Variational Inference and Mean Field

Mark Schmidt

University of British Columbia

August, 2015

Summary of Weeks 1 and 2

- We used structured prediction to motivate studying UGMs:

Input: 

Output: "Paris"

Summary of Weeks 1 and 2

- We used structured prediction to motivate studying UGMs:

Input: 

Output: "Paris"

- Week 1: exact inference:
  - Exact decoding, inference, and sampling.
  - Smal graphsl, tree, junction trees, semi-Markov, graph cuts.

Summary of Weeks 1 and 2

- We used structured prediction to motivate studying UGMs:

Input: 

Output: "Paris"

- Week 1: exact inference:
  - Exact decoding, inference, and sampling.
  - Smal graphsl, tree, junction trees, semi-Markov, graph cuts.
- Week 2: learning and approximate inference:
  - Learning based on maximum likelihood.
  - Approximate decoding with local search.
  - Approximate sampling with MCMC.
  - Hidden variables.
  - Structure learning.

Summary of Weeks 1 and 2

- We used structured prediction to motivate studying UGMs:

Input: 

Output: "Paris"

- Week 1: exact inference:
  - Exact decoding, inference, and sampling.
  - Smal graphsl, tree, junction trees, semi-Markov, graph cuts.
- Week 2: learning and approximate inference:
  - Learning based on maximum likelihood.
  - Approximate decoding with local search.
  - Approximate sampling with MCMC.
  - Hidden variables.
  - Structure learning.
- Week 3:
  - Approximate inference with variational methods.
  - Approximate decoding with convex relaxations.
  - Learning based on structured SVMs.

Variational Inference

- "Variational inference":
  - Formulate inference problem as constrained optimization.
  - Approximate the function or constraints to make it easy.

Variational Inference

- "Variational inference":
  - Formulate inference problem as constrained optimization.
  - Approximate the function or constraints to make it easy.
- Why not use MCMC?
  - MCMC works asymptotically, but may take forever.
  - Variational methods not consistent, but very fast.

(trade off accuracy vs. computation)

Overview of Methods

- "Classic" variational inference based on intuition:
  - Mean-field: approximate log-marginal i by averaging neighbours,

$$\mu_{is}^{k+1} \propto \phi_i(s) \exp\left(\sum_{(i,j)\in E} \sum_t \mu_{jt}^k \log(\phi_{ij}(s,t))\right),$$

    comes from statistical physics.

Overview of Methods

- "Classic" variational inference based on intuition:
  - Mean-field: approximate log-marginal i by averaging neighbours,

$$\mu_{is}^{k+1} \propto \phi_i(s) \exp\left(\sum_{(i,j)\in E} \sum_t \mu_{jt}^k \log(\phi_{ij}(s,t)) \right),$$

    comes from statistical physics.
  - Loopy belief propagation: apply tree-based message passing algorithm to loopy graphs.

Overview of Methods

- "Classic" variational inference based on intuition:
  - Mean-field: approximate log-marginal i by averaging neighbours,

$$\mu_{is}^{k+1} \propto \phi_i(s) \exp\left(\sum_{(i,j)\in E} \sum_t \mu_{jt}^k \log(\phi_{ij}(s,t))\right),$$

    comes from statistical physics.
  - Loopy belief propagation: apply tree-based message passing algorithm to loopy graphs.
  - Linear programming relaxation: replace integer constraints with linear constraints.

Overview of Methods

- "Classic" variational inference based on intuition:
  - Mean-field: approximate log-marginal i by averaging neighbours,

$$\mu_{is}^{k+1} \propto \phi_i(s) \exp\left(\sum_{(i,j)\in E} \sum_t \mu_{jt}^k \log(\phi_{ij}(s,t))\right),$$

    comes from statistical physics.
  - Loopy belief propagation: apply tree-based message passing algorithm to loopy graphs.
  - Linear programming relaxation: replace integer constraints with linear constraints.
- But we are developing theoretical tools to understand these:
  - Has lead to new methods with better properties.

Overview of Methods

- "Classic" variational inference based on intuition:
  - Mean-field: approximate log-marginal i by averaging neighbours,

$$\mu_{is}^{k+1} \propto \phi_i(s) \exp\left(\sum_{(i,j)\in E} \sum_t \mu_{jt}^k \log(\phi_{ij}(s,t))\right),$$

    comes from statistical physics.
  - Loopy belief propagation: apply tree-based message passing algorithm to loopy graphs.
  - Linear programming relaxation: replace integer constraints with linear constraints.
- But we are developing theoretical tools to understand these:
  - Has lead to new methods with better properties.
- This week will follow the variational inference monster paper:

Wainwright & Jordan. **Graphical Models, Exponential Families, and Variational Inference**. Foundations and Trends in Machine Learning. 1(1-2), 2008.

Exponential Families and Cumulant Function

- We will again consider log-linear models:

$$P(X) = \frac{\exp(w^T F(X))}{Z(w)},$$

but view them as exponential family distributions,

$$P(X) = \exp(w^T F(X) - A(w)),$$

where $A(w) = \log(Z(w))$.

Exponential Families and Cumulant Function

- We will again consider log-linear models:

$$P(X) = \frac{\exp(w^T F(X))}{Z(w)},$$

  but view them as exponential family distributions,

$$P(X) = \exp(w^T F(X) - A(w)),$$

  where $A(w) = \log(Z(w))$.

- Log-partition $A(w)$ is called the cumulant function,

$$\nabla A(w) = \mathbb{E}[F(X)], \quad \nabla^2 A(w) = \mathbb{V}[F(X)],$$

  which implies convexity.

Convex Conjugate and Entropy

- The convex conjugate of a function A is given by

$$A^*(\mu) = \sup_{w \in \mathcal{W}} \{\mu^T w - A(w)\}.$$

Convex Conjugate and Entropy

- The convex conjugate of a function A is given by

$$A^*(\mu) = \sup_{w \in \mathcal{W}} \{\mu^T w - A(w)\}.$$

- E.g., in CPSC 540 we did this for logistic regression:

$$A(w) = \log(1 + \exp(w)),$$

Convex Conjugate and Entropy

- The convex conjugate of a function A is given by

$$A^*(\mu) = \sup_{w \in \mathcal{W}} \{\mu^T w - A(w)\}.$$

- E.g., in CPSC 540 we did this for logistic regression:

$$A(w) = \log(1 + \exp(w)),$$

implies that $A^*(\mu)$ satisfies $w = \log(\mu)/\log(1-\mu)$.

Convex Conjugate and Entropy

- The convex conjugate of a function A is given by

$$A^*(\mu) = \sup_{w \in \mathcal{W}} \{\mu^T w - A(w)\}.$$

- E.g., in CPSC 540 we did this for logistic regression:

$$A(w) = \log(1 + \exp(w)),$$

implies that $A^*(\mu)$ satisfies $w = \log(\mu)/\log(1-\mu)$.

  - When $0 < \mu < 1$ we have

$$\begin{aligned} A^*(\mu) &= \mu \log(\mu) + (1-\mu)\log(1-\mu) \\ &= -H(p_\mu), \end{aligned}$$

  negative entropy of binary distribution with mean μ.

  - If μ does not satisfy boundary constraint, $A^*(\mu) = \infty$.

Convex Conjugate and Entropy

- More generally, if $A(w) = \log(Z(w))$ then

$$A^*(\mu) = -H(p_\mu),$$

subject to boundary constraints on μ and constraint:

$$\mu = \nabla A(w) = \mathbb{E}[F(X)].$$

- Convex set satisfying these is called marginal polytope $\mathcal{M}$.

Convex Conjugate and Entropy

- More generally, if $A(w) = \log(Z(w))$ then

$$A^*(\mu) = -H(p_\mu),$$

subject to boundary constraints on μ and constraint:

$$\mu = \nabla A(w) = \mathbb{E}[F(X)].$$

- Convex set satisfying these is called marginal polytope $\mathcal{M}$.
- If A is convex (and LSC), $A^{**} = A$. So we have

$$A(w) = \sup_{\mu \in \mathcal{U}} \{w^T \mu - A^*(\mu)\}.$$

Convex Conjugate and Entropy

- More generally, if $A(w) = \log(Z(w))$ then

$$A^*(\mu) = -H(p_\mu),$$

subject to boundary constraints on μ and constraint:

$$\mu = \nabla A(w) = \mathbb{E}[F(X)].$$

- Convex set satisfying these is called marginal polytope $\mathcal{M}$.
- If A is convex (and LSC), $A^{**} = A$. So we have

$$A(w) = \sup_{\mu \in \mathcal{U}} \{w^T\mu - A^*(\mu)\}.$$

and when $A(w) = \log(Z(w))$ we have

$$\log(Z(w)) = \sup_{\mu \in \mathcal{M}} \{w^T\mu + H(p_\mu)\}.$$

- We've written inference as a convex optimization problem.

Detour: Maximum Likelihood and Maximum Entropy

- The maximum likelihood parameters w satisfy:

$$
\begin{aligned}
&\min_{w \in \mathbb{R}^d} -w^T F(D) + \log(Z(w)) \\
&= \min_{w \in \mathbb{R}^d} -w^T F(D) + \sup_{\mu \in \mathcal{M}} \{w^T \mu + H(p_\mu)\} \quad \text{(convex conjugate)} \\
&= \min_{w \in \mathbb{R}^d} \sup_{\mu \in \mathcal{M}} \{-w^T F(D) + w^T \mu + H(p_\mu)\} \\
&= \sup_{\mu \in \mathcal{M}} \{ \min_{w \in \mathbb{R}^d} -w^T F(D) + w^T \mu + H(p_\mu)\} \quad \text{(convex/concave)}
\end{aligned}
$$

Detour: Maximum Likelihood and Maximum Entropy

- The maximum likelihood parameters w satisfy:

$$\begin{aligned}
&\min_{w \in \mathbb{R}^d} -w^T F(D) + \log(Z(w)) \\
&= \min_{w \in \mathbb{R}^d} -w^T F(D) + \sup_{\mu \in \mathcal{M}} \{w^T \mu + H(p_\mu)\} \quad \text{(convex conjugate)} \\
&= \min_{w \in \mathbb{R}^d} \sup_{\mu \in \mathcal{M}} \{-w^T F(D) + w^T \mu + H(p_\mu)\} \\
&= \sup_{\mu \in \mathcal{M}} \{ \min_{w \in \mathbb{R}^d} -w^T F(D) + w^T \mu + H(p_\mu)\} \quad \text{(convex/concave)}
\end{aligned}$$

which is $-\infty$ unless $F(D) = \mu$ (e.g., Max Likelihood), so we have

Detour: Maximum Likelihood and Maximum Entropy

- The maximum likelihood parameters w satisfy:

$$\begin{aligned}
&\min_{w \in \mathbb{R}^d} -w^T F(D) + \log(Z(w)) \\
&= \min_{w \in \mathbb{R}^d} -w^T F(D) + \sup_{\mu \in \mathcal{M}} \{w^T \mu + H(p_\mu)\} \quad \text{(convex conjugate)} \\
&= \min_{w \in \mathbb{R}^d} \sup_{\mu \in \mathcal{M}} \{-w^T F(D) + w^T \mu + H(p_\mu)\} \\
&= \sup_{\mu \in \mathcal{M}} \{\min_{w \in \mathbb{R}^d} -w^T F(D) + w^T \mu + H(p_\mu)\} \quad \text{(convex/concave)}
\end{aligned}$$

which is $-\infty$ unless $F(D) = \mu$ (e.g., Max Likelihood), so we have

$$\begin{aligned}
&\min_{w \in \mathbb{R}^d} -w^T F(D) + \log(Z(w)) \\
&= \max_{\mu \in \mathcal{M}} H(p_\mu),
\end{aligned}$$

subject to $F(D) = \mu$.

Detour: Maximum Likelihood and Maximum Entropy

- The maximum likelihood parameters w satisfy:

$$
\begin{aligned}
&\min_{w \in \mathbb{R}^d} -w^T F(D) + \log(Z(w)) \\
&= \min_{w \in \mathbb{R}^d} -w^T F(D) + \sup_{\mu \in \mathcal{M}} \{w^T \mu + H(p_\mu)\} \quad \text{(convex conjugate)} \\
&= \min_{w \in \mathbb{R}^d} \sup_{\mu \in \mathcal{M}} \{-w^T F(D) + w^T \mu + H(p_\mu)\} \\
&= \sup_{\mu \in \mathcal{M}} \{ \min_{w \in \mathbb{R}^d} -w^T F(D) + w^T \mu + H(p_\mu)\} \qquad \text{(convex/concave)}
\end{aligned}
$$

  which is $-\infty$ unless $F(D) = \mu$ (e.g., Max Likelihood), so we have

$$
\begin{aligned}
&\min_{w \in \mathbb{R}^d} -w^T F(D) + \log(Z(w)) \\
&= \max_{\mu \in \mathcal{M}} H(p_\mu),
\end{aligned}
$$

  subject to $F(D) = \mu$.

- Maximum likelihood $\Rightarrow$ maximum entropy + moment constraints.
- Converse: MaxEnt + fit feature frequencies $\Rightarrow$ ML(log-linear).

Difficulty of Variational Formulation

- We wrote inference as a convex optimization:

$$\log(Z)) = \sup_{\mu \in \mathcal{M}} \{w^T\mu + H(p_\mu)\},$$

Difficulty of Variational Formulation

- We wrote inference as a convex optimization:

$$\log(Z)) = \sup_{\mu \in \mathcal{M}} \{w^T \mu + H(p_\mu)\},$$

- Did this make anything easier?

Difficulty of Variational Formulation

- We wrote inference as a convex optimization:

$$\log(Z)) = \sup_{\mu \in \mathcal{M}} \{w^T \mu + H(p_\mu)\},$$

- Did this make anything easier?
  - Computing entropy $H(p_\mu)$ seems as hard as inference.
  - Characterizing marginal polytope $\mathcal{M}$ becomes hard with loops.

Difficulty of Variational Formulation

- We wrote inference as a convex optimization:

$$\log(Z)) = \sup_{\mu \in \mathcal{M}} \{w^T \mu + H(p_\mu)\},$$

- Did this make anything easier?
  - Computing entropy $H(p_\mu)$ seems as hard as inference.
  - Characterizing marginal polytope $\mathcal{M}$ becomes hard with loops.
- Practical variational methods:
  - Work with approximation to marginal polytope $\mathcal{M}$.
  - Work with approximation/bound on entropy A^*.

Difficulty of Variational Formulation

- We wrote inference as a convex optimization:

$$\log(Z)) = \sup_{\mu \in \mathcal{M}} \{w^T \mu + H(p_\mu)\},$$

- Did this make anything easier?
  - Computing entropy $H(p_\mu)$ seems as hard as inference.
  - Characterizing marginal polytope $\mathcal{M}$ becomes hard with loops.
- Practical variational methods:
  - Work with approximation to marginal polytope $\mathcal{M}$.
  - Work with approximation/bound on entropy A^*.
- Comment on notation when discussing inference with fixed "w":
  - Put everything "inside" w to discuss general log-potentials:

$$\log(Z) = \sup_{\mu \in \mathcal{M}} \left\{ \sum_i \sum_s \mu_{i,s} \log \phi_i(s) + \sum_{(i,j) \in E} \sum_{s,t} \mu_{ij,st} \log \phi_{ij}(s,t) - \sum_X p_u(X) \log(p_u(X)) \right\},$$

and we have all μ values even with parameter tieing.

Mean Field Approximation

- Mean field approximation assumes

  $$\mu_{ij,st} = \mu_{i,s}\mu_{j,t},$$

  for all edges, which means

  $$p(x_i = s, x_j = t) = p(x_i = s)p(x_j = t),$$

  and that variables are independent.

Mean Field Approximation

- Mean field approximation assumes

$$\mu_{ij,st} = \mu_{i,s}\mu_{j,t},$$

for all edges, which means

$$p(x_i = s, x_j = t) = p(x_i = s)p(x_j = t),$$

and that variables are independent.

- Entropy is simple under mean field approximation:

$$\sum_X p(X) \log p(X) = \sum_i \sum_{x_i} p(x_i) \log p(x_i).$$

Mean Field Approximation

- Mean field approximation assumes

$$\mu_{ij,st} = \mu_{i,s}\mu_{j,t},$$

for all edges, which means

$$p(x_i = s, x_j = t) = p(x_i = s)p(x_j = t),$$

and that variables are independent.

- Entropy is simple under mean field approximation:

$$\sum_X p(X)\log p(X) = \sum_i \sum_{x_i} p(x_i)\log p(x_i).$$

- Marginal polytope is also simple:

$$\mathcal{M}_F = \{\mu \mid \mu_{i,s} \geq 0,\ \sum_s \mu_{i,s} = 1,\ \mu_{ij,st} = \mu_{i,s}\mu_{j,t}\}.$$

Entropy of Mean Field Approximation

- Entropy form is from distributive law and probabilities sum to 1:

$$
\begin{aligned}
\sum_X p(X)\log p(X) &= \sum_X p(X)\log(\prod_i p(x_i)) \\
&= \sum_X p(X)\sum_i \log(p(x_i)) \\
&= \sum_i \sum_X p(X)\log p(x_i) \\
&= \sum_i \sum_X \prod_j p(x_j)\log p(x_i) \\
&= \sum_i \sum_X p(x_i)\log p(x_i)\prod_{j\neq i} p(x_j) \\
&= \sum_i \sum_{x_i} p(x_i)\log p(x_i) \sum_{x_j | j\neq i}\prod_{j\neq i} p(x_j) \\
&= \sum_i \sum_{x_i} p(x_i)\log p(x_i).
\end{aligned}
$$

Mean Field as Non-Convex Lower Bound

- Since $\mathcal{M}_F \subseteq \mathcal{M}$, yields a lower bound on $\log(Z)$:

$$\sup_{\mu \in \mathcal{M}_F} \{w^T\mu + H(p_\mu)\} \leq \sup_{\mu \in \mathcal{M}} \{w^T\mu + H(p_\mu)\} = \log(Z).$$

Mean Field as Non-Convex Lower Bound

- Since $\mathcal{M}_F \subseteq \mathcal{M}$, yields a lower bound on $\log(Z)$:

$$\sup_{\mu \in \mathcal{M}_F} \{w^T \mu + H(p_\mu)\} \leq \sup_{\mu \in \mathcal{M}} \{w^T \mu + H(p_\mu)\} = \log(Z).$$

- Since $\mathcal{M}_F \subseteq \mathcal{M}$, it is an inner approximation:

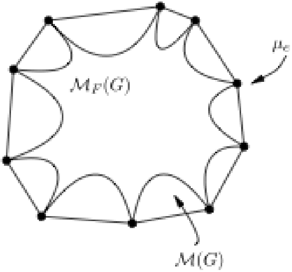


Fig. 5.3 Cartoon illustration of the set $\mathcal{M}_F(G)$ of mean parameters that arise from tractable distributions is a nonconvex inner bound on $\mathcal{M}(G)$. Illustrated here is the case of discrete random variables where $\mathcal{M}(G)$ is a polytope. The circles correspond to mean parameters that arise from delta distributions, and belong to both $\mathcal{M}(G)$ and $\mathcal{M}_F(G)$.

Mean Field as Non-Convex Lower Bound

- Since $\mathcal{M}_F \subseteq \mathcal{M}$, yields a lower bound on $\log(Z)$:

$$\sup_{\mu \in \mathcal{M}_F} \{w^T\mu + H(p_\mu)\} \leq \sup_{\mu \in \mathcal{M}} \{w^T\mu + H(p_\mu)\} = \log(Z).$$

- Since $\mathcal{M}_F \subseteq \mathcal{M}$, it is an inner approximation:

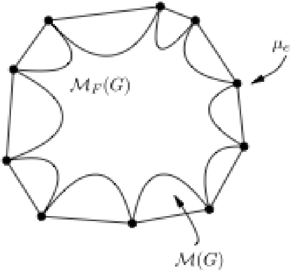


Fig. 5.3 Cartoon illustration of the set $\mathcal{M}_F(G)$ of mean parameters that arise from tractable distributions is a nonconvex inner bound on $\mathcal{M}(G)$. Illustrated here is the case of discrete random variables where $\mathcal{M}(G)$ is a polytope. The circles correspond to mean parameters that arise from delta distributions, and belong to both $\mathcal{M}(G)$ and $\mathcal{M}_F(G)$.

- Constraints $\mu_{ij,st} = \mu_{i,s}\mu_{j,t}$ make it non-convex.

Mean Field Algorithm

- The mean field free energy is defined as

$$
\begin{aligned}
-E_{MF} &\triangleq w^T \mu + H(p_\mu) \\
&= \sum_i \sum_s \mu_{i,s} w_{i,s} + \sum_{(i,j)\in E} \sum_{s,t} \mu_{i,s}\mu_{i,t} w_{ij,st} - \sum_i \sum_s \mu_{i,s} \log \mu_{i,s}.
\end{aligned}
$$

- Last term is entropy, first two terms sometimes called ‘energy’.

Mean Field Algorithm

- The mean field free energy is defined as

$$
\begin{aligned}
-E_{MF} &\triangleq w^T\mu + H(p_\mu) \\
&= \sum_i \sum_s \mu_{i,s} w_{i,s} + \sum_{(i,j)\in E} \sum_{s,t} \mu_{i,s}\mu_{i,t} w_{ij,st} - \sum_i \sum_s \mu_{i,s} \log \mu_{i,s}.
\end{aligned}
$$

- Last term is entropy, first two terms sometimes called ‘energy’.
- Mean field algorithm is coordinate descent on this objective,

$$
-\nabla_{i,s} E_{MF} = w_{i,s} + \sum_{j|(i,j)\in E} \sum_t \mu_{i,j} w_{ij,st} - \log(\mu_{i,s}) - 1.
$$

- Equating to zero for all s and solving for $\mu_{i,s}$ gives update

$$
\mu_{i,s} \propto \exp(w_{i,s} + \sum_{j|(i,j)\in E} \sum_t \mu_{i,j} w_{ij,st}).
$$

Discussion of Mean Field and Structured MF

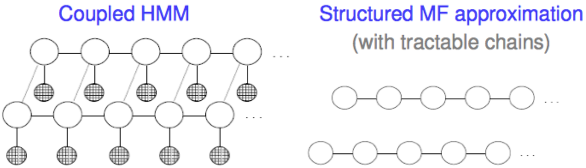
- Mean field is weird:
  - Non-convex approximation to a convex problem.
  - For learning, we want upper bounds on $\log(Z)$.
- Alternative interpretation of mean field:
  - Minimize KL divergence between independent distribution and p.

Discussion of Mean Field and Structured MF

- Mean field is weird:
  - Non-convex approximation to a convex problem.
  - For learning, we want upper bounds on $\log(Z)$.
- Alternative interpretation of mean field:
  - Minimize KL divergence between independent distribution and p.
- Structured mean field:
  - Cost of computing entropy is similar to cost of inference.

Discussion of Mean Field and Structured MF

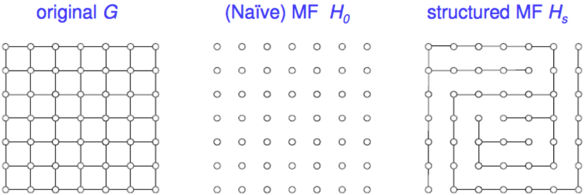
- Mean field is weird:
  - Non-convex approximation to a convex problem.
  - For learning, we want upper bounds on $\log(Z)$.
- Alternative interpretation of mean field:
  - Minimize KL divergence between independent distribution and p.
- Structured mean field:
  - Cost of computing entropy is similar to cost of inference.
  - Use a subgraph where we can perform exact inference.



http://courses.cms.caltech.edu/cs155/slides/cs155-14-variational.pdf

Structured Mean Field with Tree

More edges means better approximation of $\mathcal{M}$ and $H(p_\mu)$:



http://courses.cms.caltech.edu/cs155/slides/cs155-14-variational.pdf

Discussion

- Variational methods write inference as optimization:
  - But optimization seems as hard as original problem.
- We relax the objective/constraints to obtain tractable problems.
- Mean field methods are one way to construct lower-bounds.

For tomorrow, Chapter 4:

Wainwright & Jordan. **Graphical Models, Exponential Families, and Variational Inference**.
Foundations and Trends in Machine Learning. 1(1-2), 2008.