Graphical Models, ExpFam, Variational Inference
Chapter 5: Mean Field Methods

20 April 2015

Wittawat Jitkrittum

Gatsby Machine Learning Journal Club

Exponential Family (review)

$$p(x|\theta) = \exp\left(\langle \theta, \phi(x) \rangle - A(\theta)\right)$$
$$A(\theta) = \log \int \exp(\langle \theta, \phi(x) \rangle)\, \mathrm{d}x$$

- Variational principle

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

- Marginal polytope (feasible mean parameters)

$$\mathcal{M} = \left\{ \mu \in \mathbb{R}^d \mid \exists q \text{ s.t. } \mathbb{E}_q[\phi(X)] = \mu \right\}$$

- Negative entropy: $A^*(\mu) = -H(p)$.

Exponential Family (review)

Variational representation (from chapter 3):

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

Legendre duality:

$$\nabla A^*(\mu) = \theta,$$
$$\nabla A(\theta) = \mu,$$

for dually coupled (θ, μ) i.e., $\mu = \mathbb{E}_\theta[\phi(x)]$.

BP, EP and Mean Field Methods

Variational principle:

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

- $\mathcal{M}$ characterized by exponentially many half-space constraints.
- BP and EP approximates $A(\theta)$ by relaxing $\mathcal{M}$ and $A^*(\mu)$.
- BP relaxes $\mathcal{M}$ to $\mathbb{L}(G)$ (locally consistent distributions).
- A^* relaxed to A^*_{Bethe} (only pairwise interaction).

Mean field:

- Also approximate the variational principle.
- Consider subset of distributions for which $\mathcal{M}$ and A^* are easy to characterize e.g., tractable distributions.
- Simplest choice = product distributions. Give naive mean field.

BP, EP and Mean Field Methods

Variational principle:

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

- $\mathcal{M}$ characterized by exponentially many half-space constraints.
- BP and EP approximates $A(\theta)$ by relaxing $\mathcal{M}$ and $A^*(\mu)$.
- BP relaxes $\mathcal{M}$ to $\mathbb{L}(G)$ (locally consistent distributions).
- A^* relaxed to A^*_{Bethe} (only pairwise interaction).

Mean field:

- Also approximate the variational principle.
- Consider subset of distributions for which $\mathcal{M}$ and A^* are easy to characterize e.g., tractable distributions.
- Simplest choice = product distributions. Give naive mean field.

5.1 Tractable Families (p. 128)

- ExpFam with sufficient statistics $\phi = (\phi_\alpha, \alpha \in \mathcal{I})$ on cliques of $G = (V, E)$.
- Consider a subgraph $F = (V_F, E_F) \subseteq G$ i.e., $V_F \subseteq V$ and $E_F \subseteq E$.
- $\mathcal{I}(F) \subseteq \mathcal{I}$: the subset of sufficient statistics associated with F.
- {Distributions following F} = sub-family with subspace of canonical parameters

$$\Omega(F) := \{\theta \in \Omega \mid \theta_\alpha = 0,\ \forall \alpha \in \mathcal{I} \backslash \mathcal{I}(F)\}.$$

Marginal polytope:

$$\mathcal{M}_F(G) := \left\{\mu \in \mathbb{R}^d \mid \mu = \mathbb{E}_\theta[\phi(x)],\ \text{for some } \theta \in \Omega(F)\right\}.$$

- $\mathcal{M}_F$ is an inner approximation to $\mathcal{M}$, unlike $\mathbb{L}(G)$ in BP.

Example 5.1: Tractable Subgraphs

- Ising model with $G = (V, E)$. $X_s \in \{0, 1\}$.

$$p_\theta(x) \propto \exp\left(\sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{st} x_s x_t\right),$$

$$\phi(x) = (x_s, s \in V;\ x_s x_t, (s,t) \in E) \in \{0,1\}^{|V|+|E|}.$$

- Consider $F_0 = (V, \emptyset)$ (completely disconnected subgraph).
- Permissible parameters:

$$\Omega(F_0) = \{\theta \in \Omega \mid \theta_{st} = 0,\ \forall (s,t) \in E\}.$$

- Densities in the sub-family fully factorized:

$$p_\theta(x) = \prod_{s \in V} p(x_s | \theta_s) \propto \exp\left(\sum_{s \in V} \theta_s x_s\right)$$

5.2.1 Generic Mean Field Procedure

Given θ, the mean field solves

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

where A_F^* is A^* restricted to $\mathcal{M}_F(G)$.

Properties of mean field:

1. $A(\theta) \geq A_F(\theta)$ because

$$\begin{aligned} A(\theta) &= \sup_{\mu \in \mathcal{M}} \langle \mu, \theta \rangle - A^*(\mu) \text{ (variational principle)} \\ &\geq \sup_{\mu \in \mathcal{M}_F} \langle \mu, \theta \rangle - A^*(\mu) \text{ (mean field)} \end{aligned}$$

   because $\mathcal{M}_F \subset \mathcal{M}$.

2. Approximate μ with the best match in $\mathcal{M}_F$ in the KL sense.

5.2.1 Generic Mean Field Procedure

Given θ, the mean field solves

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

where A_F^* is A^* restricted to $\mathcal{M}_F(G)$.

Properties of mean field:

1. $A(\theta) \geq A_F(\theta)$ because

$$\begin{aligned} A(\theta) &= \sup_{\mu \in \mathcal{M}} \langle \mu, \theta \rangle - A^*(\mu) \ \text{(variational principle)} \\ &\geq \sup_{\mu \in \mathcal{M}_F} \langle \mu, \theta \rangle - A^*(\mu) \ \text{(mean field)} \end{aligned}$$

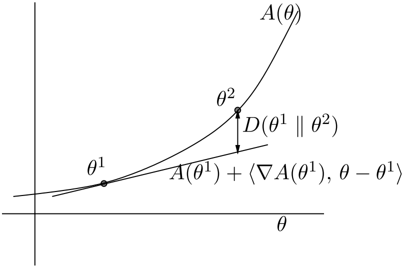
   because $\mathcal{M}_F \subset \mathcal{M}$.

2. Approximate μ with the best match in $\mathcal{M}_F$ in the KL sense.

KL on Exponential Family Distributions

- Consider $p_{\theta^1}, p_{\theta^2} \in \mathrm{ExpFam}$ where $p_\theta(x) = \exp\left(\langle \theta, \phi(x) \rangle - A(\theta)\right).$

$$
\begin{aligned}
\mathrm{D_{KL}}(\theta^1 \| \theta^2) &= \mathbb{E}_{\theta^1}\left[\log \frac{p_{\theta^1}(x)}{p_{\theta^2}(x)}\right] = \mathbb{E}_{\theta^1}\left[\log p_{\theta^1}(x) - \log p_{\theta^2}(x)\right] \\
&= \mathbb{E}_{\theta^1}\left[\left\langle \theta^1, \phi(x) \right\rangle - A(\theta^1) - \left\langle \theta^2, \phi(x) \right\rangle + A(\theta^2)\right] \\
&= A(\theta^2) - A(\theta^1) - \left\langle \mu^1, \theta^2 - \theta^1 \right\rangle .
\end{aligned}
$$



- $\nabla A(\theta^1) = \mu^1 = \mathbb{E}_{\theta^1}[\phi(x)]$
- An instance of Bregman divergence with the convex function $A(\theta)$.

5.2.2 Mean Field and KL Divergence

- Let (θ, μ) be a **dual couple** i.e., $\mu = \mathbb{E}_\theta[\phi(x)]$.
- Given θ', mean field approximates its couple μ' by

$$\begin{aligned}
\mu' &\approx \arg \sup_{\mu \in \mathcal{M}_F(G)} \langle \mu, \theta' \rangle - A^*(\mu) \\
&\overset{(a)}{=} \arg \sup_{\mu \in \mathcal{M}_F(G)} \langle \mu, \theta' \rangle - (\langle \mu, \theta \rangle - A(\theta)) \\
&= \arg \sup_{\mu \in \mathcal{M}_F(G)} A(\theta) + \langle \mu, \theta' - \theta \rangle \\
&\overset{(b)}{=} \arg \inf_{\mu \in \mathcal{M}_F(G)} A(\theta') - A(\theta) - \langle \mu, \theta' - \theta \rangle \\
&= \arg \inf_{\mu \in \mathcal{M}_F(G)} \mathrm{D}_{\mathrm{KL}}(\theta \| \theta').
\end{aligned}$$

  - (a): $A^*(\mu) = \langle \mu, \theta \rangle - A(\theta)$ by variational principle.
  - (b): Negate. Then add $A(\theta')$, a constant.
- **Mean field**: Approximate $p_{\theta'}$ with a distribution in $\mathcal{M}_F(G)$. Quality measured by KL.

Example 5.2 Naive Mean Field for Ising Model (p. 134) I

- **Naive mean field:** $p_\theta(x_{1:m}) := \prod_{s \in V} p(x_s; \theta_s)$.
- Ising model:
  - Sufficient statistics: $(x_s,\ s \in V)$ and $(x_s x_t,\ (s,t) \in E)$. Binary x_s.
  - Mean parameters: $\mu_s = \mathbb{E}[X_s] = P[X_s = 1]$ and $\mu_{st} = \mathbb{E}[X_s X_t]$.
- $F_0 :=$ fully disconnected graph.

$$\mathcal{M}_{F_0}(G) := \{\mu \in \mathbb{R}^{|V|+|E|} \mid \mu_{st} = \mu_s \mu_t,\ 0 \le \mu_s \le 1 \text{ for all } s, t\}$$

- Dual function: $A^*_{F_0}(\mu) = -\sum_{s \in V} H_s(\mu_s)$.

Example 5.2 Naive Mean Field for Ising Model (p. 134) II

- Variational problem:

$$A(\theta) \geq \max_{\{\mu_i \in [0,1]\}_i} \left\{ \sum_{s \in V} \theta_s \mu_s + \sum_{(s,t) \in E} \theta_{st} \mu_s \mu_t + \sum_{s \in V} H_s(\mu_s) \right\},$$

  strictly concave w.r.t. μ_s when $\{\mu_t\}_{t \neq s}$ are fixed.
- Equate the derivative to 0:

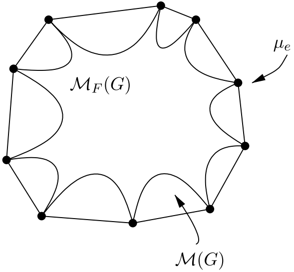
$$\mu_s \leftarrow \sigma\left(\theta_s + \sum_{t \in N(s)} \theta_{st} \mu_t\right), \qquad (5.17)$$

  where $\sigma(\cdot)$ is the logistic function.
- Coordinate ascent with unique max for every update.
- Guaranteed to converge.
- Not jointly concave in $\{\mu_t\}_t$. Sensitive to initialization.

Claim (Nonconvexity of Mean Field)

If the domain $\mathcal{X}^m$ is finite, and $\mathcal{M}_F(G) \subsetneq \mathcal{M}(G)$, then $\mathcal{M}_F(G)$ is not a convex set.

- Assume $\mathcal{X}^m$ is finite, and $\mathcal{M}_F(G) \subsetneq \mathcal{M}(G)$.
- Assume $\mathcal{M}_F(G)$ is convex.
- $\mathcal{M}_F(G)$ contains all the extreme points $\mu_x = \phi(x)$ of $\mathcal{M}(G)$ i.e., point mass distributions.
- Since $\mathcal{M}_F(G)$ is convex, it must contain $\mathrm{conv}\{\phi(x),\, x \in \mathcal{X}^m\}$ which is $\mathcal{M}(G)$.
- $\mathcal{M}_F(G) \supset \mathcal{M}(G)$ is a contradiction.



5.5 Structured Mean Field (p. 142)

- Tractable distributions based on an arbitrary subgraph F.
- $\mathcal{I}(F) :=$ subset of indices of suff. stats. associated with F.
- $\mu(F) := (\mu_\alpha,\ \alpha \in \mathcal{I}(F))$, subvector of μ.
- $\mathcal{M}(F) :=$ set of realizable means defined by F.

Observation:

- A_F^* depends only on $\mu(F)$, and not on μ_α for $\alpha \in \mathcal{I}(G)\backslash\mathcal{I}(F)$.
  - In Ising model, naive MF does not depend on μ_{st}.
  - μ_s, μ_t determines μ_{st}. $\alpha = (s,t)$.
- For each $\alpha \in \mathcal{I}(G)\backslash\mathcal{I}(F)$,

$$\mu_\alpha = g_\alpha(\mu(F))$$

  for some nonlinear g_α.
- Ex: $\mu_{st} = \mu_s\mu_t = g_{st}(\mu_1, \ldots, \mu_m)$ in naive MF on Ising model.

5.5 Structured Mean Field (p. 142)

- Tractable distributions based on an arbitrary subgraph F.
- $\mathcal{I}(F) :=$ subset of indices of suff. stats. associated with F.
- $\mu(F) := (\mu_\alpha,\ \alpha \in \mathcal{I}(F))$, subvector of μ.
- $\mathcal{M}(F) :=$ set of realizable means defined by F.

Observation:

- A_F^* depends only on $\mu(F)$, and not on μ_α for $\alpha \in \mathcal{I}(G) \backslash \mathcal{I}(F)$.
  - In Ising model, naive MF does not depend on μ_{st}.
  - μ_s, μ_t determines μ_{st}. $\alpha = (s, t)$.
- For each $\alpha \in \mathcal{I}(G) \backslash \mathcal{I}(F)$,

$$\mu_\alpha = g_\alpha(\mu(F))$$

  for some nonlinear g_α.
- Ex: $\mu_{st} = \mu_s \mu_t = g_{st}(\mu_1, \ldots, \mu_m)$ in naive MF on Ising model.

Nonlinear Constraints on Mean Parameters (p. 143) I

- MF variational problem:

$$\max_{\mu(F)\in\mathcal{M}(F)} \sum_{\beta\in\mathcal{I}(F)} \theta_\beta\mu_\beta + \sum_{\alpha\in\mathcal{I}(G)\setminus\mathcal{I}(F)} \theta_\alpha g_\alpha(\mu(F)) - A_F^*(\mu(F))$$

$$:= \max_{\mu(F)\in\mathcal{M}(F)} f(\mu(F))$$

(recall θ_β is param. of the original distribution)

- Derivative for $\beta \in \mathcal{I}(F)$:

$$\frac{\partial f}{\partial \mu_\beta}(\mu(F)) = \theta_\beta + \sum_{\alpha\in\mathcal{I}(G)\setminus\mathcal{I}(F)} \theta_\alpha \frac{\partial g_\alpha}{\partial \mu_\beta}(\mu(F)) - \underbrace{\frac{\partial A_F^*}{\partial \mu_\beta}(\mu(F))}_{:=\gamma_\beta(F)}$$

where $(\gamma_\beta, \mu_\beta)$ is a dual couple.

Nonlinear Constraints on Mean Parameters (p. 143) II

- $\frac{\partial f}{\partial \mu_\beta}(\mu(F)) = 0$ and rearranging:

$$\gamma_\beta(F) \leftarrow \theta_\beta + \sum_{\alpha \in \mathcal{I}(G) \setminus \mathcal{I}(F)} \theta_\alpha \frac{\partial g_\alpha}{\partial \mu_\beta}(\mu(F)). \quad (5.27)$$

- Need to adjust all mean parameters that depend on γ_β e.g., via junction tree updates.

MF Updates in Terms of $\mu(F)$ (p. 144)

- By exploiting duality of (A_F, A_F^*),

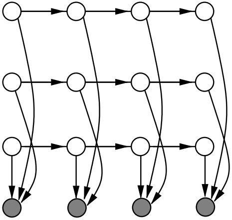
$$\gamma_\beta(F) \leftarrow \theta_\beta + \sum_{\alpha \in \mathcal{I}(G) \setminus \mathcal{I}(F)} \theta_\alpha \frac{\partial g_\alpha}{\partial \mu_\beta}(\mu(F)). \quad (5.27)$$

  becomes

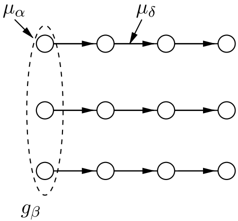
$$\mu_\beta(F) \leftarrow \frac{\partial A_F}{\partial \gamma_\beta} \left(\overbrace{\theta + \sum_{\alpha \in \mathcal{I}(G) \setminus \mathcal{I}(F)} \theta_\alpha \frac{\partial g_\alpha}{\partial \mu(F)}(\mu(F))}^{=\gamma=(5.27)=\text{dual couple of } \mu} \right) \quad (5.28)$$

  which involves only the mean parameters $\mu(F)$.
- With (5.28), we get Ising model naive MF updates when $g_{st}(\mu_1, \ldots, \mu_m) = \mu_s \mu_t$. See example 5.5.

Example 5.6 Structured MF for Factorial HMMs (p. 146)



(a)



(c)

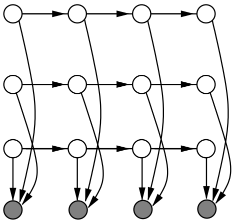
- M latent chains independent a priori (a).
- Common observations induces a coupling (by graph moralization).
- Approximation: decoupling M chains.
- M latent variables coupled at each time (c).
- Assume binary latent. $g_{stu}(\mu) = \mu_s \mu_t \mu_u$. $\beta = (s, t, u)$.
- g_β does not depend on μ_δ. $\frac{\partial g_\beta}{\partial \mu_\delta} = 0$.

$$\gamma_\delta(F) \leftarrow \theta_\delta + \sum_{\beta \in \mathcal{I}(G) \setminus \mathcal{I}(F)} \theta_\beta \frac{\partial g_\beta}{\partial \mu_\delta}(\mu(F)). \quad (5.27)$$

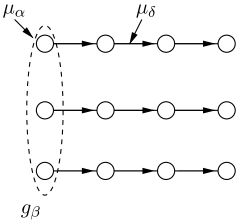
$\gamma_\delta = \theta_\delta$ meaning edge potentials θ_δ from the original distribution remains unchanged.

- Make sense from the approximation choice.

Example 5.6 Structured MF for Factorial HMMs (p. 146)



(a)



(c)

- M latent chains independent a priori (a).
- Common observations induces a coupling (by graph moralization).
- Approximation: decoupling M chains.
- M latent variables coupled at each time (c).
- Assume binary latent. $g_{stu}(\mu) = \mu_s \mu_t \mu_u$. $\beta = (s, t, u)$.
- g_β does not depend on μ_δ. $\frac{\partial g_\beta}{\partial \mu_\delta} = 0$.

$$\gamma_\delta(F) \leftarrow \theta_\delta + \sum_{\beta \in \mathcal{I}(G) \backslash \mathcal{I}(F)} \theta_\beta \frac{\partial g_\beta}{\partial \mu_\delta}(\mu(F)). \quad (5.27)$$

$\gamma_\delta = \theta_\delta$ meaning edge potentials θ_δ from the original distribution remains unchanged.

- Make sense from the approximation choice.

Summary of Mean Field

Inner approximation $\mathcal{M}_F$ to $\mathcal{M}$ in the variational principle:

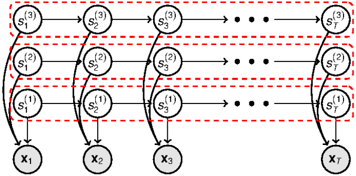
$$A(\theta) = \sup_{\mu \in \mathcal{M}} \langle \mu, \theta \rangle - A^*(\mu).$$

- Equivalently, approximate μ with the best match in $\mathcal{M}_F$ in the KL sense.
- Generally nonconvex.
- Fast updates for naive mean field.
- Structured mean field preserves more interaction with higher computational cost.

Factorial HMM Updates

(Stolen from Maneesh's ML course).

Stuctured FHMM



For the FHMM we can factor the chains:

$$q(s_{1:T}^{1:M}) = \prod_m q^m(s_{1:T}^m)$$

$$q^m(s_{1:T}^m) \propto \exp \left\langle \log P(\mathbf{s}_{1:T}^{1:M}, \mathbf{x}_{1:T}) \right\rangle_{\prod_{\neg m} q^{m'}(s_{1:T}^{m'})}$$

$$= \exp \left\langle \sum_\mu \sum_t \log P(s_t^\mu | s_{t-1}^\mu) + \sum_t \log P(\mathbf{x}_t | s_t^{1:M}) \right\rangle_{\prod_{\neg m} q^{m'}}$$

$$\propto \exp \left[\sum_t \log P(s_t^m | s_{t-1}^m) + \sum_t \left\langle \log P(\mathbf{x}_t | s_t^{1:M}) \right\rangle_{\prod_{\neg m} q^{m'}(s_t^{m'})} \right]$$

$$= \prod_t P(s_t^m | s_{t-1}^m) \quad \prod_t e^{\left\langle \log P(\mathbf{x}_t | s_t^{1:M}) \right\rangle_{\prod_{\neg m} q^{m'}(s_t^{m'})}}$$

This looks like a standard HMM joint, with a modified likelihood term ⇒ cycle through multiple forward-backward passes, updating likelihood terms each time.

References I

- Chapter 5. Wainwright & Jordan technical report.
- https://www.eecs.berkeley.edu/~wainwrig/Papers/WaiJor08_FTML.p