Chapter 5: Mean Field Methods

Abstract

In our continuing crusade to turn everything into an optimization problem, we return to our consideration of cumulant function A under the variational principle $A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\}$ and approximate the problem by considering only the subset of $\mathcal{M}$ for which A^* is tractable. In this chapter review, we discuss selecting the subset of $\mathcal{M}$, interpreting the resulting (non-convex) optimization problem, and ultimately solving the optimization problem.

1 Tractable Families

Consider an exponential family defined by a graph $G = (V, E)$ and the collection of sufficient statistics $\phi = (\phi_\alpha, \alpha \in \mathcal{I})$ associated with the cliques of G. A subgraph F then has a subset $\mathcal{I}(F) \subseteq \mathcal{I}$ associated with the cliques of F. (F, ϕ) defines a sub-family of the full exponential family and is parameterized by

$$\Omega(F) = \{\theta \in \Omega \mid \theta_\alpha = 0 \; \forall \; \alpha \in \mathcal{I} \setminus \mathcal{I}(F)\}. \tag{1}$$

A subgraph F is tractable if $A^*(\theta)$ is easy to compute for all $\theta \in \Omega(F)$. Associated with (G, ϕ) is the set of all mean parameters realizable by any distribution[1],

$$\mathcal{M}(G; \phi) = \left\{\mu \in \mathbb{R}^d \mid \exists \; p \text{ s.t. } \mathbb{E}_p \left[\phi_\alpha(X)\right] = \mu_\alpha \; \forall \; \alpha \in \mathcal{I}\right\}. \tag{2}$$

For the subset of exponential family densities $\{p_\theta, \theta \in \Omega(F)\}$ defined by (F, ϕ), the set of mean parameters that can be obtained is

$$\mathcal{M}_F(G; \phi) = \left\{\mu \in \mathbb{R}^d \mid \exists \; \theta \in \Omega(F) \text{ s.t. } \mu = \mathbb{E}_\theta \left[\phi(x)\right]\right\}. \tag{3}$$

Note that, by definition, $\mathcal{M}_F(G; \phi) = \nabla A(\Omega(F))$. It follows that

$$\mathcal{M}_F(G; \phi) \subseteq \mathcal{M}^\circ(G; \phi). \tag{4}$$

We say $\mathcal{M}_F$ is the inner approximation of $\mathcal{M}$.

2 Optimization and Lower Bounds

Suppose we are interested in approximating p_θ where $\theta \in \Omega$. Mean field methods lower bound $A(\theta)$ and approximate the mean parameters of p_θ.

Proposition 1 *Any mean parameter $\mu \in \mathcal{M}^\circ$ yields*

$$A(\theta) \geq \langle \theta, \mu \rangle - A^*(\mu). \tag{5}$$

Equality holds if and only if $\mu = \mathbb{E}_\theta \left[\phi(X)\right]$.

[1] While not stated in the book, I think the dependency on G arises from $\mathcal{I}(G) = \mathcal{I}$. Contrast this with Eq. (26)

Proof. For any $\mu \in \mathcal{M}$, there exists some distribution q for which $\mathbb{E}_q[\phi(X)] = \mu$. It follows that

$$A(\theta) = \log \int_{\mathcal{X}^m} q(x) \frac{\exp\{\langle \theta, \phi(x) \rangle\}}{q(x)} \nu(dx) \tag{6}$$

$$\geq \int_{\mathcal{X}^m} q(x) \left[\langle \theta, \phi(x) \rangle - \log q(x) \right] \nu(dx) \tag{7}$$

$$= \langle \theta, \mu \rangle + H(q), \tag{8}$$

where $H(q) = -\mathbb{E}_q[\log q(X)]$. If q belongs to the exponential family $q = p_{\theta(\mu)}$, then the entropy can be re-expressed as the dual function $H(q) = -A^*(\mu)$. Equality proof presented in Chapter 3. □

Let A_F^* denote the dual function restricted to the domain $\mathcal{M}_F(G)$, carefully chosen such that A^* is tractable. The best lower bound from within $\mathcal{M}_F(G)$ is

$$\max_{\mu \in \mathcal{M}_F(G)} \{\langle \mu, \theta \rangle - A_F^*(\mu)\}. \tag{9}$$

3 Mean Field and Kullback-Leibler

We wish to quantify the divergence between our target distribution p and approximating distribution q using KL divergence

$$D(q \parallel p) = \int_{\mathcal{X}^m} \left[\log \frac{q(x)}{p(x)} \right] q(x) \nu(dx). \tag{10}$$

For two canonical parameters $\theta^1, \theta^2 \in \Omega$ and corresponding (dually coupled) mean parameters μ^1, μ^2, the primal form of the KL divergence is

$$D(\theta^1 \parallel \theta^2) = A(\theta^2) - A(\theta^1) - \langle \mu^1, \theta^2 - \theta^1 \rangle. \tag{11}$$

Recall from Chapter 3 that $\nabla A(\theta^1) = \mu^1$. KL divergence is an example of a Bregman distance, where the function of choice is the log partition function. The dual form of the KL divergence is

$$D(\mu^1 \parallel \mu^2) = A^*(\mu^1) - A^*(\mu^2) - \langle \theta^2, \mu^1 - \mu^2 \rangle. \tag{12}$$

This suggests an alternative interpretation that the KL divergence is a Bregman distance, where the function of choice is the negative entropy function. Finally, the mixed form of the KL divergence is

$$D(\mu^1 \parallel \theta^2) = A(\theta^2) + A^*(\mu^1) - \langle \mu^1, \theta^2 \rangle. \tag{13}$$

This exposes KL divergence as the gap in the variational lower bound. This suggests an alternative representation of the variational objective as

$$0 = \min_{\mu \in \mathcal{M}} \{A(\theta) + A^*(\mu) - \langle \theta, \mu \rangle\} = \min_{\mu \in \mathcal{M}} D(\mu \parallel \theta). \tag{14}$$

When $\mu \in \mathcal{M}_F(G)$ and for fixed target θ, maximizing $\langle \mu, \theta \rangle - A^*(\mu)$ is equivalent to minimizing minimizing $D(\mu \parallel \theta) = A(\theta) + A^*(\mu) - \langle \mu, \theta \rangle$. As a reminder that the feasible set is restricted to $\mathcal{M}_F(G)$, we replace future references of A^* with $A_F^*(\mu)$, which is simply A^* restricted to $\operatorname{dom} A_F^* = \mathcal{M}_F(G)$. The mean field approximation finds the best KL divergence approximation to p_θ from a tractable distribution family.

4 Naive Mean Field Algorithms

We focus on the case where the subgraph $F_0 = (V, \varnothing)$ is fully disconnected. This defines a set of distributions that are fully factorized.

4.1 Naive Mean Field for Ising Model

An Ising model is characterized by sufficient statistics $(x_s, s \in V)$ and $(x_s x_t, (s,t) \in E)$. The mean parameters are $\mu_s = \mathbb{E}[X_s], \mu_{st} = \mathbb{E}_{X_s X_t}$. For binary Ising models,

$$\mathcal{M}_{F_0}(G) = \{\mu \in \mathbb{R}^{|V|+|E|} \mid 0 \leq \mu_s \leq 1 \ \forall \ s \in V, \mu_{st} = \mu_s \mu_t \ \forall \ (s,t) \in E\}. \tag{15}$$

For $\mu \in \mathcal{M}_{F_0}(G)$, the entropy is easy to compute

$$H(\mu) = -A^*_{F_0}(\mu) = \sum_s H_s(\mu_s) = -\sum_s [\mu_s \log \mu_s + (1-\mu_s)\log(1-\mu_s)]. \tag{16}$$

The naive mean field problem is

$$A(\theta) \geq \max_{\mu \in [0,1]^m} \left\{ \sum_s \theta_s \mu_s + \sum_{(s,t)} \theta_{st} \mu_s \mu_t + \sum_s H_s(\mu_s) \right\}, \tag{17}$$

where $m = |V|$. Using coordinate ascent, the update is simply

$$\mu_s \leftarrow \sigma(\theta_s + \sum_{t \in N(s)} \theta_{st} \mu_t), \tag{18}$$

where $\sigma(z) = (1 + \exp(-z))^{-1}$ is the logistic function. Convergence to local optimum guaranteed. Interestingly, the naive mean field approximation is asymptotically exact for certain models when $m \to \infty$.

4.2 Gaussian Mean Field

A Gaussian Markov random field is characterized by mean parameters $\mu = \mathbb{E}[X] \in \mathbb{R}^m$ and $\Sigma = \mathbb{E}[XX^\top] \in \mathcal{S}^m_+$. For the disconnected subgraph F_0,

$$\mathcal{M}_{F_0}(G) = \{(\mu, \Sigma) \in \mathbb{R}^m \times \mathcal{S}^m_+ \mid \Sigma - \mu\mu^\top = \mathrm{diag}(\Sigma - \mu\mu^\top) \succeq 0\}, \tag{19}$$

where $\Sigma - \mu\mu^\top$ is the covariance matrix. For $\mu \in \mathcal{M}_{F_0}(G)$, the entropy is simply

$$-A^*_{F_0}(\mu, \Sigma) = \frac{m}{2} \log 2\pi e + \frac{1}{2} \sum_s \log(\Sigma_{ss} - \mu_s^2). \tag{20}$$

The naive mean field problem is

$$\max_{\mu, \Sigma} \left\{ \langle \theta, \mu \rangle + \frac{1}{2} \langle\!\langle \Theta, \Sigma \rangle\!\rangle + \frac{1}{2} \sum_m (\Sigma_{ss} - \mu_s^2) + \frac{m}{2} \log 2\pi e \right\} \tag{21}$$

$$\text{s.t. } \Sigma_{ss} - \mu_s^2 > 0, \Sigma_{st} = \mu_s \mu_t. \tag{22}$$

For each vertex $s \in V$, the stationary conditions are

$$\frac{1}{2(\mu_{ss} - \mu_s^2)} = -\Theta_{ss}, \frac{\mu_s}{2(\mu_{ss} - \mu_s^2)} = \theta_s + \Sigma_{t \in N(s)} \Theta_{st} \mu_t. \tag{23}$$

The condition is succinctly expressed as $-\Theta\mu = \theta$. We solve for this fixed point by iteratively applying

$$\mu_s \leftarrow -\frac{1}{\Theta_{ss}} \left\{ \theta_s + \Sigma_{t \in N(s)} \Theta_{st} \mu_t \right\}. \tag{24}$$

For specific choice of ordering, this update procedure is equivalent to the Gauss-Jacobi or Gauss-Seidel method. If converges, then it achieves the global optimum. Convergence is guaranteed when $-\Theta$ is strictly diagonally dominant (i.e. $|\Theta_{ss}| > \sum_{t \neq s} |\Theta_{st}| \ \forall \ s$).

5 Non-Convexity of Mean Field

Mean field optimization is always non-convex for exponential families with finite state space $\mathcal{X}^m$ because the inner approximation $\mathcal{M}_F(G)$ is a nonconvex set. Recall from Chapter 3 that $\mathcal{M}(G)$ is the finite convex hull

$$\mathcal{M}(G) = \text{conv}\left\{\phi(e), e \in \mathcal{X}^m\right\}. \tag{25}$$

Each extreme points of the polytope is a mean parameter $\mu_x = \phi(x)$ corresponding to distribution with its entire probability mass on some $x \in \mathcal{X}^m$. $\mathcal{M}_F(G)$ always contains the extreme points of $\mathcal{M}(G)$. If $\mathcal{M}_F(G)$ is convex, then $\mathcal{M}_F(G) = \mathcal{M}(G)$. By contraposition, if $\mathcal{M}_F(G)$ is a strict subset of $\mathcal{M}(G)$, then it cannot be a convex set.

6 Structured Mean Field

In contrast to naive mean field, which uses F_0, we now consider subgraphs with additional structure. We now discuss general procedure for performing parameter updates for an arbitrary subgraph F of the original graph G. We do not claim that these updates are best in practice.

Using $\mathcal{I}(F)$ as the subset of indices corresponding to the sufficient statistics associated with F, let $\mu(F) = (\mu_\alpha, \alpha \in \mathcal{I}(F))$. Let

$$\mathcal{M}(F;\phi) = \left\{\mu \in \mathbb{R}^d \mid \exists\ p \text{ s.t. } \mathbb{E}_p\left[\phi_\alpha(X)\right] = \mu_\alpha\ \forall\ \alpha \in \mathcal{I}(F)\right\}. \tag{26}$$

Note the difference between $\mathcal{M}(F) \in \mathbb{R}^{|\mathcal{I}(F)|}$ and $\mathcal{M}_F(G) \in \mathbb{R}^{|\mathcal{I}|}$. Note that

1. $\mu(F)$ can be an arbitrary member of $\mathcal{M}(F)$.
2. A_F^* only depends on $\mu(F)$, and not on $(\mu_\beta, \beta \in \mathcal{I}^c(F) = \mathcal{I}(G) \setminus \mathcal{I}(F))$.

More precisely, for each $\beta \in \mathcal{I}^c(F)$, there is some nonlinear function g_β for which $\mu_\beta = g_\beta(\mu(F))$. For example, in naive mean field for the binary Ising model, $\mu_{st} = g_{st}(\mu(F_0)) = \mu_s\mu_t$. We rewrite the optimization problem $\max_{\mu \in \mathcal{M}_F(G)}\{\langle\theta, \mu\rangle - A_F^*(\mu)\}$ as

$$\max_{\mu(F)\in\mathcal{M}(F)} \underbrace{\left\{\sum_{\alpha\in\mathcal{I}(F)} \theta_\alpha\mu_\alpha + \sum_{\alpha\in\mathcal{I}^c(F)} \theta_\alpha g_\alpha(\mu(F)) - A_F^*(\mu(F))\right\}}_{f(\mu(F))}. \tag{27}$$

Taking partial derivatives with respect to $\mu_\beta \in \mathcal{I}(F)$ yields

$$\frac{\partial f}{\partial \mu_\beta}(\mu(F)) = \theta_\beta + \sum_{\alpha\in\mathcal{I}^c(F)} \theta_\alpha \frac{\partial g_\alpha}{\partial \mu_\beta}(\mu(F)) - \frac{A_F^*}{\partial \mu_\beta}(\mu(F)). \tag{28}$$

Setting the derivatives to zero yields the fixed point condition

$$\nabla A_F^*(\mu(F)) = \theta + \sum_{\alpha\in\mathcal{I}^c(F)} \theta_\alpha \nabla g_\alpha(\mu(F)). \tag{29}$$

Recall from Chapter 3 that ∇A_F^* defines the forward mapping from from mean to canonical parameters coupled with $\mu(F)$. Let $\gamma(F)$ denote the canonical parameters. We can write the fixed point update as

$$\gamma_\beta(F) \leftarrow \theta_\beta + \sum_{\alpha\in\mathcal{I}^c(F)} \theta_\alpha \frac{\partial g_\alpha}{\partial \mu_\beta}(\mu(F)). \tag{30}$$

Intuitively, the second term in the update accounts for the influence that μ_β has on the implicitly determined mean parameters $(\mu_\alpha, \alpha \in \mathcal{I}^c(F))$. After each update, use junction tree to determine $\mu(F)$. Alternatively, since $\nabla A_F(\gamma(F)) = \mu(F)$ defines the backward mapping

$$\mu_\beta(F) \leftarrow \frac{\partial A_F}{\partial \gamma_\beta}\left(\theta + \sum_{\alpha\in\mathcal{I}^c(F)} \theta_\alpha \nabla g_\alpha(\mu(F))\right). \tag{31}$$

6.1 Structured Mean Field for Factorial Hidden Markov Models

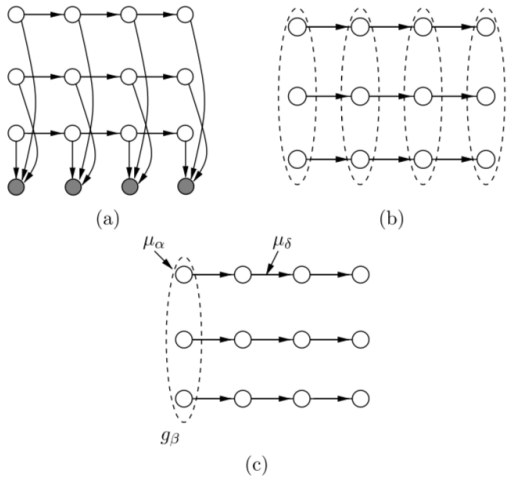

Figure 1: Structured mean field approximation for a factorial HMM. (a) Original model consists of three hidden Markov models coupled at each time step by a common observation. (b) Equivalent model where the observations are absorbed into the potential. (c) We approximate the distribution using a product distribution of the three chains.

Consider a binary hidden Markov model. Let $T = \{1, \ldots, n\}$ index the time steps and $M = \{1, 2, 3\}$ index the separate chains. Let $V = T \times M$ index the unobserved vertices. The binary factorial HMM can be described by the following exponential family with canonical parameters $(\theta_\alpha, \theta_\beta, \theta_\delta) = (\rho, \kappa, \lambda)$. Here, we use (α, β, δ) to index the subvectors corresponding to (ρ, κ, λ). Note that the unnormalized distribution is

$$p_\theta(x \mid y) \propto p_\theta(x, y) \tag{32}$$

$$\propto \exp\left\{ \sum_{(t,i)\in V} \rho_i^{(t)} x_i^{(t)} + \sum_{t\in T} \kappa^{(t)} \prod_{i\in M} x_i^{(t)} + \sum_{t\in T\setminus\{n\}} \sum_{i\in M} \lambda_i^{(t)} x_i^{(t)} x_i^{(t+1)} \right\} \tag{33}$$

For fixed θ, we wish to approximate $p_\theta(x \mid y)$ with the exponential distribution sub-family defined by the subgraph F with canonical parameters $\gamma(F)$, where $\gamma_\beta(F) = 0$. Let canonical parameters $(\gamma_\alpha(F), \gamma_\beta(F), \gamma_\delta(F))$ have corresponding (dually coupled) mean parameters noted as $(\mu_\alpha, \mu_\beta, \mu_\delta) = (\nu, \omega, \eta)$. Note that $\mathcal{I}(F) = \{\alpha, \delta\}$ and $\mathcal{I}^c(F) = \{\beta\}$. The subgraph F enforces that the mean parameter μ_β is a function of μ_α

$$\omega^{(t)} = g^{(t)}(\mu(F)) = \nu_1^{(t)} \nu_2^{(t)} \nu_3^{(t)}. \tag{34}$$

Note that $\nabla_{\mu_\delta} g^{(t)} = 0$ for all $t \in T$. The update rule Eq. (30) suggests that the canonical parameters for $\gamma_\delta(F)$ should always be set to $\gamma_\delta(F) = \theta_\delta$.