Lecture 2: Exponential Families and Variational Representation

Tianpei Xie

Aug. 25th., 2022

Contents

1 Background knowledge

- The commonly used function representation for distributions are the exponential famlity. The joint distribution $p(\boldsymbol{x})$ follows the canonical form ***exponential famlity*** of distribution

$$
\begin{aligned}
p(x_1, \ldots, x_m) = p(\boldsymbol{x}; \boldsymbol{\eta}) &= \exp\left(\langle \boldsymbol{\eta}\,,\, \boldsymbol{\phi}(\boldsymbol{x})\rangle - A(\boldsymbol{\eta})\right) h(\boldsymbol{x})\nu(d\boldsymbol{x}) \\
&= \exp\left(\sum_{\alpha} \eta_\alpha \phi_\alpha(\boldsymbol{x}) - A(\boldsymbol{\eta})\right) \qquad (1)
\end{aligned}
$$

  where ϕ is a feature map and $\boldsymbol{\phi}(\boldsymbol{x})$ defines a set of ***sufficient statistics*** (or ***potential functions***). The normalization factor is defined as

$$
A(\boldsymbol{\eta}) := \log \int \exp\left(\langle \boldsymbol{\eta}\,,\, \boldsymbol{\phi}(\boldsymbol{x})\rangle\right) h(\boldsymbol{x})\nu(d\boldsymbol{x}) = \log Z(\boldsymbol{\eta})
$$

  $A(\boldsymbol{\eta})$ is also referred as ***log-partition function*** or *cumulant function*. The parameters $\boldsymbol{\eta} = (\eta_\alpha)$ are called ***natural parameters*** or *canonical parameters*. $\boldsymbol{\eta} \in \{\boldsymbol{\eta} \in \mathbb{R}^d : A(\boldsymbol{\eta}) < \infty\}$, which is called *natural parameter space*. Note that $A(\boldsymbol{\eta})$ is a convex function.

  In exponential family, due to property of exponent, we can formulate the (unnormalized) local functions as a exponential family too

$$
\phi_C(\boldsymbol{x}_C; \boldsymbol{\eta}_C) = \exp\left(\sum_{k_C} \eta_{k_C} \phi_{k_C}(\boldsymbol{x}_C)\right)
$$

  Commonly known distribution and there natural parameterization:

  - Bernoulli distribution $\mathrm{B}(x; p)$: $\nu =$ Counting measure, $\eta = \log(p/(1-p))$, $\phi(x) = x$

$$
\begin{aligned}
\langle \eta\,,\, \phi(x)\rangle &= \log\left(\frac{p}{1-p}\right) x \\
A(\eta) &= -\log(1-p) = \log(1 + \exp(\eta))
\end{aligned}
$$

  - Gaussian distribution $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$: $\nu =$ Lebesgue measure $\mathbb{R}^d$, $h(\boldsymbol{x}) = \frac{1}{(2\pi)^d}$,

$$
\begin{aligned}
\boldsymbol{\eta} &= \left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, -\frac{1}{2}\mathrm{vec}(\boldsymbol{\Sigma}^{-1})\right) := \left(\boldsymbol{\theta}, -\frac{1}{2}\mathrm{vec}(\boldsymbol{\Theta})\right) \qquad (2) \\
\boldsymbol{\phi}(\boldsymbol{x}) &= (\boldsymbol{x}, \mathrm{vec}(\boldsymbol{x}\boldsymbol{x}^T)) \qquad (3) \\
\langle \boldsymbol{\eta}\,,\, \boldsymbol{\phi}(x)\rangle &= \boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{\theta} - \frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Theta}\boldsymbol{x} \\
A(\boldsymbol{\eta}) &= \frac{1}{2}\left(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \log\det|\boldsymbol{\Sigma}|\right) = \frac{1}{2}\left(\boldsymbol{\theta}^T\boldsymbol{\Theta}^{-1}\boldsymbol{\theta} - \log\det|\boldsymbol{\Theta}|\right) \qquad (4)
\end{aligned}
$$

  - Poisson distribution Possion(λ): $\nu =$ Counting measure $h(x) = 1/(x!)$, $\eta = \log(\lambda)$, $\phi(x) = x$

$$
\begin{aligned}
\langle \boldsymbol{\eta}\,,\, \boldsymbol{\phi}(x)\rangle &= \log(\lambda)x \\
A(\eta) &= \lambda = \exp(\eta)
\end{aligned}
$$

- Gamma distribution $\Gamma(\alpha, \lambda)$: $\nu =$ Lebesgue measure $(0, \infty)$, $\boldsymbol{\eta} = (-\lambda, \alpha - 1)$ and $\boldsymbol{\phi}(x) = (x, \log(x))$

$$\begin{aligned}\langle \boldsymbol{\eta} \, , \, \boldsymbol{\phi}(x) \rangle &= -\lambda x + (\alpha - 1) \log(x) \\ A(\eta) &= \log(\Gamma(\alpha)) - \alpha \log(\lambda) = \log(\Gamma(\eta_2 + 1)) - (\eta_2 + 1) \log(-\eta_1)\end{aligned}$$

- We can *re-parameterize* the exponential family by choosing $\boldsymbol{\theta}$ as parameter when $\boldsymbol{\eta} := \boldsymbol{\eta}(\boldsymbol{\theta})$. In (1), if $\boldsymbol{\eta} := \boldsymbol{\eta}(\boldsymbol{\theta}) = \boldsymbol{\theta}$, we call it canonical form.

$$\begin{aligned}p(x_1, \ldots, x_m) = p(\boldsymbol{x}; \boldsymbol{\theta}) &= \exp\left(\langle \boldsymbol{\eta}(\boldsymbol{\theta}) \, , \, \boldsymbol{\phi}(\boldsymbol{x}) \rangle - A(\boldsymbol{\eta}(\boldsymbol{\theta}))\right) \\ &= \exp\left(\sum_{\alpha} \eta_\alpha(\boldsymbol{\theta}) \phi_\alpha(\boldsymbol{x}) - A(\boldsymbol{\eta}(\boldsymbol{\theta}))\right) \qquad (5)\end{aligned}$$

From [Wainwright et al., 2008] we can see that the form in (1) and (5) are both ***conjugate*** to each other based on convex analysis.

  - A special form of exponential family is the ***generalized linear models (GLMs)***, when $p(x_s|x_C)$ follows exponential family, $\phi(\boldsymbol{x}) = \boldsymbol{x}$,

$$\boldsymbol{\theta} = \mathbb{E}\left[\boldsymbol{\phi}(\boldsymbol{x})\right] = g^{-1}\left(\langle \boldsymbol{\eta} \, , \, \boldsymbol{x} \rangle\right) \qquad (6)$$

    where g is called the ***link function***, $\langle \boldsymbol{\eta} \, , \, \boldsymbol{x} \rangle$ is referred as linear predictor or system components.

- ***Minimal***: It is typical to define an exponential family with a vector of sufficient statistics $\boldsymbol{\phi}(\boldsymbol{x})$ for which there **does not exist** a nonzero vector $\boldsymbol{a} \in \mathbb{R}^d$ such that the linear combination

$$\sum_{\alpha \in \mathcal{I}} a_\alpha \phi_\alpha(x) = \text{const.} \quad (\nu\text{-almost everywhere})$$

This condition gives rise to a so-called ***minimal representation***, in which there is a unique parameter vector $\boldsymbol{\mu}$ associated with each distribution.

2 Exponential family via maximum entropy

2.1 Maximum entropy estimation

The exponential family (1) is the **unique solution** to the following ***maximum entropy estimation*** problem:

$$\min_{q \in \Delta} \quad \mathbb{KL}\left(q \,\|\, p_0\right) \tag{7}$$

$$\text{s.t.} \quad \mathbb{E}_q\left[\phi_\alpha(X)\right] = \mu_\alpha \quad \forall\, \alpha \in \mathcal{I} \tag{8}$$

where $\mathbb{KL}\left(q \,\|\, p_0\right) = \int \log(\frac{q}{p_0}) q dx = \mathbb{E}_q\left[\log \frac{q}{p_0}\right]$ is the relative entropy or the Kullback-Leibler divergence of q w.r.t. p_0. To see this, we have the Lagrangian function

$$\mathcal{L}(q, \{\eta_\alpha\}) := \mathbb{KL}\left(q \,\|\, p_0\right) - \sum_\alpha \eta_\alpha \left[\mathbb{E}_q\left[\phi_\alpha(X)\right] - \mu_\alpha\right]$$

$$\frac{\partial \mathcal{L}}{\partial q} = \log\left(\frac{q}{p_0}\right) + 1 - \sum_\alpha \eta_\alpha \phi_\alpha(x) = 0$$

The equation gives the exponential family in canonical form

$$q(x) = \exp\left(\sum_\alpha \eta_\alpha \phi_\alpha(x) - A(\boldsymbol{\eta})\right)$$

Also note that $\mathbb{KL}\left(q \,\|\, p_0\right)$ is **convex** w.r.t. q, therefore the optimal solution is unique.

The canonical parameter $\{\eta_\alpha\}$ forms a **canonical parameter space**

$$\Omega = \left\{\boldsymbol{\eta} \in \mathbb{R}^d : A(\boldsymbol{\eta}) < \infty\right\} \tag{9}$$

The mean constraint (**moment matching** conditions) (8) defines a set of ***mean parameters*** $\{\mu_\alpha\}_{\alpha \in \mathcal{I}}$ one for each of the $|\mathcal{I}| = d$ sufficient statistics ϕ_α, with respect to an arbitrary density q. An interesting object is the set of all such vectors $\boldsymbol{\mu} \in \mathbb{R}^d$ traced out as the underlying density q is varied. More formally, we define the **mean parameter space**

$$\mathcal{M} := \left\{\boldsymbol{\mu} \in \mathbb{R}^d : \exists q \text{ s.t. } \mathbb{E}_q\left[\phi_\alpha(X)\right] = \mu_\alpha \quad \forall\, \alpha \in \mathcal{I}\right\} \tag{10}$$

We see that $\mathcal{M}$ is the *feasible region* of the maximum entropy optimization (7). We see that $\mathcal{M}$ is a *convex hull* spanned by sufficient statistics $\{\phi_\alpha\}_{\alpha \in \mathcal{I}}$

$$\begin{aligned}\mathcal{M} &= \left\{\boldsymbol{\mu} \in \mathbb{R}^d : \sum_{x \in \mathcal{X}^m} q(x)\phi_\alpha(x) = \mu_\alpha \text{ for some } \boldsymbol{q} \in \Delta_{|\mathcal{X}|},\ \forall\, \alpha \in \mathcal{I}\right\} \\ &= \operatorname{conv}\left\{\phi_\alpha(x),\ x \in \mathcal{X},\ \alpha \in \mathcal{I}\right\}\end{aligned}$$

It is thus a **convex polytope**.

Note that if the sufficient statistics are chosen as indicator functions of variables, the expectation constraint (8) becomes marginal distribution constraint.

$$\phi_{s;j}(x_s) = \mathbb{1}\left\{x_s = j\right\}$$

Moreover, for each edge (s,t) and pair of values $(j,k) \in \mathcal{X} \times \mathcal{X}$, define the sufficient statistics

$$\phi_{st;jk}(x_s, x_t) = \mathbb{1}\{x_s = j \wedge x_t = k\}$$

Thus the mean constraints becomes

$$\begin{aligned}\mathbb{E}_q\left[\mathbb{1}\{x_s = j \wedge x_t = k\}\right] = \mu_{st;jk} = \mathbb{Q}(X_s = j \wedge X_t = k), \quad \forall s,t,j,k \\ \mathbb{E}_q\left[\mathbb{1}\{x_s = j\}\right] = \mu_{s;j} = \mathbb{Q}(X_s = j), \quad \forall s,j\end{aligned}$$

Thus $\mathcal{M}$ defined in (10) is also referred as the ***marginal polytope*** associated with the graph $\mathcal{G}$.

2.2 Properties of log-partition function

For the log-partition function (or cumulant function) $A(\boldsymbol{\eta})$ we have the following theorem:

Theorem 2.1 *The log-partition function $A(\boldsymbol{\eta})$ is defined as*

$$A(\boldsymbol{\eta}) := \log \int \exp\left(\langle \boldsymbol{\eta}\,,\, \boldsymbol{\phi}(\boldsymbol{x})\rangle\right) h(\boldsymbol{x})\nu(d\boldsymbol{x}) \tag{11}$$

associated with any regular *exponential family has the following properties:*

- *It has derivatives of* ***all orders*** *on its domain Ω. The first two derivatives yield the* ***cumulants*** *of the random vector $\phi(X)$ as follows:*

$$\frac{\partial A}{\partial \eta_\alpha} = \mathbb{E}_{\boldsymbol{\eta}}\left[\phi_\alpha(X)\right] := \int_{\mathcal{X}^m} \phi_\alpha(\boldsymbol{x}) q(\boldsymbol{x};\boldsymbol{\eta}) d\boldsymbol{x} \tag{12}$$

$$\frac{\partial^2 A}{\partial \eta_\alpha \partial \eta_\beta} = \mathbb{E}_{\boldsymbol{\eta}}\left[\phi_\alpha(X)\phi_\beta(X)\right] - \mathbb{E}_{\boldsymbol{\eta}}\left[\phi_\alpha(X)\right]\mathbb{E}_{\boldsymbol{\eta}}\left[\phi_\beta(X)\right] \tag{13}$$

- *Moreover, A is a* ***convex*** *function of $\boldsymbol{\eta}$ on its domain Ω, and strictly so if the representation is* ***minimal****.*

From (12), we see that the **gradient of log-partition function** define a mapping $\nabla A : \Omega \to \mathcal{M}$ from canonical parameters $\boldsymbol{\eta}$ to the mean parameters $\boldsymbol{\mu}$. This mapping is called ***forward mapping*** [Wainwright et al., 2008]. The forward mapping ∇A is ***one-to-one mapping*** when the exponential family is minimal.

Proposition 2.2 *The gradient mapping $\nabla A : \Omega \to \mathcal{M}$ is* ***one-to-one*** *if and only if the exponential representation is* ***minimal****.*

In particular, we say that the pair $(\boldsymbol{\eta}, \boldsymbol{\mu})$ are ***dual coupled*** if $\nabla A(\boldsymbol{\eta}) = \boldsymbol{\mu}$.

We now consider the image $\nabla A(\Omega)$ of the domain of valid canonical parameters Ω under the gradient mapping ∇A. We have the following theorem:

Theorem 2.3 *In a minimal exponential family, the gradient map ∇A is* ***onto*** *the* ***interior*** *of $\mathcal{M}$, denoted by $\mathcal{M}^\circ$. Consequently, for each $\boldsymbol{\mu} \in \mathcal{M}^\circ$, there exists some $\boldsymbol{\eta} = \boldsymbol{\eta}(\boldsymbol{\mu}) \in \Omega$ such that $\mathbb{E}_{\boldsymbol{\eta}}\left[\phi(X)\right] = \boldsymbol{\mu}$*

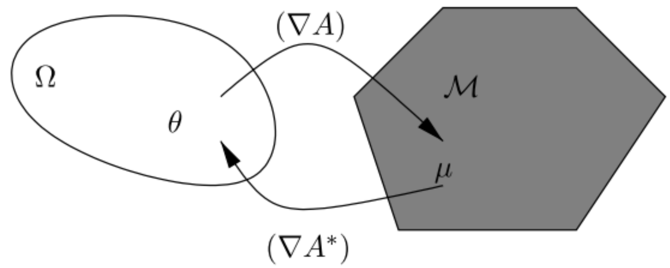

Fig. 3.8 Idealized illustration of the relation between the set Ω of valid canonical parameters, and the set $\mathcal{M}$ of valid mean parameters. The gradient mappings ∇A and ∇A^* associated with the conjugate dual pair (A, A^*) provide a bijective mapping between Ω and the interior $\mathcal{M}^\circ$.

Figure 1: The forward and backward mapping between the cannoical parameter region to the marginal polytope.

This implies that besides the ***canonical parameterization*** via $\boldsymbol{\eta}$, the exponential family also has an equivalent parameterization: the ***mean parameterization***:

$$p(\boldsymbol{x}; \boldsymbol{\mu}) := p(\boldsymbol{x}; \boldsymbol{\eta}(\boldsymbol{\mu})) = \exp\left(\langle \boldsymbol{\eta}(\boldsymbol{\mu})\,,\, \boldsymbol{\phi}(\boldsymbol{x}) \rangle - A(\boldsymbol{\eta}(\boldsymbol{\mu}))\right)$$

For instance $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a mean parameterization of the normali distribution.

This fact is remarkable: *it means that (disregarding boundary points)* ***all mean parameters $\mathcal{M}$ that are realizable by some distribution can be realized by a member of the exponential family***. From this point of view, the maximum entropy problem (7) is just to **project** the prior distribution p_0 into the space of exponential families $\mathcal{M}$. The moment matching conditions (8) are identical to those defining the maximum likelihood problem.

2.3 Conjugate Duality: Maximum Likelihood and Maximum Entropy

The convex ***conjugate dual*** of log-partition function A is defined as

$$A^*(\boldsymbol{\mu}) := \sup_{\boldsymbol{\eta} \in \Omega} \left\{ \langle \boldsymbol{\mu}\,,\, \boldsymbol{\eta} \rangle - A(\boldsymbol{\eta}) \right\} \tag{14}$$

Here $\boldsymbol{\mu} \in \mathbb{R}^d$ is a fixed vector of so-called **dual variables** of the same dimension as $\boldsymbol{\eta}$.

The conjugate dual function (14) is the **negative entropy**. When $\boldsymbol{\mu} \in \mathcal{M}^\circ$, then

$$A^*(\boldsymbol{\mu}) = \int \log q(\boldsymbol{x}; \boldsymbol{\eta}(\boldsymbol{\mu}))\, q(\boldsymbol{x}; \boldsymbol{\eta}(\boldsymbol{\mu}))\nu(d\boldsymbol{x}) := -H(q_{\boldsymbol{\eta}(\boldsymbol{\mu})}),$$

where the q is the exponential family (1) with canonical parameter $\boldsymbol{\eta}(\boldsymbol{\mu})$ and the moment matching conditions are met

$$\mathbb{E}_{\boldsymbol{\eta}(\boldsymbol{\mu})}\left[\boldsymbol{\phi}(X)\right] = \boldsymbol{\mu} \tag{15}$$

This fact is essential in the use of **variational methods**: it guarantees that any optimization problem involving the dual function can be reduced to an optimization problem over $\mathcal{M}$.

Theorem 2.4 *[Wainwright et al., 2008]*

1. *For any $\boldsymbol{\mu} \in \mathcal{M}^\circ$, denote by $\boldsymbol{\eta}(\boldsymbol{\mu})$ the unique canonical parameter satisfying the dual matching condition* (19). *The conjugate dual function A^* takes the form*

$$A^*(\boldsymbol{\mu}) = \begin{cases} -H(q_{\boldsymbol{\eta}(\boldsymbol{\mu})}) & \text{if } \boldsymbol{\mu} \in \mathcal{M}^\circ \\ +\infty & \text{if } \boldsymbol{\mu} \notin \overline{\mathcal{M}} \end{cases} \tag{16}$$

For any boundary point $\boldsymbol{\mu} \in \partial\mathcal{M} := \overline{\mathcal{M}} \setminus \mathcal{M}^\circ$ we have

$$A^*(\boldsymbol{\mu}) = \lim_{n\to\infty} A^*(\boldsymbol{\mu}_n) \tag{17}$$

taken over any sequence $\{\boldsymbol{\mu}_n\} \subset \mathcal{M}^\circ$ converging to μ.

2. *In terms of this dual, the log-partition function has the* ***variational representation***

$$A(\boldsymbol{\eta}) = \sup_{\boldsymbol{\mu}\in\mathcal{M}} \{\langle \boldsymbol{\eta}\,,\, \boldsymbol{\mu}\rangle - A^*(\boldsymbol{\mu})\} \tag{18}$$

3. *For all $\boldsymbol{\eta} \in \Omega$, the supremum in Equation* (18) *is attained uniquely at the vector $\boldsymbol{\mu} \in \mathcal{M}^\circ$ specified by the moment matching conditions*

$$\mathbb{E}_{\boldsymbol{\eta}}\left[\boldsymbol{\phi}(X)\right] = \int_{\mathcal{X}^m} \boldsymbol{\phi}(\boldsymbol{x})q(\boldsymbol{x};\boldsymbol{\eta})\nu(d\boldsymbol{x}) = \boldsymbol{\mu} \tag{19}$$

The above theorm establishes the duality between log-partition function A and negative entropy A^*. Note that A^* is a function of mean parameter $\boldsymbol{\mu}$ not like usual entropy as function of distribution. Moreover, the value $-A^*(\boldsymbol{\mu})$ corresponds to the **optimal value** of maximum entropy optimization (7). Thus (18) formulate the maximum entropy estimation problem in (7). Third, above theorem also clarifies the precise nature of the ***bijection*** between the sets Ω and $\mathcal{M}^\circ$, which holds for any minimal exponential family. In particular, the gradient mapping ∇A maps Ω in a one-to-one manner onto $\mathcal{M}^\circ$, whereas the ***inverse mapping*** from $\mathcal{M}^\circ$ to Ω is given by the gradient ∇A^* of the dual function. The mapping $\nabla A^* : \mathcal{M}^\circ \to \Omega$ is called ***backward mapping***. See Figure 1 for an idealized illustration of this **bijective** correspondence based on the gradient mappings $(\nabla A, \nabla A^*)$.

With conjugate dual, we see that the ***maximum likelihood estimation*** problem is essentially the **dual problem** of the maximum entropy estimation (7).

$$\begin{aligned} &\max_{\boldsymbol{\eta}} \frac{1}{N}\sum_{n=1}^{N} \log q_{\boldsymbol{\eta}}(X_n) \\ \Rightarrow &\max_{\boldsymbol{\eta}} \langle \bar{\boldsymbol{\mu}}\,,\, \boldsymbol{\eta}\rangle - A(\boldsymbol{\eta}) \end{aligned} \tag{20}$$

where $\bar{\boldsymbol{\mu}} = \hat{\mathbb{E}}[\boldsymbol{\phi}(X)] = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{\phi}(X_n)$ fits the moment matching conditions. (20) is the right-hand side of the conjugate dual of log-partition function A^* in (14). Thus we have one statistical interpretation of this variational problem (14): A^* is the **optimal value of the rescaled log likelihood** (20).

Also see that the gradient of log-likelihood function

$$\nabla_{\boldsymbol{\eta}} \frac{1}{N}\sum_{n=1}^{N} \log q_{\boldsymbol{\eta}}(X_n) = \nabla_{\boldsymbol{\eta}}(\langle \bar{\boldsymbol{\mu}}\,,\, \boldsymbol{\eta}\rangle - A(\boldsymbol{\eta}))$$

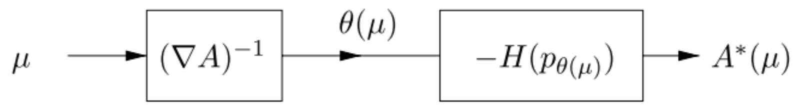

Fig. 3.9 A block diagram decomposition of A^* as the composition of two functions. Any mean parameter $\mu \in \mathcal{M}^\circ$ is first mapped back to a canonical parameter $\theta(\mu)$ in the inverse image $(\nabla A)^{-1}(\mu)$. The value of $A^*(\mu)$ corresponds to the negative entropy $-H(p_{\theta(\mu)})$ of the associated exponential family density $p_{\theta(\mu)}$.

Figure 2: The computation of A^*.

$$
\begin{aligned}
&= \hat{\mathbb{E}}[\boldsymbol{\phi}(X)] - \mathbb{E}_{\boldsymbol{\eta}}\left[\boldsymbol{\phi}(X)\right] \qquad (21)\\
&= \bar{\boldsymbol{\mu}} - \boldsymbol{\mu} = \text{sample mean} - \text{model mean}
\end{aligned}
$$

2.4 Challenges in high dimensional setting

For general multivariate exponential families, there are two **primary challenges** associated with the **variational representation**:

1. In many cases, the constraint set $\mathcal{M}$ of realizable mean parameters is extremely difficult to characterize in an **explicit** manner. Note that even in discrete cases, the number of constraints defining $\mathcal{M}$ grows exponentially with respect to dimension of sample space $\mathcal{X}^m$.

2. The negative entropy function A^* is defined indirectly in a variational manner so that it too typically **lacks an explicit form**.

To understand the complexity inherent in evaluating the dual value $A^*(\boldsymbol{\mu})$, note that Theorem 2.4 provides only an implicit characterization of A^* as the **composition** of mappings: first, the **inverse mapping** $(\nabla A)^{-1} : \mathcal{M}^\circ \to \Omega$, in which $\boldsymbol{\mu}$ maps to $\boldsymbol{\eta}(\boldsymbol{\mu})$, corresponding to the *exponential family member* with **mean parameters** $\boldsymbol{\mu}$; and second, the mapping from $\boldsymbol{\eta}(\boldsymbol{\mu})$ to the negative entropy $-H(q_{\boldsymbol{\eta}(\boldsymbol{\mu})})$ of the associated exponential family density. This decomposition of the value $A^*(\boldsymbol{\mu})$ is illustrated in Figure 2. Computing the inverse mapping $(\nabla A)^{-1}$ as well as entropy $-H$ are both challenging in high dimensional setting. These difficulties motivate the use of **approximations** to $\mathcal{M}$ and A^*.

2.5 Primal-dual formulation of KL divergence

The conjugate duality between A and A^*, as characterized in Theorem (2.4), leads to several alternative forms of the KL divergence for *exponential family members.*

$$
\mathbb{KL}\left(q \,\|\, p\right) := \int_{\mathcal{X}^m} q(x) \log\left[\frac{q(x)}{p(x)}\right] \nu(dx)
$$

Consider two canonical parameter vectors $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \in \Omega$, we formulate the KL divergence between two distributions in exponential family

$$
\begin{aligned}
\mathbb{KL}\left(p_{\boldsymbol{\eta}_1} \,\|\, p_{\boldsymbol{\eta}_2}\right) &\equiv \mathbb{KL}\left(\boldsymbol{\eta}_1 \,\|\, \boldsymbol{\eta}_2\right) \\
&:= \mathbb{E}_{\boldsymbol{\eta}_1}\left[A(\boldsymbol{\eta}_2) - A(\boldsymbol{\eta}_1) - \langle \boldsymbol{\phi}(X)\,,\, \boldsymbol{\eta}_2 - \boldsymbol{\eta}_1\rangle\right]
\end{aligned}
$$

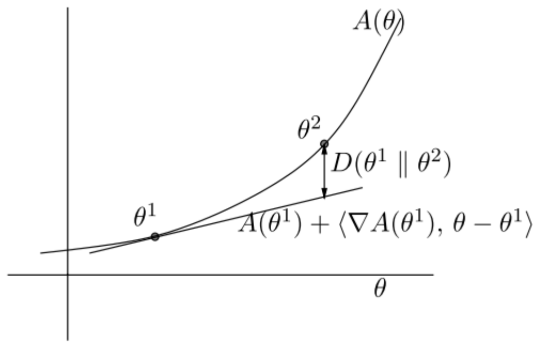

Fig. 5.1 The hyperplane $A(\theta^1) + \langle \nabla A(\theta^1), \theta - \theta^1 \rangle$ supports the epigraph of A at θ^1. The Kullback–Leibler divergence $D(\theta^1 \,\|\, \theta^2)$ is equal to the difference between $A(\theta^2)$ and this tangent approximation.

Figure 3: The geometrical intepretation using primal A and dual A^*

$$
\begin{aligned}
&= A(\boldsymbol{\eta}_2) - A(\boldsymbol{\eta}_1) - \langle \boldsymbol{\mu}_1 \,,\, \boldsymbol{\eta}_2 - \boldsymbol{\eta}_1 \rangle \\
&\equiv A(\boldsymbol{\eta}_2) - A(\boldsymbol{\eta}_1) - \langle \nabla A(\boldsymbol{\eta}_1) \,,\, \boldsymbol{\eta}_2 - \boldsymbol{\eta}_1 \rangle
\end{aligned} \tag{22}
$$

where $\boldsymbol{\mu}_1 = \mathbb{E}_{\boldsymbol{\eta}_1}[\boldsymbol{\phi}(X)] = \nabla A(\boldsymbol{\eta}_1)$ is the mean parameters for $p_{\boldsymbol{\eta}_1}$. This is the ***primal form* of the KL divergence** at $p_{\boldsymbol{\eta}_1}$. As illustrated in Figure 3, this form of the KL divergence can be interpreted as the difference between $A(\boldsymbol{\eta}_2)$ and the *hyperplane* ***tangent*** to A at $\boldsymbol{\eta}_1$ with normal $\nabla A(\boldsymbol{\eta}_1) = \boldsymbol{\mu}_1$.

A second form of the KL divergence can be obtained by using the strong duality condition (18) for dually coupled parameters.

$$
\begin{aligned}
\mathbb{KL}\left(\boldsymbol{\eta}_1 \,\|\, \boldsymbol{\eta}_2\right) \equiv \mathbb{KL}\left(\boldsymbol{\mu}_1 \,\|\, \boldsymbol{\eta}_2\right) &= A(\boldsymbol{\eta}_2) - A(\boldsymbol{\eta}_1) - \langle \boldsymbol{\mu}_1 \,,\, \boldsymbol{\eta}_2 - \boldsymbol{\eta}_1 \rangle \\
&= A(\boldsymbol{\eta}_2) - (\langle \boldsymbol{\mu} \,,\, \boldsymbol{\eta}_1 \rangle - A^*(\boldsymbol{\mu}_1)) - \langle \boldsymbol{\mu}_1 \,,\, \boldsymbol{\eta}_2 - \boldsymbol{\eta}_1 \rangle \\
&= A(\boldsymbol{\eta}_2) + A^*(\boldsymbol{\mu}_1) - \langle \boldsymbol{\mu}_1 \,,\, \boldsymbol{\eta}_2 \rangle
\end{aligned} \tag{23}
$$

This is the ***primal-dual mixed form* of the KL divergence**.

Finally, we have the ***dual form* of KL divergence**:

$$
\begin{aligned}
\mathbb{KL}\left(\boldsymbol{\mu}_1 \,\|\, \boldsymbol{\mu}_2\right) &= A(\boldsymbol{\eta}_2) + A^*(\boldsymbol{\mu}_1) - \langle \boldsymbol{\mu}_1 \,,\, \boldsymbol{\eta}_2 \rangle \\
&= \langle \boldsymbol{\mu}_2 \,,\, \boldsymbol{\eta}_2 \rangle - A^*(\boldsymbol{\mu}_2) + A^*(\boldsymbol{\mu}_1) - \langle \boldsymbol{\mu}_1 \,,\, \boldsymbol{\eta}_2 \rangle \\
&= A^*(\boldsymbol{\mu}_1) - A^*(\boldsymbol{\mu}_2) - \langle \boldsymbol{\eta}_2 \,,\, \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \rangle \\
&\equiv A^*(\boldsymbol{\mu}_1) - A^*(\boldsymbol{\mu}_2) - \langle \nabla A^*(\boldsymbol{\mu}_2) \,,\, \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \rangle
\end{aligned} \tag{24}
$$

The dual form is related to the *Bregman divergence*, which induce the **projection operation**:

Definition Let $F : \mathcal{X} \to \mathbb{R}$ be a *continuously-differentiable*, ***strictly convex*** function defined on a convex set $\mathcal{X}$. The ***Bregman divergence*** associated with F for points $p, q \in \mathcal{X}$ is the difference between the value of F at point p and the value of the *first-order Taylor expansion* of F around point q evaluated at point p:

$$
\mathbb{D}^F\left(p \,\|\, q\right) = F(p) - F(q) - \langle \nabla F(q) \,,\, p - q \rangle \tag{25}
$$

we see that dual form $\mathbb{KL}\left(\boldsymbol{\mu}_1 \,\|\, \boldsymbol{\mu}_2\right) = \mathbb{D}^{A^*}\left(\boldsymbol{\mu}_1 \,\|\, \boldsymbol{\mu}_2\right)$, where $F = A^*$ is the negative entropy.

References

Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.