Probabilistic Graphical Models

Brown University CSCI 2950-P, Spring 2013
Prof. Erik Sudderth

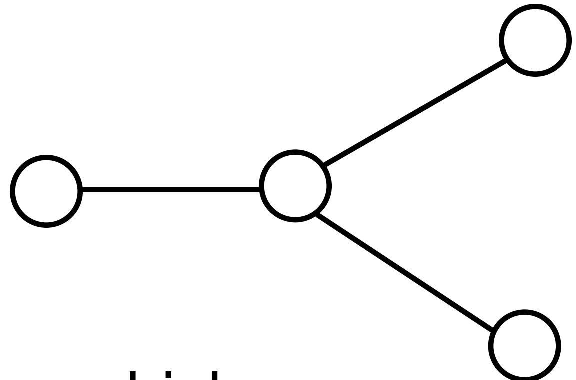
Lecture 21: Convexity, Duality, and Mean Field Methods

Some figures and examples courtesy M. Wainwright & M. Jordan, *Graphical Models, Exponential Families, & Variational Inference*, 2008.

Tree Structured Variational Methods

- Trees exactly factorize as

$$q(x) = \prod_{(s,t)\in\mathcal{E}} \frac{q_{st}(x_s, x_t)}{q_s(x_s), q_t(x_t)} \prod_{s\in\mathcal{V}} q_s(x_s)$$



- We may then optimize over all distributions which are Markov with respect to a tree-structured graph:

$$D(q \,||\, p) = -H(q) + \sum_x q(x)E(x) + \log Z$$

$$\sum_x q(x)E(x) = \sum_{(s,t)\in\mathcal{E}} \sum_{x_s, x_t} q_{st}(x_s, x_t)\phi_{st}(x_s, x_t) + \sum_{s\in\mathcal{V}} \sum_{x_s} q_s(x_s)\phi_s(x_s)$$

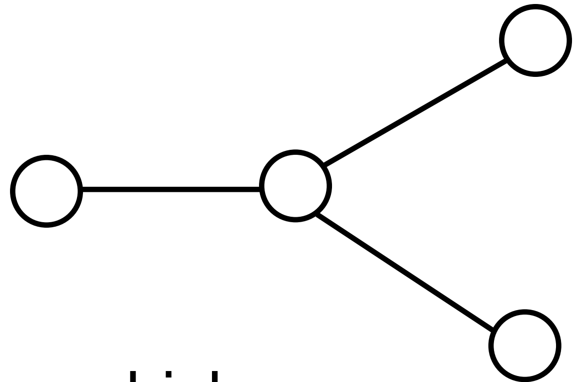
$$H(q) = \sum_{s\in\mathcal{V}} H_s(q_s) - \sum_{(s,t)\in\mathcal{E}} I_{st}(q_{st})$$

Marginal Entropies

Mutual Information

Tree Structured Variational Methods

- Trees exactly factorize as



$$q(x) = \prod_{(s,t)\in\mathcal{E}} \frac{q_{st}(x_s, x_t)}{q_s(x_s), q_t(x_t)} \prod_{s\in\mathcal{V}} q_s(x_s)$$

- We may then optimize over all distributions which are Markov with respect to a tree-structured graph:

$$D(q \,||\, p) = -H(q) + \sum_x q(x)E(x) + \log Z$$

$$\sum_x q(x)E(x) = \sum_{(s,t)\in\mathcal{E}} \sum_{x_s, x_t} q_{st}(x_s, x_t)\phi_{st}(x_s, x_t) + \sum_{s\in\mathcal{V}} \sum_{x_s} q_s(x_s)\phi_s(x_s)$$

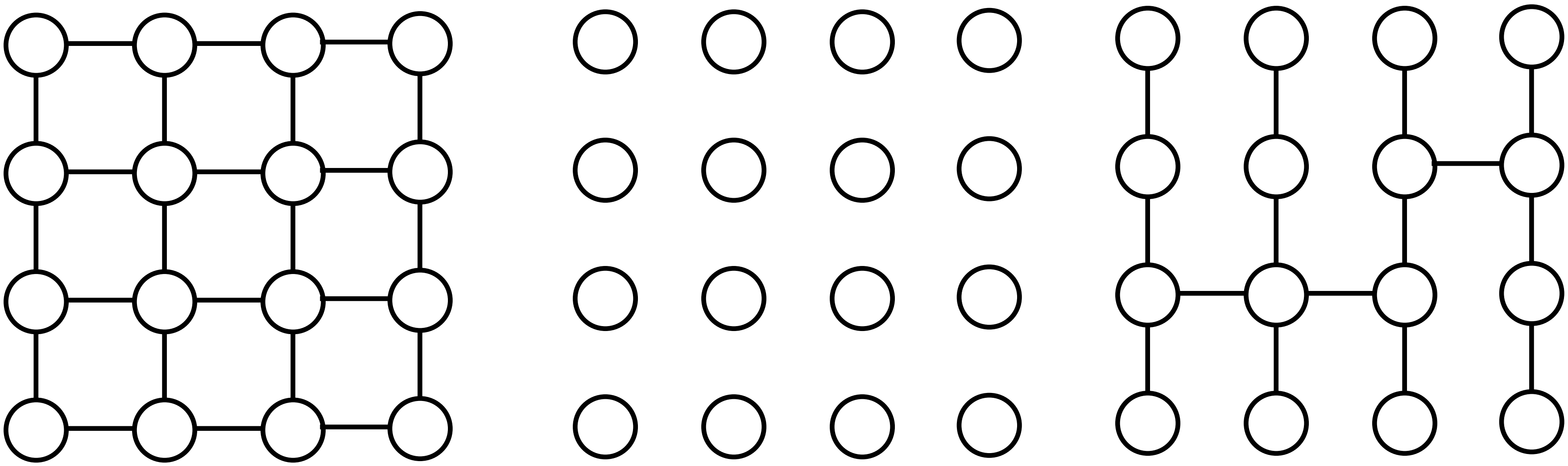
$$H(q) = \sum_{s\in\mathcal{V}} H_s(q_s) - \sum_{(s,t)\in\mathcal{E}} I_{st}(q_{st})$$

$$H_s(q_s) = -\sum_{x_s} q_s(x_s) \log q_s(x_s) \qquad I_{st}(q_{st}) = \sum_{x_s, x_t} q_{st}(x_s, x_t) \log \frac{q_{st}(x_s, x_t)}{q_s(x_s)q_t(x_t)}$$

Mean Field & Belief Propagation

$$p(x) = \frac{1}{Z} \prod_{(s,t)\in\mathcal{E}} \psi_{st}(x_s, x_t) \prod_{s\in\mathcal{V}} \psi_s(x_s)$$

$$\phi_{st}(x_s, x_t) = -\log \psi_{st}(x_s, x_t)$$
$$\phi_s(x_s) = -\log \psi_s(x_s)$$



Original Graph (Loopy BP) **Naïve Mean Field** **Structured Mean Field**

Partition the graph edges into two sets:

$\mathcal{E}_c \longrightarrow$ *core* edges, dependence directly modeled: $q_{st}(x_s, x_t)$

$\mathcal{E}_r \longrightarrow$ *residual* edges, assume nodes factorize: $q_s(x_s)q_t(x_t)$

MF & BP: Variational Objective

$$p(x) = \frac{1}{Z} \prod_{(s,t)\in\mathcal{E}} \psi_{st}(x_s, x_t) \prod_{s\in\mathcal{V}} \psi_s(x_s)$$

$$\phi_{st}(x_s, x_t) = -\log \psi_{st}(x_s, x_t)$$
$$\phi_s(x_s) = -\log \psi_s(x_s)$$

$$\mathcal{L}(q, \lambda) =$$

$$+ \sum_{s\in\mathcal{V}} \sum_{x_s} q_s(x_s)(\phi_s(x_s) + \log q_s(x_s))$$
$$+ \sum_{(s,t)\in\mathcal{E}_r} \sum_{x_s,x_t} q_s(x_s) q_t(x_t) \phi_{st}(x_s, x_t)$$
$$+ \sum_{(s,t)\in\mathcal{E}_c} \sum_{x_s,x_t} q_{st}(x_s, x_t)\left(\phi_{st}(x_s, x_t) + \log \frac{q_{st}(x_s, x_t)}{q_s(x_s) q_t(x_t)}\right)$$
$$+ \sum_{s\in\mathcal{V}} \lambda_{ss}\left(1 - \sum_{x_s} q_s(x_s)\right)$$
$$+ \sum_{(s,t)\in\mathcal{E}_c} \left[\sum_{x_s} \lambda_{ts}(x_s)\left(q_s(x_s) - \sum_{x_t} q_{st}(x_s, x_t)\right) + \sum_{x_t} \lambda_{st}(x_t)\left(q_t(x_t) - \sum_{x_s} q_{st}(x_s, x_t)\right)\right]$$

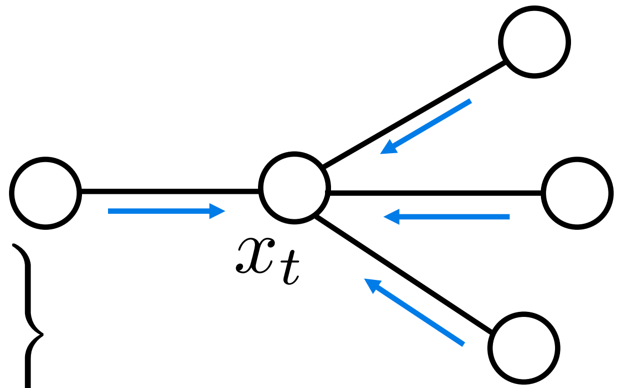
MF & BP: Message Passing

$$p(x) = \frac{1}{Z} \prod_{(s,t)\in\mathcal{E}} \psi_{st}(x_s, x_t) \prod_{s\in\mathcal{V}} \psi_s(x_s)$$

$$\phi_{st}(x_s, x_t) = -\log \psi_{st}(x_s, x_t)$$
$$\phi_s(x_s) = -\log \psi_s(x_s)$$

Beliefs: *pseudo-marginals*

$$q_t(x_t) = \frac{1}{Z_t}\psi_t(x_t) \prod_{u\in\Gamma(t)} m_{ut}(x_t)$$



MF: *residual*

$$m_{ts}(x_s) \propto \exp\left\{ -\sum_{x_t} \phi_{st}(x_s, x_t) q_t(x_t) \right\}$$

BP: *core*

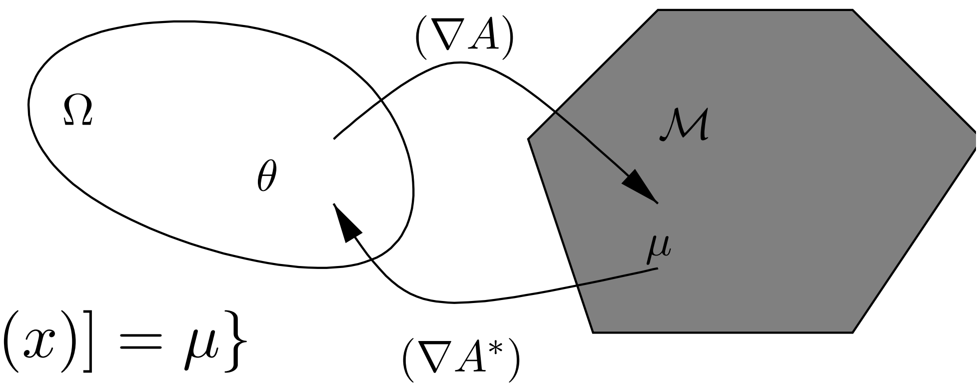
$$m_{ts}(x_s) \propto \sum_{x_t} \psi_{st}(x_s, x_t) \frac{q_t(x_t)}{m_{st}(x_t)}$$

- *Naïve mean field:* All edges in *residual*, guaranteed convergent
- *Structured mean field:* Acyclic subset of edges in *core*, remainder in *residual*, guaranteed convergent and strictly more expressive
- *Loopy belief propagation:* All edges in *core*, captures most direct dependences, but approximation uncontrolled and may not converge
- *All methods:* Exist one, or more, fixed points (possibly non-convex). Strongest convergence guarantees for *sequential* message updates.

Exponential Families: Inference & Learning

$$p(x \mid \theta) = \exp\{\theta^T \phi(x) - A(\theta)\} \qquad A(\theta) = \log \int_{\mathcal{X}} \exp\{\theta^T \phi(x)\}\, dx$$

Alternative Representations:
Canonical parameters or moments



$$\Omega \triangleq \{\theta \in \mathbb{R}^d \mid A(\theta) < +\infty\}$$

$$\mathcal{M} \triangleq \{\mu \in \mathbb{R}^d \mid \exists\, p \text{ such that } \mathbb{E}_p[\phi(x)] = \mu\}$$

Inference: Find moments of model with known parameters

$$\mu = \nabla_\theta A(\theta) = \mathbb{E}_\theta[\phi(x)] = \int_{\mathcal{X}} \phi(x) p(x \mid \theta)\, dx$$

Learning: Find model parameters matching data moments

$$\mathbb{E}_{\hat{\theta}}[\phi(x)] = \hat{\mu}$$ *inverse of mapping required for inference*

ML: $$\hat{\mu} = \frac{1}{N} \sum_{\ell=1}^{N} \phi(x^{(\ell)})$$

MAP: *(conjugate prior)* $$\hat{\mu} = \frac{1}{\alpha + N} \left(\alpha \mu_0 + \sum_{\ell=1}^{N} \phi(x^{(\ell)}) \right)$$

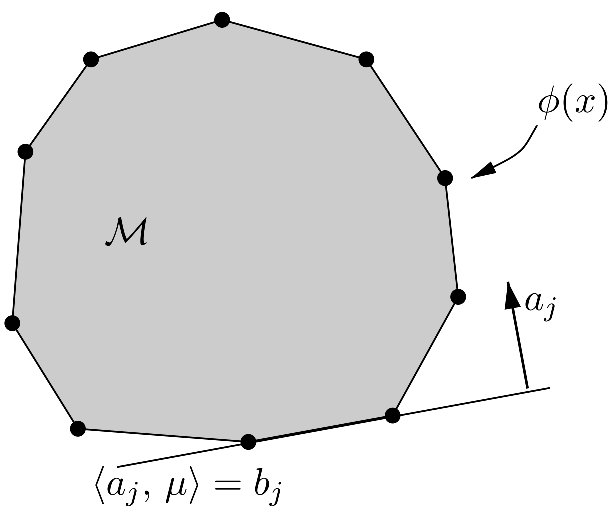
Discrete Variables & Marginal Polytopes

$$p(x \mid \theta) = \exp\{\theta^T \phi(x) - A(\theta)\} \qquad A(\theta) = \log \sum_{\mathcal{X}} \exp\{\theta^T \phi(x)\}$$

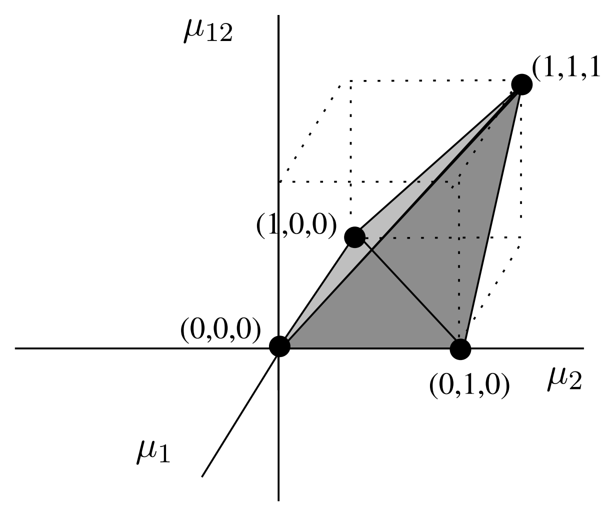
$$\mu = \nabla_\theta A(\theta) = \mathbb{E}_\theta[\phi(x)] = \sum_{\mathcal{X}} \phi(x) p(x \mid \theta)$$

$$\mathcal{M} \triangleq \{\mu \in \mathbb{R}^d \mid \exists\, p \text{ such that } \mathbb{E}_p[\phi(x)] = \mu\} \subseteq [0,1]^d$$

$$\mathcal{M} = \mathrm{conv}\{\phi(x) \mid x \in \mathcal{X}\}$$ *convex hull of possible configurations*



General Convex Polytope



Pair of Binary Variables

Marginal Polytope: Vertices & Faces

- Number of vertices always exponential in number of variables

$$\mu_s = \mathbb{E}_p[X_s] = \mathbb{P}[X_s = 1]$$

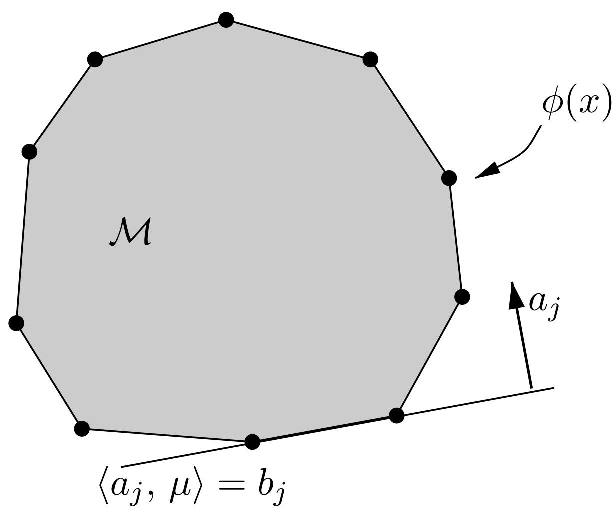
$$\mu_{st} = \mathbb{E}_p[X_s X_t] = \mathbb{P}[(X_s, X_t) = (1,1)]$$

$$\mathcal{M} = \mathrm{conv}\{\phi(x) \mid x \in \mathcal{X}\} \qquad \mathrm{conv}\{(0,0,0),(1,0,0),(0,1,0),(1,1,1)\}$$

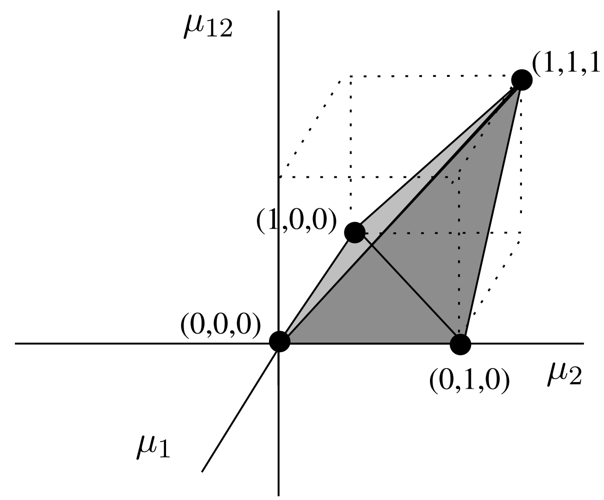
- Number of faces exponential in general, but grows *linearly* with problem size for certain graph topologies

$$\begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \\ -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_{12} \end{bmatrix} \geq \begin{bmatrix} 0 \\ 0 \\ 0 \\ -1 \end{bmatrix}$$

$$\mathcal{M} = \{\mu \in \mathbb{R}^d \mid \langle a_j, \mu \rangle \geq b_j \ \forall j \in \mathcal{J}\}$$

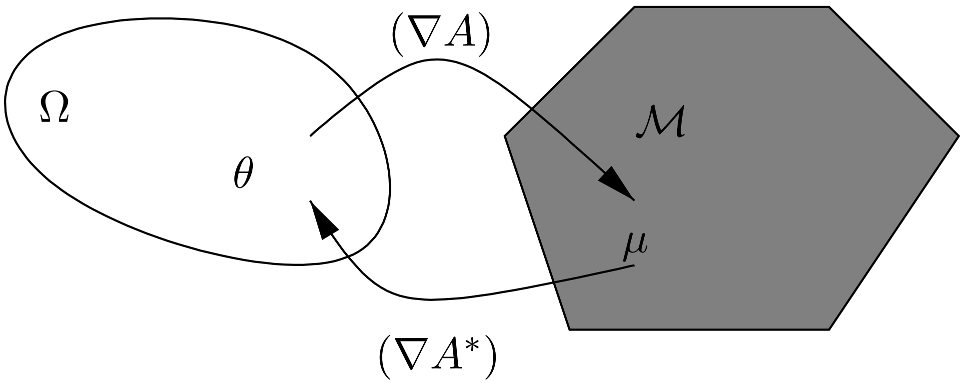


General Convex Polytope



Pair of Binary Variables

Conjugate Duality

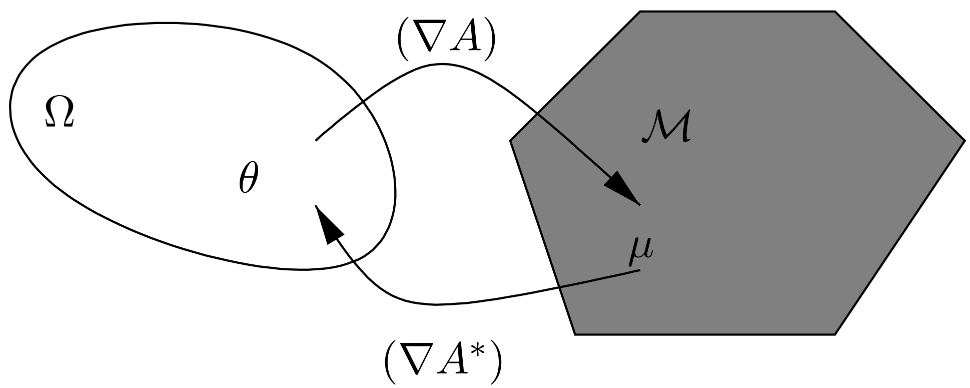


$$A^*(\mu) := \sup_{\theta \in \Omega} \{\langle \mu, \theta \rangle - A(\theta)\}$$

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \big\{\langle \theta, \mu \rangle - A^*(\mu)\big\}$$

Proposition 3.2. The gradient mapping $\nabla A : \Omega \to \mathcal{M}$ is one-to-one if and only if the exponential representation is minimal.

Conjugate Duality



$$A^*(\mu) := \sup_{\theta \in \Omega} \{\langle \mu, \theta \rangle - A(\theta)\}$$

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\}$$

$$\mu \longrightarrow \boxed{(\nabla A)^{-1}} \xrightarrow{\theta(\mu)} \boxed{-H(p_{\theta(\mu)})} \longrightarrow A^*(\mu)$$

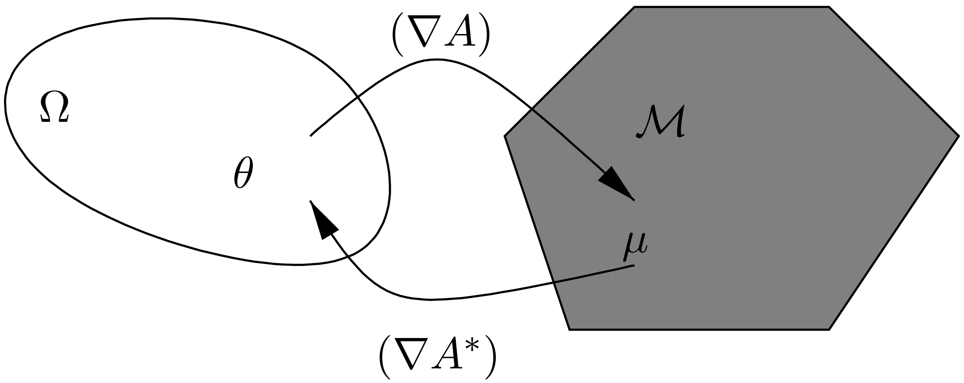
Theorem 3.3. In a minimal exponential family, the gradient map ∇A is onto the interior of $\mathcal{M}$, denoted by $\mathcal{M}^\circ$. Consequently, for each $\mu \in \mathcal{M}^\circ$, there exists some $\theta = \theta(\mu) \in \Omega$ such that $\mathbb{E}_\theta[\phi(X)] = \mu$.

For any $\mu \in \mathcal{M}^\circ$, denote by $\theta(\mu)$ the unique canonical parameter satisfying the dual matching condition (3.43). The conjugate dual function A^* takes the form

$$A^*(\mu) = \begin{cases} -H(p_{\theta(\mu)}) & \text{if } \mu \in \mathcal{M}^\circ \\ +\infty & \text{if } \mu \notin \overline{\mathcal{M}}. \end{cases} \qquad (3.44)$$

$$\mathbb{E}_{\theta(\mu)}[\phi(X)] = \nabla A(\theta(\mu)) = \mu$$

Conjugate Duality



$$A^*(\mu) := \sup_{\theta \in \Omega} \{\langle \mu, \theta \rangle - A(\theta)\}$$

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \left\{\langle \theta, \mu \rangle - A^*(\mu)\right\}$$

$$\mu \longrightarrow \boxed{(\nabla A)^{-1}} \xrightarrow{\theta(\mu)} \boxed{-H(p_{\theta(\mu)})} \longrightarrow A^*(\mu)$$

For all $\theta \in \Omega$, the supremum in Equation (3.45) is attained uniquely at the vector $\mu \in \mathcal{M}^\circ$ specified by the moment-matching conditions

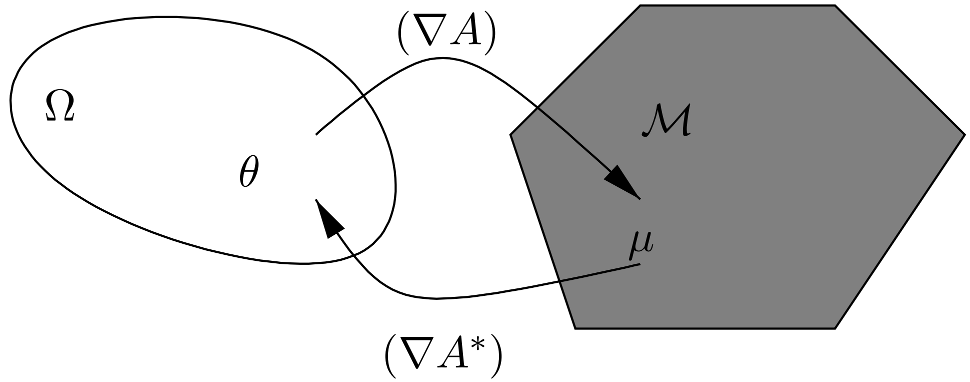
$$\mu = \int_{\mathcal{X}^m} \phi(x) p_\theta(x) \nu(dx) = \mathbb{E}_\theta[\phi(X)]$$

For any $\mu \in \mathcal{M}^\circ$, denote by $\theta(\mu)$ the unique canonical parameter satisfying the dual matching condition (3.43). The conjugate dual function A^* takes the form

$$A^*(\mu) = \begin{cases} -H(p_{\theta(\mu)}) & \text{if } \mu \in \mathcal{M}^\circ \\ +\infty & \text{if } \mu \notin \overline{\mathcal{M}}. \end{cases} \qquad (3.44)$$

$$\mathbb{E}_{\theta(\mu)}[\phi(X)] = \nabla A(\theta(\mu)) = \mu$$

Duality and Variational Inference



$$\mu \longrightarrow \boxed{(\nabla A)^{-1}} \xrightarrow{\theta(\mu)} \boxed{-H(p_{\theta(\mu)})} \longrightarrow A^*(\mu)$$

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \left\{ \langle \theta, \mu \rangle - A^*(\mu) \right\} \qquad A^*(\mu) := \sup_{\theta \in \Omega} \{ \langle \mu, \theta \rangle - A(\theta) \}$$

$$\mu = \int_{\mathcal{X}^m} \phi(x) p_\theta(x) \nu(dx) = \mathbb{E}_\theta[\phi(X)]$$

To infer or approximate moments for known model, we can:

- Represent, or approximate, the marginal polytope
- Compute, bound, or approximate the entropy function
- Derive algorithms for resulting constrained optimization problem

Non-Convexity of Naïve Mean Field

$$p_\theta(x) \propto \exp(\theta_1 x_1 + \theta_2 x_2 + \theta_{12} x_1 x_2) \qquad x_i \in \{-1, +1\}$$

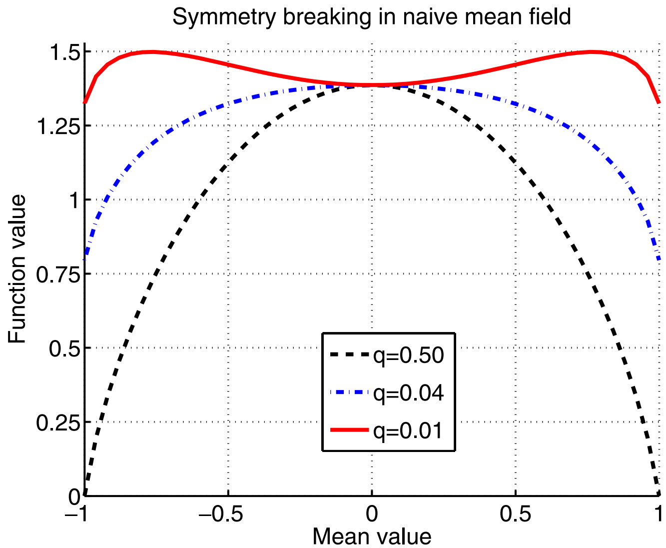
$$f(\mu_1, \mu_2; \theta) = \theta_{12}\mu_1\mu_2 + \theta_1\mu_1 + \theta_2\mu_2 + H(\mu_1) + H(\mu_2)$$

$$H(\mu_i) = -\tfrac{1}{2}(1+\mu_i)\log\tfrac{1}{2}(1+\mu_i) - \tfrac{1}{2}(1-\mu_i)\log\tfrac{1}{2}(1-\mu_i)$$

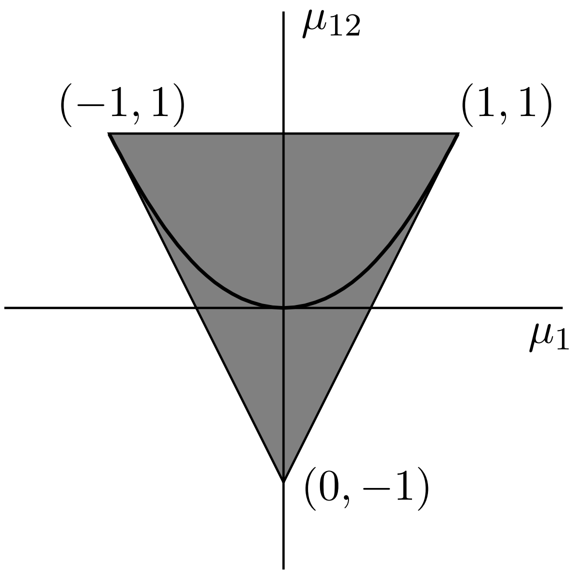
$$(\theta_1, \theta_2, \theta_{12}) = \left(0,\ 0,\ \frac{1}{4}\log\frac{q}{1-q}\right) =:\ \theta(q)$$

$$\mathbb{E}[X_1] = \mathbb{E}[X_2] = 0$$

$$q = \mathbb{P}[X_1 = X_2]$$



$$f(\tau, -\tau; \theta(q))$$



$$\mu_{12} \le 1, \quad \mu_{12} \ge 2\mu_1 - 1, \quad \mu_{12} \ge -2\mu_1 - 1$$