Convex Foundations for Generalized MaxEnt Models

Rafael Frongillo* and Mark D. Reid†

*Microsoft Research, New York City, NY, United States
†The Australian National University & NICTA, Canberra, ACT, Australia

Abstract. We present an approach to maximum entropy models that highlights the convex geometry and duality of GEFs and their connection to Bregman divergences. Using our framework, we are able to resolve a puzzling aspect of the bijection of [1] between classical exponential families and what they call *regular* Bregman divergences. Their regularity condition rules out all but Bregman divergences generated from log-convex generators. We recover their bijection and show that a much broader class of divergences correspond to GEFs via two key observations: 1) Like classical exponential families, GEFs have a "cumulant" C whose subdifferential contains the mean: $\mathbb{E}_{o\sim p_\theta}[\phi(o)] \in \partial C(\theta)$; 2) Generalized relative entropy is a C-Bregman divergence between parameters: $D_F(p_\theta, p_{\theta'}) = D_C(\theta, \theta')$, where D_F becomes the KL divergence for $F = -H$. We also show that every *incomplete* market with cost function C (see [2]) can be expressed as a complete market, where the prices are constrained to be a GEF with cumulant C. This provides an entirely new interpretation of prediction markets, relating their design back to the principle of maximum entropy.

Keywords: convexity, exponential families, Bregman divergences, prediction markets
PACS: Statistics, 02.50.-r; Convex sets, 02.40.Ft; Entropy (in information theory), 89.70.Cf

INTRODUCTION

A key property of exponential families is that they are *maximum entropy* distributions; given a statistic, the distribution with a particular mean of the statistic which is of maximum entropy forms an exponential family [3]. In 2004, Grünwald and Dawid [4] introduced *generalized* exponential families (GEFs) as maximum entropy distributions for other entropy functions — that is, they noted that one may *define* classical exponential families as maximum entropy distributions for the usual Shannon entropy, thus making the generalization to other notions of "entropy" immediate. We will work with these GEFs, though our definition will slightly depart from theirs by following the classical exponential family presentation more closely while emphasizing their convex geometry. Our aim is to highlight connections between those two literatures and show how those connections naturally generalize to GEFs via two results: Theorem 2 generalizes a result of Banerjee et al. [1], showing that there is a bijection between Bregman divergences and classes of GEFs; and Theorem 3 shows that the generalized relative entropy between two members of a GEF is the Bregman divergence with respect to the cumulant of the two corresponding parameters (cf. [5, 6, 7]). Our convex perspective on exponential families also lets us make connections to the prediction market literature (e.g. [2]), with Theorem 4 showing that GEFs naturally arise when relating prices in complete and incomplete versions of the same market.

We now give the basic definitions and results from convex analysis that will be used throughout this paper. Terms not defined here can be found in standard references (e.g., [8, 9]) or in [10]. Throughout, $\mathscr{O}$ is a (possibly infinite) set of outcomes, $\mathscr{P} \subseteq \Delta(\mathscr{O})$ is a convex set of probability measures, and $\mathscr{R}$ is a convex report space. Our results will involve a statistic $\phi : \mathscr{O} \to \mathscr{R} \subset \mathbb{R}^k$, which plays the role of a payoff function in the prediction market setting.

For the majority of the paper, owing to our heavy use of convex conjugate duality, we will work with *dual pairs* of vector spaces [11, §5.14]. These are pairs $(\mathscr{V}, \mathscr{V}^*)$ of vector spaces equipped with a bilinear form $\langle\cdot,\cdot\rangle : \mathscr{V} \times \mathscr{V}^* \to \mathbb{R}$ that *separates points*, that is, $v = 0$ whenever $\langle v, v^*\rangle = 0$ for all $v^* \in \mathscr{V}^*$ and $v^* = 0$ whenever $\langle v, v^*\rangle = 0$ for all $v \in \mathscr{V}$.

Central to our analysis is the notion of a *convex conjugate* $F^* : \mathscr{V}^* \to \overline{\mathbb{R}}$ of a given function $F : \mathscr{V} \to \overline{\mathbb{R}}$ defined in terms of dual pairs via $F^*(v^*) := \sup_{v\in\mathscr{V}} \langle v, v^*\rangle - F(v)$. Here $\overline{\mathbb{R}}$ denotes the extended reals $\overline{\mathbb{R}} \doteq \mathbb{R} \cup \{-\infty, +\infty\}$. A classical result about the convex conjugate F^* is that it "encodes" the gradients of F in the sense that F^* measures the gap between F and any tangent v^* of F at v. Formally, we will make use of this result via the following lemma.

Lemma 1. *If $F : \mathscr{V} \to \overline{\mathbb{R}}$ is convex then $F^*(v^*) = \langle v^*, v\rangle - F(v) \iff v^* \in \partial F_v$ for all $v \in \mathscr{V}$ and $v^* \in \mathscr{V}^*$.*

Although we aim for generality, we note that all results in this paper still hold if one assumes the convex functions we speak of are all proper, strictly convex, and differentiable. A reader more interested in the conceptual rather

than technical content may safely make this single set of assumptions throughout, perhaps even further assuming $\mathscr{V} = \mathscr{V}^* = \mathbb{R}^k$ for concreteness.

GENERALIZED EXPONENTIAL FAMILIES

Generalized exponential families (GEFs) are an extension of classical[1] exponential families as maximum entropy distributions for non-standard entropy functions. Below we take the reader through the very basics of the exponential family derivation, and show how one may generalize them in a natural way for other choices of entropy.

The goal of this section is two-fold. First, we wish to develop the theory of generalized exponential families beyond the excellent foundation of Grünwald and Dawid [4]. Second, we seek the machinery necessary to relate GEFs to divergences and prediction markets. In many cases, the relationships we uncover are surprising; as it happens, GEFs are the "answer" to many natural questions one may ask about the other constructs.

We will write $\phi^\top \theta$ to mean $o \mapsto \langle \phi(o), \theta \rangle$. We will also be assuming $\mathscr{P} \subseteq \mathscr{W}$ for the dual pair $(\mathscr{W}, \mathscr{W}^*)$ given by the duality $\langle p, f \rangle = \mathbb{E}_p[f]$. A major source of seemingly mysterious results is the simple observation that by linearity we may pass between inner products in $(\mathscr{W}, \mathscr{W}^*)$ to those in $(\mathscr{V}, \mathscr{V}^*)$. Specifically, we will constantly appeal to the fact that $\left\langle p, \phi^\top \theta \right\rangle = \left\langle \mathbb{E}_p[\phi], \theta \right\rangle$.

Before defining generalized exponential families, we recall some basic concepts from the literature on classical exponential families. In this case, we have a some σ-algebra Σ on $\mathscr{O}$ and base measure ν on $(\mathscr{O}, \Sigma)$, a statistic $\phi : \mathscr{O} \to \mathbb{R}^k$, and a parameter space Θ called the *natural parameters*. The *exponential family* $\{p_\theta\}_\Theta$ is defined by

$$p_\theta(o) = \exp\{\phi(o)^\top \theta - \Psi(\theta)\}, \tag{1}$$

where the *log partition function* $\Psi(\theta)$ is chosen to normalize p_θ, namely $\Psi(\theta) = \log \int_{\mathscr{O}} \exp\{\phi(o)^\top \theta\} \, d\nu(o)$. Typically the parameter space Θ is defined in terms of Ψ, by letting $\Theta \doteq \{\theta \in \mathbb{R}^k : \Psi(\theta) < \infty\}$. If the set Θ is open then the family $\{p_\theta\}_\Theta$ is called *regular*.

Many interesting characteristics of exponential families are known (see e.g. [3, 12]), but for our exploration two are especially relevant: exponential families have alternate parameterizations in terms of the mean of the statistic ϕ, and they can also be viewed as maximum entropy distributions under a mean constraint. Very briefly, one can check that, somewhat surprisingly, $\nabla\Psi(\theta) = \mathbb{E}_{p_\theta}[\phi]$; that is, the derivative of Ψ at θ is precisely the ϕ-mean for p_θ. This allows one to reparametrize the family by $\mathbb{E}_{p_\theta}[\phi]$. Moreover, one can derive this mean parametrization via a maximum entropy calculation. We briefly sketch this argument now.

The widely-used notion of entropy in probability theory is that of *Shannon entropy*, defined as $H(p) \doteq -\int_{\mathscr{O}} p(o) \log p(o) \, d\nu(o)$. The principle of maximum entropy states that given some data with empirical mean $\hat{\mu}$, to estimate the distribution from which the data was generated, one should compute the distribution p of maximum entropy $H(p)$ under the constraint $\mathbb{E}_p[\phi] = \hat{\mu}$. Formally, we wish to define p via the following optimization:

$$p \in \operatorname{argsup}\{H(p) : p \in \mathscr{P}, \mathbb{E}_p[\phi] = \hat{\mu}\}.$$

To solve this problem, we may turn to variational analysis and Lagrange multipliers, yielding the solution (1) [13, Thm 12.1.1], where θ is the vector of Lagrange multipliers from the calculation. However, one can also derive (1) via convex analysis. As it turns out, Ψ is a convex function, and happens to equal $(-H)^*$, the convex conjugate of (negative) Shannon entropy, applied to a particular point. To see this, first compute the entropy dual, $(-H)^*(q) = \log \int_{\mathscr{O}} \exp\{q(o)\} \, d\nu(o)$, and from there we can check that indeed $\Psi(\theta) = (-H)^*(\phi^\top \theta)$. More surprisingly, we can rederive (1) via the *derivative* of $(-H)^*$ at the same point:

$$\nabla_q (-H)^*(o) = \frac{\exp\{q(o)\}}{\int_{\mathscr{O}} \exp\{q(o')\} \, d\nu(o')} = \exp\{q(o) - (-H)^*(q)\}, \tag{2}$$

whence we have $\nabla_{\phi^\top \theta}(-H)^*(o) = \exp\{\phi(o)^\top \theta - \Psi(\theta)\} = p_\theta(o)$ which is precisely the definition of exponential families given in (1). As we will see below, this derivation is essentially the same as the maximum entropy calculation.

The main theme of this paper is that all of the observations and properties of classical exponential families mention above can be extended to families derived from other entropy functions other than Shannon entropy via convex duality. We are heavily influenced by Grünwald and Dawid [4], who introduced the idea of generalized exponential families,

[1] Throughout, we use the term "classical" to mean the standard definition in the literature.

along with many of the ideas we will explore. Our approach is different, however, relying much more on convex analysis. As a result, our setting will be slightly less general, but the extra regularity will go a long way.

Our results depend on the duality that is induced by convex conjugacy between $\mathscr{V}$ and $\mathscr{V}^*$ "going both ways", that is, we require that $F^{**} \doteq (F^*)^* = F$. The Fenchel-Moreau theorem (see, e.g., [14]) states that this is the case if and only if $F : \mathscr{V} \to \overline{\mathbb{R}}$ is convex, *lower semi-continuous* — i.e., $\liminf_{x \to x_0} F(x) \geq F(x_0)$ for all $x_0 \in \mathsf{dom}(F)$ — and *proper* — i.e., $F(x) > -\infty$. Thus, requiring that $F^{**} = F$ limits what we will consider an alternate entropy.[2]

Definition 1. *A function $F : \mathscr{P} \to \mathbb{R}$ is a* (generalized) entropy function *if it is convex, l.s.c., and proper.*

We also adopt the convention that $F(x) = \infty$ for $x \notin \mathsf{dom}(F)$, i.e., $F(p) = \infty$ for all $p \notin \mathscr{P}$. We can now define our generalized version of an exponential family. Of all the derivations above for classical families, the final derivation of $\nabla_{\phi^\top \theta}(-H)^*(o) = \exp\{\phi(o)^\top \theta - \Psi(\theta)\} = p_\theta(o)$ lends itself most to generalization. We employ this strategy, replacing $-H$ with our alternate entropy F.

Definition 2. *Let F be a given generalized entropy function, and let statistic $\phi : \mathscr{O} \to \mathbb{R}^k$ be given. Then a family of distributions $P_\Theta = \{p_\theta \in \partial F^*(\phi^\top \theta)\}_{\theta \in \Theta}$ is a F-*generalized exponential family (F-GEF) *with* cumulant $C(\theta) \doteq F^*(\phi^\top \theta)$*, where* $\Theta \doteq \mathsf{dom}(C)$.

To see that an F-GEF is indeed a collection of probability distributions (i.e., $P_\Theta \subseteq \mathscr{P}$) we first note that Lemma 1 applies to both F and F^* since $F = F^{**}$ and so $d \in \partial F_w \iff w \in \partial F^*_d$. This means for any $d \in \mathsf{dom}(F^*)$ we have $w \in \partial F^*_d \implies d \in \partial F_w$ for all $w \in \mathscr{W}$. Since $\partial F_w = \emptyset$ for $w \notin \mathsf{dom}(F)$, we must have $\partial F^* \subseteq \mathsf{dom}(F) = \mathscr{P}$. In particular then, $p_\theta \in \mathscr{P}$ for all $\theta \in \Theta$.

The following lemma is a well-known result about affine restrictions of convex functions on $\mathbb{R}^k$ [9, Thm E.2.1.1] which we have generalized to the infinite dimensional setting (see e.g. [10] for a proof).

Lemma 2. *If F is an entropy and $G(v) := \inf_{p \in \mathscr{P}}\{F(p) : \mathbb{E}_{o \sim p}[\phi(o)] = v\}$ then G is convex and $G^*(\theta) = F^*(\phi^\top \theta)$.*

Thus, Lemma 2 implies that the cumulant C of an F-GEF is the convex conjugate of $G(v) \doteq \inf\{F(p) : p \in \mathscr{P}, \mathbb{E}_p[\phi] = v\}$ which can be interpreted as a "maximum entropy" value under an appropriate change of sign. It is interesting to compare this result to our review of classical exponential families above; there we had $C = \Psi$, and as we saw, $\Psi(\theta) = (-H)^*(\phi^\top \theta)$, a result we of course recover by setting $F = -H$.

We now show a generalization of another classical result, that $\nabla \Psi(\theta) = \mathbb{E}_{p_\theta}[\phi]$. We first require a mild restriction on our class of F-GEFs that will also play a role in the later section on the bijection between GEFs and Bregman divergences. For that section we also need a a stronger restriction that is based on the concept of *Legendre type* from convex analysis (cf. [15, §26]) which is the following property of a set and function pair (X, F): (a) X is nonempty and open, (b) F is strictly convex and differentiable on X, and (c) $\lim_{x \to b} \|\nabla f(x)\| = \infty$ for $x \in X$ and $b \in \mathrm{bd}(X)$, the boundary of X.

Definition 3. *The F-GEF $\{p_\theta\}_{\theta \in \Theta}$ is* regular *if its cumulant C is l.s.c. and proper. A regular F-GEF is* Legendre *if $(\mathsf{dom}(C), C)$ of Legendre type.*

Note that even when we assume regularity our entropy might not be differentiable in any sense, so we state our generalisation of the classical exponential family result in terms of subgradients of the cumulant C.

Theorem 1. *A regular F-GEF $\{p_\theta\}$ with statistic ϕ and cumulant C satisfies $\mathbb{E}_{p_\theta}[\phi] \in \partial C_\theta$ for all θ.*

Proof. We first prove that if $g \doteq f \circ A^\top$ where $f : X \to \mathbb{R}$ is convex $A : X \to Y$ linear then $d \in \partial f(A^\top y) \implies Ad \in \partial g(y)$. Since $d \in \partial f(A^\top y)$ we have $\forall x\, f(x) \geq f(A^\top y) + \langle d, x - A^\top y \rangle$ which implies $\forall y'\, f(A^\top y') \geq f(A^\top y) + \langle d, A^\top y' - A^\top y \rangle$. However, this last inequality is equivalent to $g(y') \geq g(y) + \langle Ad, y' - y \rangle$ for all y', establishing the claim. Applying this to $f = F^*$, $g = C$, $A : p \mapsto \mathbb{E}_p[\phi]$ (and thus $A^\top = \phi^\top$) yields $p \in \partial F^*(\phi^\top \theta) \implies \mathbb{E}_p[\phi] \in \partial C(\theta)$, from which the result follows. □

[2] Note: our entropies are *convex*, unlike Shannon entropy which is concave; the reader familiar with the latter may need to mentally insert negations.

A BIJECTION BETWEEN GEFS AND BREGMAN DIVERGENCES

We introduce Bregman divergences and show how, due to convex duality, they are very closely related to F-GEFs.

Definition 4. *A* generalized Bregman divergence *on space X is a function $D_{G,dG} : X \times X \to \overline{\mathbb{R}}$ given by*

$$D_{G,dG}(x,x') = G(x) - G(x') - dG_{x'}(x-x'), \tag{3}$$

where $G : \mathsf{conv}(X) \to \overline{\mathbb{R}}$ is convex with $G(X) \subseteq \mathbb{R}$, and dG is a subgradient of G. If G is differentiable on X we simply write D_G and call D_G a Bregman divergence*.*

When G is continuously differentiable, the form (3) is simply called a *Bregman divergence*. Hence, Definition 4 is merely a natural extension to the nondifferentiable case, and has been studied in machine learning [16].

We now apply the above definitions and machinery to show a bijection between generalized exponential families and generalized Bregman divergences. This investigation is inspired by Banerjee et al. [1], which shows a similar result for (classical) exponential families. In fact, our result will in some sense generalize theirs to other entropies besides Shannon entropy. We first need to define some restrictions on the class of Bregman divergences.

Definition 5. *A generalized Bregman divergence $D_{G,dG}$ for a convex $G : \mathscr{V} \to \overline{\mathbb{R}}$ is F-*regular *for a convex $F : \mathscr{P} \to \overline{\mathbb{R}}$ if G is proper and l.s.c., and there exists some statistic $\phi : \mathscr{O} \to \mathbb{R}^k$ such that $G(v) = \inf_{p \in \mathscr{P}}\{F(p) : \mathbb{E}_p[\phi] = v\}$. We will also say G itself is F-regular with statistic ϕ. An F-regular Bregman divergence $D_{G,dG}$ is F-*Legendre *if $(\mathrm{int}(\mathsf{dom}(G)), G)$ is of Legendre type.*

The bijection we present which generalizes that of [1] ties Bregman divergences to *equivalence classes* of GEFs:

Definition 6. *The* cumulant class *of F-GEFs with cumulant C is the set of F-GEFs whose cumulant is C.*

Of course, two F-GEFs with the same cumulant may be very different from one another, but in some sense the disparity is only with regard to $\mathscr{O}$, not Θ and $\mathscr{R}$. Note that our definition of regularity, Definition 3, really is a property of the cumulant, and hence naturally applies to cumulant classes; we will say a cumulant class is regular to mean its cumulant satisfies the same. We can now state our generalization of Banerjee et al.'s bijection.

Theorem 2. *Fix entropy function F. The set of F-regular Bregman divergences is in bijection with the set of regular cumulant classes of F-GEFs. Furthermore, a bijection also holds between F-Legendre Bregman divergences and cumulant classes of Legendre F-GEFs.*

Proof. We will show that each F-regular Bregman divergence $D_{G,dG}$ yields the cumulant class of $C(\theta) \doteq G^*(\theta)$, and as the convex conjugation operator is invertible, thereby establishing our bijection.

Let F-regular Bregman divergence $D_{G,dG}$ be given, with statistic ϕ. Then simply take the cumulant class of the F-GEF $P_\Theta = \{p_\theta \in \partial F^*(\phi^\top \theta)\}_{\theta \in \Theta}$. By Lemma 2 and Definition 5, we have that the cumulant of P_Θ is G^*. Finally, by regularity of $D_{G,dG}$ and the Fenchel-Moreau theorem, we have $(G^*)^{**} = (G^{**})^* = G^*$, so P_Θ is regular. Now, given any cumulant class of regular F-GEFs with cumulant C and statistic ϕ, we take the Bregman divergence with convex function $G \doteq C^*$. Again by regularity and Fenchel-Moreau, we have that G is l.s.c. and proper. Also, since $G^* = C^{**} = C$, we have $G^*(\theta) = F^*(\phi^\top \theta)$, so by Lemma 2 we have $G(v) = \inf_{p : \mathbb{E}_p[\phi] = v} F(p)$, meaning G is F-regular.

A classic result of convex analysis states that (X,F) is of Legendre type if and only if (X^*,F^*) is [15, Thm 26.5]. Applying this to $(\mathsf{dom}(C),C) = (\mathsf{dom}(G^*),G^*)$ in the argument above shows the bijection is preserved when restricted to F-Legendre divergences and Legendre F-GEFs. □

To see that the result of Banerjee et al. [1] is indeed a special case of ours we need only check that, in the case of $F = -H$, the classes of F-Legendre Bregman divergences and Legendre F-GEFs coincide with their definitions of regular exponential families and regular Bregman divergences, respectively. Their Lemma 1 shows that any (classically) regular exponential family has a cumulant of Legendre type, and is thus representative of a Legendre $(-H)$-GEF cumulant class. Conversely, a Legendre $(-H)$-GEF cumulant C will have open $\Theta = \mathsf{dom}(C)$ and every exponential family in its class is thus regular in the classical sense.

Their regularity condition on divergences D_G is, in our notation, that $G = C^*$ for some strictly convex C which (after applying a result due to Devinatz) is of the form $C(\theta) = \log \int_{\mathbb{R}^k} \exp\{\langle x, \theta \rangle\} d\mu(x)$ for some unique, bounded, non-negative measure μ. But since a $(-H)$-Legendre Bregman divergence guarantees a statistic ϕ such that $C(\theta) = (-H)^*(\phi^\top \theta)$ and $(\mathsf{dom}(C),C)$ is of Legendre type we see that these two classes coincide.

It has been noted that one can express the relative entropy between two members of a classical exponential family as a divergence between their corresponding parameters [5, 17, 7]. Specifically, one can write

$$\mathrm{KL}(p_\theta \| p_{\theta'}) = D_\Psi(\theta', \theta). \tag{4}$$

We now generalize this result, showing that the *generalized relative entropy*, given by D_F, can be written analogously.

Theorem 3. *Let F-GEF $\{p_\theta\}$ be given with cumulant C. Then there exist subradients dF and dC such that for all $\theta, \theta' \in \Theta$, $D_{F,dF}(p_{\theta'}, p_\theta) = D_{C,dC}(\theta, \theta')$.*

Proof. By Lemma 1, and the fact that $p_\theta \in \partial F^*(\phi^\top \theta)$, we have $F(p_\theta) = \langle p_\theta, dF_{p_\theta} \rangle - F^*(\phi^\top \theta)$ for all $\theta \in \Theta$ and any choice of subgradient dF. Selecting $dF_{p_\theta} = \phi^\top \theta$ and applying this to the definition of $D_{F,dF}$ gives $D_{F,dF}(p_{\theta'}, p_\theta) = \langle p_{\theta'}, \phi^\top \theta' \rangle - F^*(\phi^\top \theta') - \langle p_\theta, \phi^\top \theta \rangle + F^*(\phi^\top \theta) - \langle p_{\theta'} - p_\theta, \phi^\top \theta \rangle$. By Lemma 2, we see $D_{F,dF}(p_{\theta'}, p_\theta) = C(\theta) - C(\theta') + \langle \mathbb{E}_{p_{\theta'}}[\phi], \theta' - \theta \rangle$. Finally, by Theorem 1 we may select $dC_\theta = \mathbb{E}_{p_\theta}[\phi]$, completing the proof. □

Theorem 3 gives life to our bijection in Theorem 2. We see that not only are F-GEFs in bijection with divergences, but these divergences exactly capture the geometry of the GEF, and succinctly so, in terms of their parameters.

PREDICTION MARKETS, PRICES, AND GEFS

In finance, securities and specifically futures markets are often thought to in some sense reveal the beliefs of the traders. A *prediction market* makes this connection more explicit: traders buy and sell contracts in specific outcomes, which pay out according to what happens on some future date. The simplest such market is the *complete market*, which offers a contract for each of a set $\mathscr{O}$ of mutually exclusive outcomes, with the contract for $o \in \mathscr{O}$ paying out \$1 if outcome o occurs and \$0 otherwise. One can see that the prices in such a market should form a probability distribution over $\mathscr{O}$ (otherwise a trader could arbitrage), intuitively reflecting the consensus belief of the market. Similarly, we may extend this idea to an *incomplete market*, where there are merely k contracts offered (intuitively $k \ll |\mathscr{O}|$), and the payoff of contract i given outcome $o \in \mathscr{O}$ is determined by a *payoff function* $\phi : \mathscr{O} \to \mathbb{R}^k$, namely $\phi(o)_i$.

It is common to model, and implement, a prediction market via an automated market maker, a central entity from which traders make all purchases (cf. [2]). Formally, we will define our market maker in terms of a potential function. Given outcome space $\mathscr{O}$, price space $\mathscr{R}$, share space $\mathscr{Q}$, payoff function $\phi : \mathscr{O} \to \mathscr{R}$, and set of probability measures $\mathscr{P} \subseteq \Delta(\mathscr{O})$, a *prediction market potential* is a function $\mathfrak{P} : \mathscr{Q} \times \mathscr{O} \to \overline{\mathbb{R}}$ defined by

$$\mathfrak{P}(q, o) \doteq \langle q, \phi(o) \rangle - C(q), \tag{5}$$

where $C = G^*$ for some convex $G : \mathsf{conv}(\mathscr{R}) \to \overline{\mathbb{R}}$ with $G(\mathscr{R}) \subseteq \mathbb{R}$ and $\partial G(\mathscr{R}) = \mathscr{Q}$. This potential captures the "one-round" incentives of a single trader; the usual dynamic mechanism of buying and selling can be recovered by considering differences in the potential. Specifically, according to a standard framework [2], a trader who purchases a bundle $r \in \mathbb{R}^k$ (i.e., r_i shares of each security i), when the total number of shares sold so far in the market is $q \in \mathbb{R}^k$, must pay the market maker $C(q+r) - C(q)$, but will receive payoff $\phi(o) \cdot r$ upon outcome o being revealed. The net payoff of this trade then is precisely $\mathfrak{P}(q+r, o) - \mathfrak{P}(q, o)$; iterating this process for a sequence of trades and tracking the total shares sold gives the usual model.

We now ask a natural question: given a payoff function $\phi : \mathscr{O} \to \mathscr{R}$, when can one design an incomplete prediction market for ϕ via reduction to a complete market on $\mathscr{O}$? Specifically, when can we map trades in $\mathscr{Q}$ to trades in $\mathbb{R}^{\mathscr{O}}$ in such a way as to replicate the payoffs in the original market? As we will see, GEFs play a crucial role in the answer.

Theorem 4. *Let $C : \mathscr{V}^* \to \overline{\mathbb{R}}$ and $B : \mathscr{W}^* \to \overline{\mathbb{R}}$ be given convex, l.s.c., and proper functions, and define $\mathfrak{P}^{\mathscr{P}}(w^*, o) \doteq w^*(o) - B(w^*)$ and $\mathfrak{P}(v^*, o) \doteq \langle \phi(o), v^* \rangle - C(v^*)$. Then there is some $\hat{q} : \mathscr{V}^* \to \mathscr{W}^*$ such that $\mathfrak{P}(\cdot, \cdot) \equiv \mathfrak{P}^{\mathscr{P}}(\hat{q}(\cdot), \cdot)$ if and only if C^* is B^*-regular for statistic ϕ.*

Proof. We observe that the condition $\mathfrak{P}(\cdot, \cdot) \equiv \mathfrak{P}^{\mathscr{P}}(\hat{q}(\cdot), \cdot)$ breaks into two: (a) $\forall o \in \mathscr{O}\ \hat{q}(v^*)(o) = \langle \phi(o), v^* \rangle$, and (b) $C(v^*) = B(\hat{q}(v^*))$. Condition (a) is equivalent to $\hat{q}(v^*) = \phi^\top v^*$ by definition. Using this, we can reduce (b) to $C(v^*) = B(\phi^\top v^*)$. Now letting $\theta \doteq v^*$, we can apply Lemma 2, and as has been argued previously (since the Fenchel–Moreau theorem holds under our assumptions), we see that (b) is equivalent to C^* being B^*-regular. □

While the proof of Theorem 4 does not mention GEFs explicitly, they surface when one considers the induced *price space* for the complete market $\mathfrak{P}^{\mathscr{P}}$. That is, rather than the whole of ∂B, we consider only the possible prices attained

by varying v^*, namely $\bigcup_{v^* \in \mathcal{V}^*} \partial B(\hat{q}(v^*)) = \bigcup_{\theta \in \mathcal{V}^*} \partial F^*(\phi^\top \theta)$, where $F = B^*$. We can now see that, if we pick a price p_θ corresponding to each θ, the prices form an F-GEF! In particular, if F^* is differentiable, the prices will be unique (often a desirable property for a prediction market mechanism), and will correspond to a unique F-GEF.

This latter observation deserves attention. When an incomplete prediction market $\mathfrak{P}$ is induced from a complete one $\mathfrak{P}^{\mathscr{P}}$, the prices are constrained to be a generalized exponential family, where the statistic ϕ is simply the payoff function of market $\mathfrak{P}$. Hence, traders with beliefs about the mean of the payoff function ϕ place bets in the market in exactly the same way as if they were betting directly on the outcomes, but with the prices constrained to a particular GEF. In particular, when $\mathfrak{P}^{\mathscr{P}}$ is LMSR (the complete market with $B = (-H)^*$), the most widely-used automated prediction market mechanism, agents are essentially trading on the mean parameters of classical exponential families! Furthermore, any prediction market with $C = G^*$ for a $(-H)$-regular G, or a G which is regular in the sense of Banerjee et al. [1], can be expressed as an instance of LMSR in this way.

Finally, we turn our discussion back to the principle of maximum entropy. By the above one could interpret an incomplete market as taking the current market price, which intuitively corresponds to the mean of the payoff function ϕ under their consensus belief, and attempting to reconstruct this consensus belief via the maximum entropy principle. Indeed, the costs and payouts of the incomplete market, viewed in terms of the corresponding complete market, are exactly consistent with this interpretation.

DISCUSSION AND FUTURE WORK

The above exploration of generalized exponential families certainly opens more doors than it closes. It is in particular quite natural to ask, for other qualities of classical exponential families, does a generalization of this quality hold for GEFs too? Such extensions would immediately apply to the other concepts discussed here. While a number of specific questions and relationships are addressed in [10] we are still pursuing several other conjectures including connections between "generalized Bayesian updating" of GEFs and Vovk's aggregating algorithm [18] that may perhaps shed light on alternative notions of mixability [19] — a key concept in the sequential aggregation of predictions.

ACKNOWLEDGMENTS

Mark Reid is partly funded by an Australian Research Council DECRA (DE130101605). We thank Matus Telgarsky and Bob Williamson for helpful technical discussions. Some aspects of our study of GEFs are inspired by unpublished work of Sébastien Lahaie.

REFERENCES

1. A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh, *The Journal of Machine Learning Research* **6**, pp. 1705–1749, 2005.
2. J. Abernethy, Y. Chen, and J. Wortman Vaughan, "An optimization-based framework for automated market-making," in *Proceedings of the 12th ACM conference on Electronic commerce*, 2011.
3. O. E. Barndorff-Nielsen, *Information and exponential families: in statistical theory*, Wiley, Chichester, 1978.
4. P. Grünwald, and A. Dawid, *The Annals of Statistics* **32**, pp. 1367—1433, 2004.
5. S. Amari, *Differential-geometrical methods in statistics*, Springer-Verlag, Berlin; New York, 1985.
6. K. S. Azoury, and M. K. Warmuth, *Machine Learning* **43**, pp. 211–246, 2001.
7. F. Nielsen, and R. Nock, "Entropies and cross-entropies of exponential families," in *Intl. Conf. on Image Proc. (ICIP)*, 2010.
8. R. Rockafellar, *Convex analysis*, Princeton University Press, 1997.
9. J.-B. Hiriart-Urruty, and C. Lemaréchal, *Fundamentals of Convex Analysis*, Springer, Berlin, 2001.
10. R. M. Frongillo, *Eliciting Private Information from Selfish Agents*, Ph.D. thesis, University of California, Berkeley, 2013.
11. C. D. Aliprantis, and K. C. Border, *Infinite Dimensional Analysis: A Hitchhiker's Guide*, Springer, 2007.
12. M. J. Wainwright, and M. I. Jordan, *Foundations and Trends® in Machine Learning* **1**, 2008.
13. T. Cover, J. Thomas, J. Wiley, et al., *Elements of information theory*, vol. 6, Wiley Online Library, 1991.
14. C. Zălinescu, *Convex Analysis in General Vector Spaces*, World Scientific Publishing Company, Incorporated, 2002.
15. R. Rockafellar, *Convex analysis*, vol. 28 of *Princeton Mathematics Series*, Princeton University Press, 1997.
16. R. Iyer, and J. Bilmes, *Uncertainity in Artificial Intelligence*, 2013.
17. J.-D. Boissonnat, F. Nielsen, and R. Nock, *CoRR* **abs/0709.2196**, 2007.
18. V. G. Vovk, "A game of prediction with expert advice," in *Proceedings of the eighth annual conference on Computational learning theory*, COLT '95, ACM, New York, NY, USA, 1995.
19. T. van Erven, M. D. Reid, and R. C. Williamson, *Journal of Machine Learning Research* **13**, pp. 1639–1663, 2012.