Convex Foundations for Generalized MaxEnt Models

Rafael Frongillo[1] Mark D. Reid[2]

[1]Microsoft Research, New York

[2]The Australian National University & NICTA

December 16th, 2013

Convex Foundations for Generalized MaxEnt Models

Rafael Frongillo[1] Mark D. Reid[2]

[1]Microsoft Research, New York

[2]The Australian National University & NICTA

December 16^{th}, 2013

Eliciting Private Information from Selfish Agents
(Ph.D. – U.C. Berkeley, 2013)

Motivation

This work came about when Rafael and I tried to understand this:

Theorem 6 (Banerjee *et al.*, 2006)

There is a bijection between regular exponential families and regular Bregman divergences.

The bijection was based on the convex duality between the *cumulant* of the EF and the *generator* of the BD.

Motivation

This work came about when Rafael and I tried to understand this:

Theorem 6 (Banerjee *et al.*, 2006)

There is a bijection between regular exponential families and regular Bregman divergences.

The bijection was based on the convex duality between the *cumulant* of the EF and the *generator* of the BD.

Our idea:

- We are comfortable with Bregman divergences (BDs) and convexity
- ...but had little idea about exponential families (EFs)

Why not use the above result to understand EFs via BDs?

Motivation

This work came about when Rafael and I tried to understand this:

Theorem 6 (Banerjee *et al.*, 2006)

There is a bijection between regular exponential families and regular Bregman divergences.

The bijection was based on the convex duality between the *cumulant* of the EF and the *generator* of the BD.

Our idea:

- We are comfortable with Bregman divergences (BDs) and convexity
- ...but had little idea about exponential families (EFs)

Why not use the above result to understand EFs via BDs?

The rabbit hole: What does "regular" mean here?

Preliminaries

Convexity:

- *Dual pair* $(\mathcal{V}, \mathcal{V}^*)$ with bilinear $\langle \cdot, \cdot \rangle : \mathcal{V} \times \mathcal{V}^* \to \mathbb{R}$
- The *convex conjugate* of $G : \mathcal{V} \to \overline{\mathbb{R}}$ is $G^* : \mathcal{V}^* \to \overline{\mathbb{R}}$ defined by $G^*(v^*) := \sup_{v \in \mathcal{V}} \langle v, v^* \rangle - G(v)$
- **Fenchel-Moreau**: For $G : \Omega \to \overline{\mathbb{R}}$ with Ω Hausdorff & locally convex $G^{**} = G \iff G \equiv \pm\infty$ or G convex, l.s.c. & proper

Preliminaries

Convexity:

- *Dual pair* $(\mathcal{V}, \mathcal{V}^*)$ with bilinear $\langle \cdot, \cdot \rangle : \mathcal{V} \times \mathcal{V}^* \to \mathbb{R}$
- The *convex conjugate* of $G : \mathcal{V} \to \overline{\mathbb{R}}$ is $G^* : \mathcal{V}^* \to \overline{\mathbb{R}}$ defined by $G^*(v^*) := \sup_{v \in \mathcal{V}} \langle v, v^* \rangle - G(v)$
- **Fenchel-Moreau**: For $G : \Omega \to \overline{\mathbb{R}}$ with Ω Hausdorff & locally convex $G^{**} = G \iff G \equiv \pm\infty$ or G convex, l.s.c. & proper

Uncertainty:

- *Distribution* $p \in \Delta_\Omega$ over (possibly uncountable*) outcomes in Ω
  (*i.e.*, densities with measure space (Ω, Σ) and reference measure λ)
- *Random variable* or *statistic* $\phi : \Omega \to \mathcal{V} \subseteq \mathbb{R}^d$
- These are a dual pair $(\mathcal{W}, \mathcal{W}^*)$ with $\langle p, \phi \rangle = \mathbb{E}_{\omega \sim p}[\phi(\omega)]$

*This is a depature from a similar treatement for finite outcome spaces by Sears (2010).

Preliminaries

Convexity:

- *Dual pair* $(\mathcal{V}, \mathcal{V}^*)$ with bilinear $\langle \cdot, \cdot \rangle : \mathcal{V} \times \mathcal{V}^* \to \mathbb{R}$
- The *convex conjugate* of $G : \mathcal{V} \to \overline{\mathbb{R}}$ is $G^* : \mathcal{V}^* \to \overline{\mathbb{R}}$ defined by $G^*(v^*) := \sup_{v \in \mathcal{V}} \langle v, v^* \rangle - G(v)$
- **Fenchel-Moreau**: For $G : \Omega \to \overline{\mathbb{R}}$ with Ω Hausdorff & locally convex $G^{**} = G \iff G \equiv \pm\infty$ or G convex, l.s.c. & proper

Uncertainty:

- *Distribution* $p \in \Delta_\Omega$ over (possibly uncountable*) outcomes in Ω
  (*i.e.*, densities with measure space (Ω, Σ) and reference measure λ)
- *Random variable* or *statistic* $\phi : \Omega \to \mathcal{V} \subseteq \mathbb{R}^d$
- These are a dual pair $(\mathcal{W}, \mathcal{W}^*)$ with $\langle p, \phi \rangle = \mathbb{E}_{\omega \sim p}[\phi(\omega)]$

Connecting Two Dual Pairs:

$$\langle \mathbb{E}_p[\phi], \theta \rangle_{\mathcal{V}} = \langle \langle p, \phi \rangle_{\mathcal{W}}, \theta \rangle_{\mathcal{V}} = \langle p, \langle \phi, \theta \rangle_{\mathcal{V}} \rangle_{\mathcal{W}} = \mathbb{E}_p\left[\phi^\top \theta\right]$$

* This is a depature from a similar treatement for finite outcome spaces by Sears (2010).

A Quick Review

Exponential Family

For *statistic* $\phi : \Omega \to \mathbb{R}^d$ an exponential family (w.r.t. some measure λ) is a set $\mathcal{F} = \{p_\theta : \theta \in \Theta\}$ of densities of the form

$$p_\theta(\omega) := \exp\left(\langle \phi(\omega), \theta \rangle - C(\theta)\right)$$

with finite *cumulant* $C(\theta) := \log \int_\Omega p_\theta(\omega)\, d\lambda(\omega)$. The parameters $\theta \in \Theta$ are natural parameters. The family $\mathcal{F}$ is regular if Θ is an open set.

A Quick Review

Exponential Family

For *statistic* $\phi : \Omega \to \mathbb{R}^d$ an exponential family (w.r.t. some measure λ) is a set $\mathcal{F} = \{p_\theta : \theta \in \Theta\}$ of densities of the form

$$p_\theta(\omega) := \exp\left(\langle \phi(\omega), \theta \rangle - C(\theta)\right)$$

with finite *cumulant* $C(\theta) := \log \int_\Omega p_\theta(\omega)\, d\lambda(\omega)$. The parameters $\theta \in \Theta$ are natural parameters. The family $\mathcal{F}$ is regular if Θ is an open set.

Bregman Divergence

A (generalised) Bregman divergence on X is the function

$$D_{F,dF}(x, x) = F(x) - F(x') - dF_{x'}(x - x')$$

where its *generator* $F : X \to \mathbb{R}$ is convex and $dF \in \partial F$ a *subgradient* of F.

A Quick Review

Exponential Family

For *statistic* $\phi : \Omega \to \mathbb{R}^d$ an exponential family (w.r.t. some measure λ) is a set $\mathcal{F} = \{p_\theta : \theta \in \Theta\}$ of densities of the form

$$p_\theta(\omega) := \exp\left(\langle \phi(\omega), \theta \rangle - C(\theta)\right)$$

with finite *cumulant* $C(\theta) := \log \int_\Omega p_\theta(\omega)\, d\lambda(\omega)$. The parameters $\theta \in \Theta$ are natural parameters. The family $\mathcal{F}$ is regular if Θ is an open set.

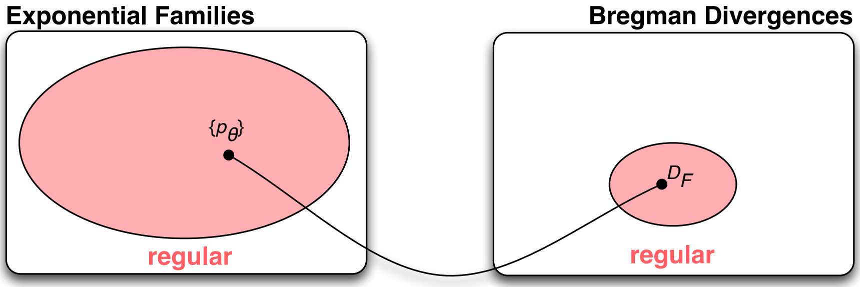
Bregman Divergence

A (generalised) Bregman divergence on X is the function

$$D_F(x, x) = F(x) - F(x') - \left\langle \nabla F(x'), x - x' \right\rangle$$

where its *generator* $F : X \to \mathbb{R}$ is convex and differentiable.

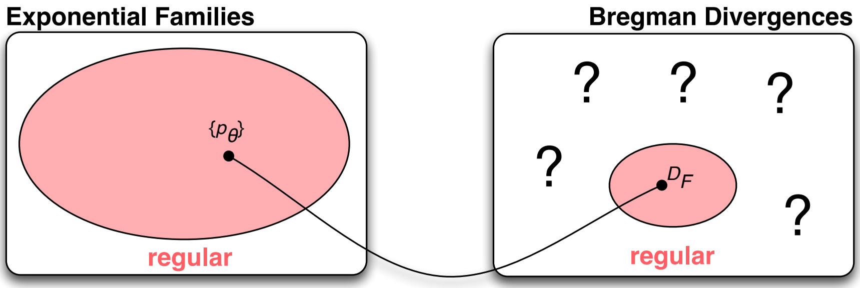
The Mystery



- Regularity is not such a strong constraint on EFs ($\Longleftrightarrow \Theta$ is open)
- Regularity for a BD D_F requires its generator F to be strictly convex and satisfy $F(x) = \log G^*(x)$ where

$$G(\theta) = \log \int_X \exp(\langle x, \theta \rangle)\, d\nu(x)$$

The Mystery



- Regularity is not such a strong constraint on EFs ($\Longleftarrow \Theta$ is open)
- Regularity for a BD D_F requires its generator F to be strictly convex and satisfy $F(x) = \log G^*(x)$ where

$$G(\theta) = \log \int_X \exp(\langle x, \theta \rangle)\, d\nu(x)$$

So what do all the other Bregman divergences correspond to?

A Clue: Exponential Families via Maximum Entropy

Maximum Entropy

Define the Shannon entropy as the concave function

$$H(p) = \begin{cases} -\int_\Omega p(\omega) \log p(\omega)\, d\lambda(\omega) & \text{for } p \in \Delta_\Omega \\ -\infty & \text{otherwise} \end{cases}$$

For a given mean value $r \in \mathbb{R}^d$ define the maximum entropy solution

$$p_r = \arg\sup\{H(p) : \mathbb{E}_p[\phi] = r\}$$

A Clue: Exponential Families via Maximum Entropy

Maximum Entropy

Define the Shannon entropy as the concave function

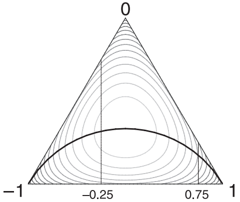
$$H(p) = \begin{cases} -\int_\Omega p(\omega)\log p(\omega)\,d\lambda(\omega) & \text{for } p \in \Delta_\Omega \\ -\infty & \text{otherwise} \end{cases}$$

For a given mean value $r \in \mathbb{R}^d$ define the maximum entropy solution

$$p_r = \arg\sup\{H(p) : \mathbb{E}_p[\phi] = r\}$$

Example [Grünwald & Dawid (2004)]:
$\Omega = \{-1, 0, 1\}$ with statistic $\phi(\omega) = \omega$.

- Each constraint $\mathbb{E}_p[\phi] = r \in [-1, 1]$ yields vertical slice of Δ_Ω.
- Choose p maximising H over slice.



Exponential Families via Convex Duality

The dual of the maximum entropy problem gives an alternative definition:

Exponential Families via Convexity

For statistic $\phi : \Omega \to \mathbb{R}^d$ each p_θ in the exp. family for ϕ can be written as

$$p_\theta = \nabla(-H)^*(\phi^\top \theta)$$

and $C(\theta) = (-H^*)(\phi^\top \theta)$ where $\phi^\top \theta \in \mathcal{W}^*$ denotes $\omega \mapsto \langle \phi(\omega), \theta \rangle$.

Exponential Families via Convex Duality

The dual of the maximum entropy problem gives an alternative definition:

Exponential Families via Convexity

For statistic $\phi : \Omega \to \mathbb{R}^d$ each p_θ in the exp. family for ϕ can be written as

$$p_\theta = \nabla(-H)^*(\phi^\top \theta)$$

and $C(\theta) = (-H^*)(\phi^\top \theta)$ where $\phi^\top \theta \in \mathcal{W}^*$ denotes $\omega \mapsto \langle \phi(\omega), \theta \rangle$.

Straight-forward to check that for any $q : \Omega \to \mathbb{R}$:

$$\nabla(-H^*)(q)_\omega = \frac{\exp(q(\omega))}{\int_\Omega \exp\left(q(o)\right)\, d\lambda(o)} \in \Delta_\Omega$$

Exponential Families via Convex Duality

The dual of the maximum entropy problem gives an alternative definition:

Exponential Families via Convexity

For statistic $\phi : \Omega \to \mathbb{R}^d$ each p_θ in the exp. family for ϕ can be written as

$$p_\theta = \nabla(-H)^*(\phi^\top\theta)$$

and $C(\theta) = (-H^*)(\phi^\top\theta)$ where $\phi^\top\theta \in \mathcal{W}^*$ denotes $\omega \mapsto \langle\phi(\omega), \theta\rangle$.

Straight-forward to check that for any $q : \Omega \to \mathbb{R}$:

$$\nabla(-H^*)(q)_\omega = \frac{\exp(q(\omega))}{\int_\Omega \exp\left(q(o)\right)\, d\lambda(o)} \in \Delta_\Omega$$

But! the Shannon entropy H is not so special: p_θ are distributions because

$\partial F^*(q) \subset \mathrm{dom}(F) \subseteq \Delta_\Omega$ for any convex, l.s.c. $F : \Delta_\Omega \to \mathbb{R}$

We will **define** an *entropy* to be a convex, l.s.c. function $F : \Delta_\Omega \to \mathbb{R}$.

Generalised Exponential Families

Generalised Exponential Family (GEF)

Let $F : \Delta_\Omega \to \mathbb{R}$ be an entropy and $\phi : \Omega \to \mathcal{V} \subseteq \mathbb{R}^d$ be a statistic. Then

$$\mathcal{F} := \{p_\theta \in \partial F^*(\phi^\top \theta)\}_{\theta \in \Theta} \subseteq \Delta_\Omega$$

is an *F-GEF* with cumulant $C(\theta) := F^*(\phi^\top \theta)$ and $\Theta := \mathrm{dom}(C)$.

Generalised Exponential Families

Generalised Exponential Family (GEF)

Let $F : \Delta_\Omega \to \mathbb{R}$ be an entropy and $\phi : \Omega \to \mathcal{V} \subseteq \mathbb{R}^d$ be a statistic. Then

$$\mathcal{F} := \{p_\theta \in \partial F^*(\phi^\top \theta)\}_{\theta \in \Theta} \subseteq \Delta_\Omega$$

is an F-GEF with cumulant $C(\theta) := F^*(\phi^\top \theta)$ and $\Theta := \mathrm{dom}(C)$.

Several properties of classical exponential families are easily recovered

Theorem 1: Subgradients Contain Means

A regular F-GEF with statistic ϕ has cumulant C s.t. $\mathbb{E}_{p_\theta}[\phi] \in \partial C(\theta)$

Generalised Exponential Families

Generalised Exponential Family (GEF)

Let $F : \Delta_\Omega \to \mathbb{R}$ be an entropy and $\phi : \Omega \to \mathcal{V} \subseteq \mathbb{R}^d$ be a statistic. Then

$$\mathcal{F} := \{p_\theta \in \partial F^*(\phi^\top \theta)\}_{\theta \in \Theta} \subseteq \Delta_\Omega$$

is an F-GEF with cumulant $C(\theta) := F^*(\phi^\top \theta)$ and $\Theta := \mathrm{dom}(C)$.

Several properties of classical exponential families are easily recovered

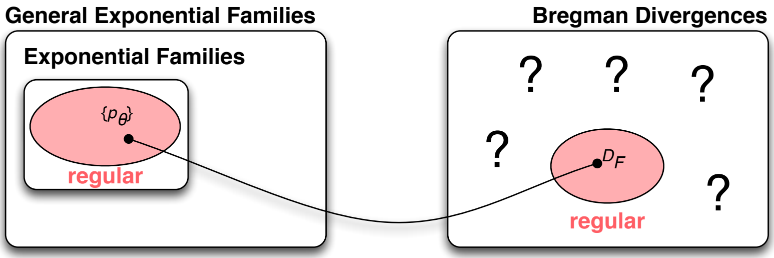
Theorem 3: Divergence Duality

For F-GEF $\mathcal{F}$ with statistic ϕ and cumulant C, for each $p_\theta, p_{\theta'} \in \mathcal{F}$

$$D_F(p_\theta, p_{\theta'}) = D_C(\theta', \theta)$$

In the special case of classical EFs $F = -H$ and $D_F(p_\theta, p_{\theta'}) = KL(p_\theta \| p_{\theta'})$.

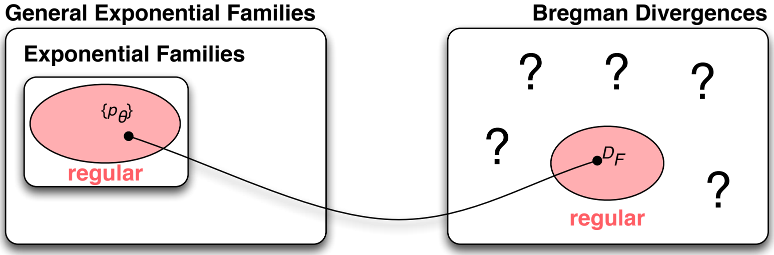
The Bigger Picture



Theorem 2: Generalised Bijection

For each entropy F, the set of F-regular Bregman divergences is in bijection with the set of regular F-GEFs.

The Bigger Picture

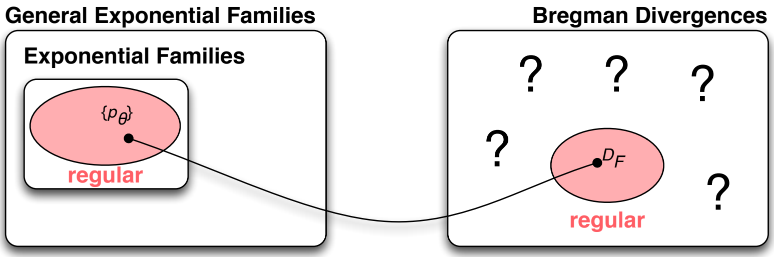


Theorem 2: Generalised Bijection

For each entropy F, the set of F-regular Bregman divergences is in bijection with the set of regular F-GEFs.

Redefining regularity:

- D_G is F-regular if there is a statistic ϕ so that G is "F-MaxEnt": $G(r) = \inf_p\{F(p) : \mathbb{E}_p[\phi] = r\}$
- An F-GEF is regular if its cumulant C is itself an entropy



Theorem 2: Generalised Bijection (Legendre Refinement)

For each entropy F, the set of F-regular (Legendre) Bregman divergences is in bijection with the set of regular (Legendre) F-GEFs.

Banerjee et al.'s bijection is recovered as a special case when $F = -H$.

Prediction Markets

Traders buy and sell contracts with payoff contingent on future outcomes (e.g., Presidential elections, horse races, box office takings) and the prices they are willing to trade at reveal their beliefs about the outcomes.

- In a k-contract market with mutually exclusive outcomes Ω, the payoff of contract $i \in \{1, \ldots, k\}$ on outcome $\omega \in \Omega$ is $\phi_i(\omega)$.
- For the bundle $r \in \mathbb{R}^k$ of contracts the payoff is $\langle r, \phi(\omega) \rangle$
- A market is complete if $k \geq |\Omega|$ and ϕ_i linearly independent

---

* Path independence, no arbitrage, information incorporation, expressiveness, instantaneous prices (Abernethy et al. (2012))

Prediction Markets

Traders buy and sell contracts with payoff contingent on future outcomes (e.g., Presidential elections, horse races, box office takings) and the prices they are willing to trade at reveal their beliefs about the outcomes.

- In a k-contract market with mutually exclusive outcomes Ω, the payoff of contract $i \in \{1, \ldots, k\}$ on outcome $\omega \in \Omega$ is $\phi_i(\omega)$.
- For the bundle $r \in \mathbb{R}^k$ of contracts the payoff is $\langle r, \phi(\omega) \rangle$
- A market is complete if $k \geq |\Omega|$ and ϕ_i linearly independent

An automated market maker (AMM) interacts with traders and adaptively prices contract bundles to aggregate the market's belief
Under some natural assumptions* AMMs **must** price bundle r as

$$\mathrm{Cost}(r) = C(q + r) - C(q)$$

where $C : \mathbb{R}^k \to \mathbb{R}$ is a convex cost function and q is net contract position

*Path independence, no arbitrage, information incorporation, expressiveness, instantaneous prices (Abernethy et al. (2012))

Prediction Market Pricing Mechanisms

Thus, the net payoff for a trader to purchase bundle r in net position q is

$$\underbrace{\langle r, \phi(\omega)\rangle}_{\text{Payoff for } r} - \underbrace{C(q+r) - C(q)}_{\text{Cost to buy } r} = V^\phi_\omega(q+r) - V^\phi_\omega(q)$$

where $V^\phi_\omega(q) = \langle q, \phi(\omega)\rangle - C(q)$ is the trader "value potential".

Prediction Market Pricing Mechanisms

Thus, the net payoff for a trader to purchase bundle r in net position q is

$$\underbrace{\langle r, \phi(\omega)\rangle}_{\text{Payoff for } r} - \underbrace{C(q+r) - C(q)}_{\text{Cost to buy } r} = V^{\phi}_{\omega}(q+r) - V^{\phi}_{\omega}(q)$$

where $V^{\phi}_{\omega}(q) = \langle q, \phi(\omega)\rangle - C(q)$ is the trader "value potential".

How does the potential V^{ϕ}_{ω} for an incomplete market with cost function C relate to V_{ω} for the underlying complete market with cost function B?

Theorem 4 : Complete and Incomplete Markets

There is an *bundle mapping* $f : \mathbb{R}^k \to \mathbb{R}^{\Omega}$ s.t. $V_{\omega}(f(q)) = V^{\phi}_{\omega}(q) \quad \forall \omega, q$
$\iff$ C^* is B^*-regular for ϕ — i.e, $C^*(r) = \inf_p\{B^*(p) : \mathbb{E}_p[\phi] = r\}$

Prediction Market Pricing Mechanisms

Thus, the net payoff for a trader to purchase bundle r in net position q is

$$\underbrace{\langle r, \phi(\omega)\rangle}_{\text{Payoff for } r} - \underbrace{C(q+r) - C(q)}_{\text{Cost to buy } r} = V_\omega^\phi(q+r) - V_\omega^\phi(q)$$

where $V_\omega^\phi(q) = \langle q, \phi(\omega)\rangle - C(q)$ is the trader "value potential".

How does the potential V_ω^ϕ for an incomplete market with cost function C relate to V_ω for the underlying complete market with cost function B?

Theorem 4 : Complete and Incomplete Markets

There is an *bundle mapping* $f : \mathbb{R}^k \to \mathbb{R}^\Omega$ s.t. $V_\omega(f(q)) = V_\omega^\phi(q) \quad \forall \omega, q$
$\iff$ C^* is B^*-regular for ϕ — i.e, $C^*(r) = \inf_p \{B^*(p) : \mathbb{E}_p[\phi] = r\}$

Interpretation: The incomplete AMM assigns "maximum entropy prices" to underlying complete market based on trade in incomplete market.

Conclusions

Several properties of (classical) exponential families can be obtained simply and with much generality (i.e., for infinite outcomes) via convex duality:

- Normalisation $\nabla(-H)^*(\phi^\top\theta) \in \Delta_\Omega$
- Means as derivatives of the cumulant $\mathbb{E}_p[\phi] = \nabla C(\theta)$
- Information geometry on natural parameters $KL(p_\theta, p_{\theta'}) = D_C(\theta', \theta)$
- (Bijection between mean and natural parameterisations)

Conclusions

Several properties of (classical) exponential families can be obtained simply and with much generality (i.e., for infinite outcomes) via convex duality:

- Normalisation $\partial F^*(\phi^\top \theta) \subseteq \Delta_\Omega$
- Means as derivatives of the cumulant $\mathbb{E}_p[\phi] \in \partial C(\theta)$
- Information geometry on natural parameters $D_F(p_\theta, p_{\theta'}) = D_C(\theta', \theta)$
- (Bijection between mean and natural parameterisations)

Moreover, the above properties all generalise to MaxEnt models (GEFs) for alternative entropies (i.e., arbitrary convex, l.s.c. functions on Δ_Ω).

Conclusions

Several properties of (classical) exponential families can be obtained simply and with much generality (i.e., for infinite outcomes) via convex duality:

- Normalisation $\partial F^*(\phi^\top \theta) \subseteq \Delta_\Omega$
- Means as derivatives of the cumulant $\mathbb{E}_p[\phi] \in \partial C(\theta)$
- Information geometry on natural parameters $D_F(p_\theta, p_{\theta'}) = D_C(\theta', \theta)$
- (Bijection between mean and natural parameterisations)

Moreover, the above properties all generalise to MaxEnt models (GEFs) for alternative entropies (i.e., arbitrary convex, l.s.c. functions on Δ_Ω).

Emphasising the convex foundations of these probabilistic families highlights connections to Bregman divergences and prediction markets.

Conclusions

Several properties of (classical) exponential families can be obtained simply and with much generality (i.e., for infinite outcomes) via convex duality:

- Normalisation $\partial F^*(\phi^\top \theta) \subseteq \Delta_\Omega$
- Means as derivatives of the cumulant $\mathbb{E}_p[\phi] \in \partial C(\theta)$
- Information geometry on natural parameters $D_F(p_\theta, p_{\theta'}) = D_C(\theta', \theta)$
- (Bijection between mean and natural parameterisations)

Moreover, the above properties all generalise to MaxEnt models (GEFs) for alternative entropies (i.e., arbitrary convex, l.s.c. functions on Δ_Ω).

Emphasising the convex foundations of these probabilistic families highlights connections to Bregman divergences and prediction markets.

Thanks!