}_*})$ *ensures normalization.*

It follows that any solution $q_* \in \mathcal{S}$ has a canonical representation in terms of its optimization parameters. The representation is as a member of the exponential family with base measure p, natural statistics $\boldsymbol{\beta}$ and parameters $\boldsymbol{\lambda}_{\mathbf{a}_*}$.

4.1. Conjugate Priors

In Bayesian statistical inference, a distribution is called a *conjugate prior* distribution if, given a likelihood, the posterior distribution is in the same family as the prior distribution (Raïffa & Schlaifer, 1968). We re-use the term "conjugate prior" for relative entropy minimization as such a prior distribution plays a similar role here.

Definition 4 *Let* $\mathcal{G}$ *represent a family of distributions. The prior distribution* $p \in \mathcal{G}$ *is a relative entropy conjugate distribution if any solution of* q_* *of* (7) *can be represented as a member of* $\mathcal{G}$.

Readers familiar with Bayesian inference will note the close similarity of the relative entropy solution with the Bayesian posterior. The following Theorem highlights this relationship in the case of conjugate prior distributions.

Theorem 5 *Let* p *denote the prior density and* $\boldsymbol{\beta}$ *denote the feature functions. Given* $\boldsymbol{\lambda}_*$, *let* f *denote any Bayesian likelihood function that satisfies*

$f \propto e^{\langle \boldsymbol{\lambda}_*, \boldsymbol{\beta}(x)\rangle}$. If p is a Bayesian conjugate prior distribution to f with posterior g, then p is a relative entropy conjugate prior distribution and the solution $q_* = g$

Proof We give a simple constructive proof. Let g denote the Bayesian posterior. Recall (Raïffa & Schlaifer, 1968) that the Bayesian posterior satisfies $g(x) \propto p(x)f(x)$. The existence of f implies that $g(x) \propto p(x)f(x) \propto p(x)e^{\langle \boldsymbol{\lambda}_*, \boldsymbol{\beta}(x)\rangle} \propto q_*$. Further, the existence of g implies that $\int_{\mathbb{X}} p(x)f(x) < \infty$, thus $\int_{\mathbb{X}} p(x)e^{\langle \boldsymbol{\lambda}_*, \boldsymbol{\beta}(x)\rangle} < \infty$. It follows from the normalization of probabilities that $g = q_*$. ■

Surprisingly, it turns out that the class of Bayesian conjugate priors is un-necessarily restrictive, as other distributions that do not satisfy Theorem 5 may still satisfy the conditions of relative entropy conjugacy. We propose a larger of relative entropy conjugate prior distributions given by:

$$p_{\boldsymbol{\eta},\nu}(x) = h(x,\nu)e^{\langle \boldsymbol{\eta}, \boldsymbol{\beta}(x)\rangle - D(\boldsymbol{\eta},\nu)}. \tag{9}$$

The following Lemma shows that the solutions q_* may be represented as member of the same family as the prior.

Lemma 6 *Let the prior density p be a member of the conjugate parametric family (9), then any solution q_* of the constrained minimum relative entropy problem can be represented as member of the same parametric family as the prior density.*

Proof The solutions q_* are given by Proposition 3. With the prior fixed in the conjugate family (9), the solutions takes the form:

$$\begin{aligned} q_*(z) &= \left(h(x,\nu)e^{\langle \boldsymbol{\eta}, \boldsymbol{\beta}(x)\rangle - D(\boldsymbol{\eta},\nu)}\right) e^{\langle \boldsymbol{\lambda}_*, \boldsymbol{\beta}(x)\rangle - G(\boldsymbol{\lambda}_*)} \\ &\propto h(x,\nu)e^{\langle \boldsymbol{\eta}, \boldsymbol{\beta}(x)\rangle + \langle \boldsymbol{\lambda}_*, \boldsymbol{\beta}(x)\rangle} \\ &= h(x,\nu)e^{\langle \boldsymbol{\eta}+\boldsymbol{\lambda}_*, \boldsymbol{\beta}(x)\rangle} \\ &= h(x,\nu)e^{\langle \boldsymbol{\theta}_*, \boldsymbol{\beta}(x)\rangle} \end{aligned}$$

where $\boldsymbol{\theta}_* = \boldsymbol{\eta} + \boldsymbol{\lambda}_*$. We note that if $\int_{\mathbb{X}} p(x)e^{\langle \boldsymbol{\lambda}_*, \boldsymbol{\beta}(x)\rangle} < \infty$, as required for $\boldsymbol{\lambda}_*$ to be a valid solution, then it follows that $\int_{\mathbb{X}} h(x,\nu)e^{\langle \boldsymbol{\theta}_*, \boldsymbol{\beta}(x)\rangle} < \infty$ as they only differ by the constant scaling $e^{-D(\boldsymbol{\eta},\nu)}$. The log-partition function is fully determined given other parameters and it takes the same parametric form as the log-partition function of the prior. Hence any q_* corresponds to a solution:

$$q_*(z) = h(x,\nu)e^{\langle \boldsymbol{\theta}_*, \boldsymbol{\beta}(x)\rangle - D(\boldsymbol{\theta}_*,\nu)}$$

represented as a member of the prior parametric family. ■

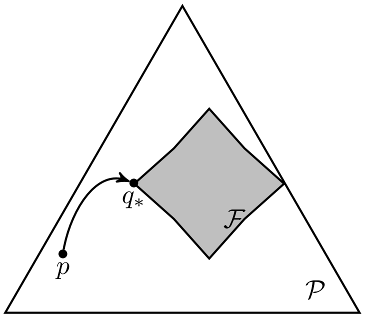

Figure 1. Constrained Relative Entropy Minimization as information projection. $\mathcal{P}$ is the probability space, $\mathcal{F}$ is the feasible set corresponding to the constraints C, p is the prior density. The resulting solution given by q_* is the information projection of p onto the set of distributions $\mathcal{F}$.

When $h(x,\nu) = \tilde{h}(x)^\nu$, the family of densities represented by (9) corresponds to the Bayesian conjugate prior for the exponential family likelihood $f = \tilde{h}(x)b(\boldsymbol{\lambda})e^{\langle \boldsymbol{\lambda}, \boldsymbol{\beta}(x)\rangle}$. We note that this family does not satisfy the conditions of Bayesian conjugacy in Theorem 5 as the density depends on a base measure $\tilde{h}$.

As in the Bayesian inference, the choice of prior distribution is a difficult one. The prior distribution should be determined based on a-priori knowledge of the problem and the data. With limited prior knowledge, the feature functions capture most of what one assumes about the distribution. The conjugate prior can be motivated as a natural choice as it intrinsically captures the distribution induced by the feature functions, although it typically includes a bias e.g. normalization factors such as a choice of base measure. Further, Theorem 5 and Lemma 6 suggest computational tractability if a conjugate prior distribution is chosen.

4.2. Representation Approach for Constrained Relative Entropy Minimization

While Proposition 3 characterizes the representation of the solutions, we will be interested in constrained relative entropy minimization by optimization with respect to the members of specified families. Towards this end, we define the *feasible set* of distributions that satisfy the constraints. Let $\mathsf{A} \subset \mathsf{C}$ represent the set of points $\mathbf{c} \in \mathsf{C}$ where $\mathbf{c}$ is bounded, and the optimization problem of (8) is finite and attained. Assuming the existence of at least one solution q_*, it follows that the set A is not empty. We associate a density function $q_{\mathbf{c}}$ with every element $\mathbf{c} \in \mathsf{A}$ given by the solution of (8). It follows that for $\mathbf{c} \in \mathsf{A}$, the minimizer is given by:

$$q_{\mathbf{c}}(x) = p(x)e^{\langle \boldsymbol{\lambda}_{\mathbf{c}}, \boldsymbol{\beta}(x)\rangle - G(\boldsymbol{\lambda}_{\mathbf{c}})}, \tag{10}$$

where $\boldsymbol{\lambda}_{\mathbf{c}}$ is the solution of the dual optimization (2) with $\epsilon = 0$. It can be shown that the solutions q_* correspond to the solutions of:

$$\min_{\mathbf{c}\in\mathsf{A}} \left[\min_{q\in\mathcal{P}} \mathrm{KL}\left(q\|p\right) \text{ s.t. } \mathrm{E}_q\left[\boldsymbol{\beta}\right] = \mathbf{c}\right].$$

We define the *feasible set* as the set of densities $\mathcal{F} = \{q_{\mathbf{c}} \,|\, \mathbf{c} \in \mathsf{A}\}$. The following theorem characterizes the solution in terms of the feasible set.

Theorem 7 *Let $\mathcal{F} = \{q_{\mathbf{c}} \,|\, \mathbf{c} \in \mathsf{A}\}$ denote the feasible set. Each member of the feasible set satisfies the representation (10). Each solution of the constrained minimum relative entropy problem is given by:*

$$q_* = \arg\min_{q\in\mathcal{F}} \mathrm{KL}\left(q\|p\right)$$

and $\exists\ \mathbf{a}_* \in \mathsf{A}$ s.t. $q_* = q_{\mathbf{a}_*}$.

Proof The exponential family representation of the feasible set is a direct application of Lemma 1 with $\epsilon = 0$. We prove that $q_* \in \mathcal{F}$ by contradiction. Suppose $q_* \notin \mathcal{F}$, then $q_* = q_{\mathbf{v}}$ for $\mathbf{v} \notin \mathsf{A}$. This is a contradiction by definition of A. ■

The representations shown in Section 4 and Section 4.1 inspire a novel approach for optimization with respect to the members of the parametric families containing the solution. Further, the representation approach may lead to a simplified optimization problem, particularly with an appropriate choice of prior distribution. To this end, we specify sufficient conditions on any subset of densities so that it contains the solutions.

Theorem 8 *Let $\mathcal{Q} \subset \mathcal{P}$ specify a subset of distributions. If the set of solutions $\mathcal{S} \subset \mathcal{Q}$, then the solutions q_* of the constrained relative entropy problem are given by:*

$$f_* = \arg\min_{f\in\mathcal{Q}} \mathrm{KL}\left(f\|p\right) \textit{ s.t. } \mathrm{E}_f\left[\boldsymbol{\beta}\right] \in \mathsf{C}$$

Proof If every solution $f_* \in \mathcal{F}$, then the proof follows directly from Theorem 7. If any $f_* \notin \mathcal{F}$, feasibility implies that f_* satisfies the constraints. Thus $\exists\ \mathbf{v}$ such that $\mathrm{E}_{f_*}\left[\boldsymbol{\beta}\right] = \mathbf{v}$ and $\mathbf{v} \notin \mathsf{A}$. This is a contradiction by definition of A. ■

One implication of Theorem 8 is that to find q_*, we may optimize over any set of distributions we choose as long as the choice contains the set of solutions $\mathcal{S}$.

We now return to the exponential family representation of Section 4. An important insight from Theorem 7 is that the members of the feasible set $\mathcal{F} = \{q_{\mathbf{c}} \,|\, \mathbf{c} \in \mathsf{A}\}$ are members of the same parametric family of distributions. determined by the prior density p and the feature functions $\boldsymbol{\beta}$ as:

$$\mathcal{F} = \left\{ q \;\middle|\; f_{\boldsymbol{\lambda}_{\mathbf{c}}}(x) = p(z) e^{\langle \boldsymbol{\lambda}_{\mathbf{c}}, \boldsymbol{\beta}(x)\rangle - G(\boldsymbol{\lambda}_{\mathbf{c}})}\ \forall\ \mathbf{c} \in \mathsf{A} \right\}.$$

The results of Theorem 8 and the parametric representation of feasible set suggests direct optimization over members of the exponential family as an approach for solving the the constrained entropy minimization problem. Let $\mathcal{E}$ specify the exponential family of distributions with measure p, natural statistics $\boldsymbol{\beta}$, parameters $\boldsymbol{\lambda}$ and log-partition function $G(\boldsymbol{\lambda})$. The following corollary shows that we can solve the constrained relative entropy minimization problem by direct optimization over the parametric family $\mathcal{E}$.

Corollary 9 *Any solution of the constrained minimum relative entropy problem $q_* \in \mathcal{S}$ is given by $q_* = f_{\boldsymbol{\lambda}_*} \in \mathcal{E}$ where:*

$$\begin{aligned} f_{\boldsymbol{\lambda}_*} &= \arg\min_{f_{\boldsymbol{\lambda}}\in\mathcal{E}} \left[\mathrm{KL}\left(f_{\boldsymbol{\lambda}}\|p\right) \textit{ s.t. } \mathrm{E}_{f_{\boldsymbol{\lambda}}}\left[\boldsymbol{\beta}\right] \in \mathsf{C}\right], \\ &\exists\ \boldsymbol{\lambda}_* \in \boldsymbol{\Lambda} \textit{ s.t.} \\ \boldsymbol{\lambda}_* &= \arg\min_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}} \left[\mathrm{KL}\left(f_{\boldsymbol{\lambda}}\|p\right) \textit{ s.t. } \mathrm{E}_{f_{\boldsymbol{\lambda}}}\left[\boldsymbol{\beta}\right] \in \mathsf{C}\right]. \end{aligned}$$

Proof Each parameter $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$ corresponds to a distribution $f_{\boldsymbol{\lambda}} \in \mathcal{E}$. Further, each $\tilde{\boldsymbol{\lambda}}_*$ corresponding to q_* must satisfy $\tilde{\boldsymbol{\lambda}}_* \in \boldsymbol{\Lambda}$ for a solution to exist. Thus, by construction, we have that $\mathcal{S} \subset \mathcal{F} \subset \mathcal{E}$. The proof follows from Theorem 8. ■

Convexity of C is sufficient for uniqueness of q_*. This follows from the strict convexity of the relative entropy. For uniqueness of the parameter $\boldsymbol{\lambda}_*$ it is sufficient that q_* is unique (alternatively, that $\boldsymbol{\lambda}_*$ is unique), and the map $\boldsymbol{\lambda} \mapsto f_{\boldsymbol{\lambda}}$ is a bijection for $f_{\boldsymbol{\lambda}} \in \mathcal{S} \subset \mathcal{E}$. For exponential family distributions, a sufficient condition for the uniqueness of the parameter mapping is affine independence of the feature functions $\boldsymbol{\beta}$.

The solution of Corollary 9 requires the evaluation of the log partition function $G(\boldsymbol{\lambda})$ which, in turn requires an integral that may be intractable. However, this computation is unnecessary if the prior distribution is a relative entropy conjugate prior with a tractable family. Alternately, the computation is simplified if computation and optimization of the KL divergence between members of the prior distribution parametric family is straightforward. The following corollary characterizes the parametric optimization solution using relative entropy conjugate priors.

Corollary 10 *Let $\mathcal{G}$ denote the set of parametrized distributions corresponding to a conjugate relative entropy prior p, and let $\boldsymbol{\Theta} \ni \boldsymbol{\theta}$ specify the domain of its parameters. Any solution of the constrained minimum relative entropy density $q_* \in \mathcal{S}$ is given by $q_* = f_{\boldsymbol{\theta}_*} \in \mathcal{G}$ where:*

$$\begin{aligned} f_{\boldsymbol{\theta}_*} &= \operatorname*{arg\,min}_{f_{\boldsymbol{\theta}} \in \mathcal{G}} \Big[\mathrm{KL}\left(f_{\boldsymbol{\theta}} \| p\right) \textit{ s.t. } \mathrm{E}_{f_{\boldsymbol{\theta}}}\left[\boldsymbol{\beta}\right] \in \mathsf{C} \Big], \\ &\exists\, \boldsymbol{\theta}_* \in \boldsymbol{\Theta} \textit{ s.t.} \\ \boldsymbol{\theta}_* &= \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \Big[\mathrm{KL}\left(f_{\boldsymbol{\theta}} \| p\right) \textit{ s.t. } \mathrm{E}_{f_{\boldsymbol{\theta}}}\left[\boldsymbol{\beta}\right] \in \mathsf{C} \Big]. \end{aligned}$$

Proof Each parameter $\boldsymbol{\theta}$ corresponds to a distribution $f_{\boldsymbol{\theta}} \in \mathcal{G}$. From Definition 4, we have that $\mathcal{S} \subset \mathcal{G}$. The proof follows from Theorem 8. ■

Stronger conditions are required for uniqueness of the solution. We briefly outline some sufficient conditions. $q_* = f_{\boldsymbol{\theta}_*}$ is unique if C is convex, and $\boldsymbol{\theta}_*$ is unique if $\boldsymbol{\theta} \mapsto f_{\boldsymbol{\theta}}$ is a bijection for $f_{\boldsymbol{\theta}} \in \mathcal{S} \subset \mathcal{G}$.

The approach outlined in this section simplifies the estimation of the distribution that minimizes the relative entropy to a specified prior distribution while satisfying a set of constraints. Constraint sets of interest may include sparsity constraints, or low rank constraints for matrix-variate data. We illustrate the utility of the results with an example.

Example 3 *Consider the constrained relative entropy minimization where the prior $p = \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ is a multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Let the constraints be given by $\mathrm{E}_q\left[\mathbf{x}\right] \in \mathsf{C}$. We note that the Gaussian prior distribution satisfies the requirements of Lemma 6 with feature functions $\boldsymbol{\beta}(\mathbf{x}) = \mathbf{x}$. Hence, the prior distribution is a conjugate relative entropy prior distribution, and the solution will be a Gaussian distribution. The constraint $\mathrm{E}_q\left[\mathbf{x}\right] \in \mathsf{C}$ corresponds to the mean constraint. Applying Corollary 10 shows that that the solution is given by $q_* = \mathcal{N}\left(\mathbf{m}_*, \mathbf{S}_*\right)$, and its parameters can be found by minimizing the KL divergence between the Gaussian distributions:*

$$\mathbf{m}_*, \mathbf{S}_* = \operatorname*{arg\,min}_{\{\mathbf{m}, \mathbf{S}\} \in \mathcal{M}} \mathrm{KL}\left(\mathcal{N}\left(\mathbf{m}, \mathbf{S}\right) \| \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)\right) \textit{ s.t. } \mathbf{m} \in \mathsf{C}.$$

where $\mathcal{M}$ represents the domain of mean parameters for the Gaussian distribution.

5. Parametric Representation for Constrained Entropy Maximization

Differential entropy maximization subject to expectation constraints is a special case of (7) where the default distribution is uniform with respect to the background measure on $\mathbb{X}$ and involves solving the optimization problem:

$$\begin{aligned} &\sup_{q \in \mathcal{P}} \mathrm{H}\left(q\right) \text{ s.t. } \mathrm{E}_q\left[\boldsymbol{\beta}\right] \in \mathsf{C} \\ &= \inf_{q \in \mathcal{P}} \; - \mathrm{H}\left(q\right) \text{ s.t. } \mathrm{E}_q\left[\boldsymbol{\beta}\right] \in \mathsf{C}. \end{aligned} \tag{11}$$

The problem can be solved using identical steps as the general constrained relative entropy minimization. Given a point $\mathbf{c} \in \mathsf{C}$, the constrained entropy maximization problem subject to equality constraints, if it exists, is given by the solution of:

$$\inf_{q \in \mathcal{P}} \Big[- \mathrm{H}\left(q\right) \text{ s.t. } \mathrm{E}_q\left[\boldsymbol{\beta}\right] = \mathbf{c} \Big]. \tag{12}$$

To avoid repetition, we present only main results. We use identical notation to highlight the similarities.

First, we show that the maximum entropy solution can be represented as a member of the exponential family. Let $\mathcal{S} = \{q_*\}$ represent the set of solutions of the constrained entropy maximization problem (12).

Proposition 11 $\exists\, \mathbf{a}_* \in \mathsf{C}$ *such that each density $q_* \in \mathcal{S}$ that optimizes the constrained maximum entropy:*

$$q_* = \operatorname*{arg\,min}_{q \in \mathcal{P}} \Big[- \mathrm{H}\left(q\right) \textit{ s.t. } \mathrm{E}_q\left[\boldsymbol{\beta}\right] \in \mathsf{C} \Big]$$

takes the parametric form:

$$q_* = e^{\langle \boldsymbol{\lambda}_{\mathbf{a}_*}, \boldsymbol{\beta}(z) \rangle - G(\boldsymbol{\lambda}_{\mathbf{a}_*})}$$

where $\boldsymbol{\lambda}_{\mathbf{a}_*}$ is the solution of (5) with $\epsilon = 0$ and $G(\boldsymbol{\lambda}_{\mathbf{a}_*})$ ensures normalization.

Thus, the solutions can be represented as members of the exponential family with natural statistics $\boldsymbol{\beta}$ and parameters $\boldsymbol{\lambda}_{\mathbf{a}_*}$.

In the following, we specify sufficient conditions on any subset $\mathcal{Q} \subset \mathcal{P}$ so that it contains the solutions q_*.

Proposition 12 *Let $\mathcal{Q} \subset \mathcal{P}$ specify an closed set of distributions. If $\mathcal{S} \subset \mathcal{Q}$, then the solutions q_* of the constrained maximum entropy problem are given by:*

$$f_* = \operatorname*{arg\,min}_{f \in \mathcal{Q}} \; - \mathrm{H}\left(f\right) \textit{ s.t. } \mathrm{E}_f\left[\boldsymbol{\beta}\right] \in \mathsf{C}$$

Thus, we may optimize over any set of distributions we choose as long as the choice contains the set of solutions $\mathcal{S}$. The exponential family representation suggests optimization over members of the parametric family given by:

$$\mathcal{E} = \left\{ f_{\boldsymbol{\lambda}} \,\middle|\, f_{\boldsymbol{\lambda}} = e^{\langle \boldsymbol{\lambda}, \boldsymbol{\beta}(z) \rangle - G(\boldsymbol{\lambda})} \, \forall\, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}. \right\}$$

where $\mathbf{\Lambda}$ represents the domain of the natural parameters of $\mathcal{E}$.

Corollary 13 *The constrained maximum entropy solution is given by* $q_* = f_{\boldsymbol{\lambda}_*}$ *where:*

$$\boldsymbol{\lambda}_* = \underset{\boldsymbol{\lambda}\in\mathbf{\Lambda}}{\arg\min}\left[-\,\mathrm{H}\left(f_{\boldsymbol{\lambda}}\right) s.t.\ \mathrm{E}_{f_{\boldsymbol{\lambda}}}\left[\boldsymbol{\beta}\right] \in \mathsf{C}\right],$$

Conjugacy does not apply to this case as the entropy maximization does not include a choice of prior distributions.

6. Conclusion and Future work

In this paper, we considered the problem of estimating of the distribution that minimizes the relative entropy to a given prior distribution subject to expectation constraints, and considered constrained entropy maximization as a special case. We showed that under mild conditions, the solution of these problems can be represented as members of an exponential family. We proposed the use of conjugate prior distributions, and showed that the distribution that minimizes the constrained relative entropy relative to a conjugate prior distribution can be represented as a member of the conjugate distribution family. These observations motivated a novel optimization approach over the members of the respective parametric families. We showed that a parametric optimization approach recovers the solution, and may significantly reduce the complexity of the optimization. Although we have focused on relative entropy, the approach outlined in this paper can be extended to other divergence metrics such as Csiszár and Bregman divergences. The presented results can also be extended to more general compact linear operators, for instance to enforce constraints on the conditional density. Practical applications of these ideas to constrained multitask regression and ranking are presented in an extended version of this paper.

Acknowledgments

Authors acknowledge support from NSF grant IIS 1016614. We thank Gustavo de Veciana for discussions that led to representation approach proposed here, and Suriya Gunasekar for suggesting improvements to the original manuscript.

References

Altun, Yasemin and Smola, Alexander J. Unifying divergence minimization and statistical inference via convex duality. In *COLT*, 2006.

Berger, Adam L., Pietra, Vincent J. Della, and Pietra, Stephen A. Della. A maximum entropy approach to natural language processing. *Comput. Linguist.*, 22 (1):39–71, March 1996. ISSN 0891-2017.

Borwein, J.M. and Zhu, Q.J. *Techniques of Variational Analysis*. CMS Books in Mathematics. Springer, 2005.

Brown, L.D. *Fundamentals of Statistical Exponential Families: With Applications in Statistical Decision Theory*. Ims Lecture Notes-Monograph Ser.: Vol.9. Inst of Mathematical Statistic, 1986.

Cover, Thomas M. and Thomas, Joy A. *Elements of information theory (2. ed.)*. Wiley, 2006.

Dudík, Miroslav, Phillips, Steven J., and Schapire, Robert E. Maximum entropy density estimation with generalized regularization and an application to species distribution modeling. *J. Mach. Learn. Res.*, 8:1217–1260, December 2007. ISSN 1532-4435.

Jaakkola, Tommi, Meila, Marina, and Jebara, Tony. Maximum entropy discrimination. In *NIPS*. MIT Press, 1999.

Jaynes, E. T. Information Theory and Statistical Mechanics. *Physical Review Online Archive (Prola)*, 106(4):620–630, May 1957.

Kullback, Solomon. *Information Theory and Statistics*. Dover, 1959.

MacKay, D.J.C. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

Nocedal, Jorge and Wright, Stephen J. *Numerical optimization*. Springer series in operations research and financial engineering. Springer, New York, NY, 2. ed. edition, 2006.

Raïffa, H. and Schlaifer, R. *Applied statistical decision theory*. M.I.T. Press, 1968.

Xu, Minjie, Zhu, Jun, and Zhang, Bo. Nonparametric max-margin matrix factorization for collaborative prediction. In *Advances in Neural Information Processing Systems 25*, pp. 64–72, 2012.

Zhu, Jun. Max-margin nonparametric latent feature models for link prediction. In *ICML*, 2012.

Zhu, Jun, Ahmed, Amr, and Xing, Eric P. Medlda: maximum margin supervised topic models for regression and classification. In *ICML*, 2009.

Zhu, Jun, Chen, Ning, and Xing, Eric P. Infinite Latent SVM for Classification and Multi-task Learning. In *NIPS*, 2011.