CPSC 540: Machine Learning

Variational Inference

Mark Schmidt

University of British Columbia

Winter 2020

Monte Carlo vs. Variational Inference

Two main strategies for approximate inference:

1. Monte Carlo methods:
   - Approximate p with empirical distribution over samples,

$$p(x) \approx \frac{1}{n}\sum_{i=1}^{n} \mathcal{I}[x^i = x].$$

   - Turns inference into sampling.
2. Variational methods:
   - Approximate p with "closest" distribution q from a tractable family,

$$p(x) \approx q(x).$$

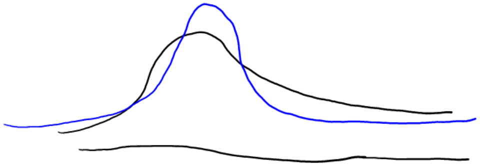
     - E.g., Gaussian, independent Bernoulli, or tree UGM.

(or mixtures of these simple distributions)

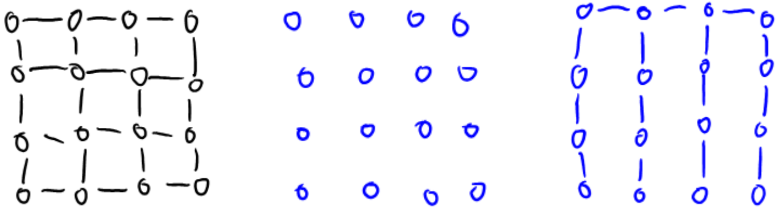
   - Turns inference into optimization.

Variational Inference Illustration

- Approximate non-Gaussian p by a Gaussian q:



- Approximate loopy UGM by independent distribution or tree-structed UGM:



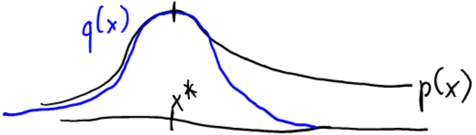
- Variational methods try to find simple distribution q that is closest to target p.
  - This isn't consistent like MCMC, but can be very fast.

Laplace Approximation

- A classic variational method is the Laplace approximation.
  1. Find an x that maximizes $p(x)$,
  $$x^* \in \underset{x}{\text{argmin}}\{-\log p(x)\}.$$
  2. Computer second-order Taylor expansion of $f(x) = -\log p(x)$ at x^*.
  $$-\log p(x) \approx f(x^*) + \underbrace{\nabla f(x^*)}_{0}^T (x - x^*) + \frac{1}{2}(x - x^*)^T \nabla^2 f(x^*)(x - x^*).$$
  3. Find Gaussian distribution q where $-\log q(x)$ has same Taylor expansion at x^*.
  $$-\log q(x) = f(x^*) + \frac{1}{2}(x - x^*)\nabla^2 f(x^*)(x - x^*),$$
  so q follows a $\mathcal{N}(x^*, \nabla^2 f(x^*)^{-1})$ distribution.
     - This is the same approximation used by Newton's method in optimization.

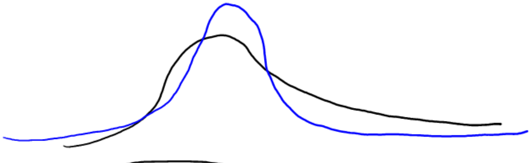
Laplace Approximation

- So Laplace approximation replaces complicated $p(x)$ with Gaussian $q(x)$.
  - Centered at mode and agreeing with 1st/2nd-derivatives of log-likelihood:



- Now you only need to compute Gaussian integrals (linear algebra for many f).
  - Very fast: just solve an optimization (compared to super-slow MCMC).
  - Bad approximation if posterior is heavy-tailed, multi-modal, skewed, etc.

- It might not even give you the "best" Gaussian approximation:



Kullback-Leibler (KL) Divergence

- How do we define "closeness" between a distribution p and q?
- A common measure is Kullback-Leibler (KL) divergence between p and q:

$$\text{KL}(p \,||\, q) = \sum_x p(x) \log \frac{p(x)}{q(x)}.$$

  - Replace sum with integral for continuous families of q distributions.
- Also called information gain: "information lost when p is approximated by q".
  - If p and q are the same, we have $KL(p \,||\, q) = 0$ (no information lost).
  - Otherwise, $KL(p \,||\, q)$ grows as it becomes hard to predict p from q.
- Unfortunately, this requires summing/integrating over p.
  - The problem we are trying to solve.

Minimizing Reverse KL Divergence

- Instead of using KL, most variational methods minimize reverse KL,

$$\mathsf{KL}(q \,||\, p) = \sum_x q(x) \log \frac{q(x)}{p(x)} = \sum_x q(x) \log \frac{q(x)}{\tilde{p}(x)} Z.$$

  which just swaps all p and q values in the definition (KL is not commutative).
  - Not intuitive: "how much information is lost when we approximate q by p".

- But, reverse KL only needs unnormalized distribution $\tilde{p}$,

$$\begin{aligned}\mathsf{KL}(q \,||\, p) &= \sum_x q(x) \log q(x) - \sum_x q(x) \log \tilde{p}(x) + \sum_x q(x) \log(Z) \\ &= \sum_x q(x) \log \frac{q(x)}{\tilde{p}(x)} + \underbrace{\log(Z)}_{\text{const. in } q}.\end{aligned}$$

- By non-negativiy of KL this also gives a lower bound on $\log(Z)$.
  - Called the ELBO ("evidence lower bound").

Coordinate Optimization: Mean Field Approximation

- This "variational lower bound" still seems difficult to work with.
  - But with appropriate q we can do coordinate optimization.

- Consider minimizing reverse KL with independent q,

$$q(x) = \prod_{j=1}^{d} q_j(x_j),$$

where we choose q to be conjugate (usually discrete or Gaussian).

  - If we fix q_{-j} and optimize the functional q_j we obtain (see Murphy's book)

$$q_j(x_j) \propto \exp\left(\mathbb{E}_{q_{-j}}[\log \tilde{p}(x)]\right),$$

  which we can use to update q_j for a particular j.

Coordinate Optimization: Mean Field Approximation

- Each iteration we choose a j and set q based on mean (of neighbours),

$$q_j(x_j) \propto \exp\left(\mathbb{E}_{q_{-j}}[\log \tilde{p}(x)]\right).$$

- This improves the (non-convex) reverse KL on each iteration.
- Applying this update is called:
  - Mean field method (graphical models).
  - Variational Bayes (Bayesian inference).

3 Coordinate-Wise Algorithms

- ICM is a coordinate-wise method for approximate decoding:
  - Choose a coordinate i to update.
  - Maximize x_i keeping other variables fixed.
- Gibbs sampling is a coordinate-wise method for approximate sampling:
  - Choose a coordinate i to update.
  - Sample x_i keeping other variables fixed.
- Mean field is a coordinate-wise method for approximate marginalization:
  - Choose a coordinate i to update.
  - Update $\underbrace{q_i(x_i)}_{\text{for all } x_i}$ keeping other variables fixed ($q_i(x_i)$ approximates $p_i(x_i)$).

3 Coordinate-Wise Algorithms

- Consider a pairwise UGM:

$$p(x_1, x_2, \ldots, x_d) \propto \left(\prod_{i=1}^{d} \phi_i(x_i)\right)\left(\prod_{(i,j)\in E} \phi_{ij}(x_i, x_j)\right),$$

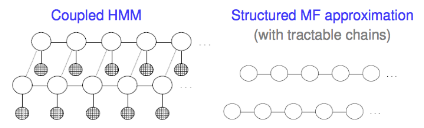
- ICM for updating a node i with 2 neighbours (j and k).
  1. Compute $M_i(x_i) = \phi_i(x_i)\phi_{ij}(x_i, x_j)\phi_{ik}(x_i, x_k)$ for all x_i.
  2. Set x_i to the largest value of $M_i(x_i)$.

- Gibbs for updating a node i with 2 neighbours (j and k).
  1. Compute $M_i(x_i) = \phi_i(x_i)\phi_{ij}(x_i, x_j)\phi_{ik}(x_i, x_k)$ for all x_i.
  2. Sample x_i proportional to $M_i(x_i)$.

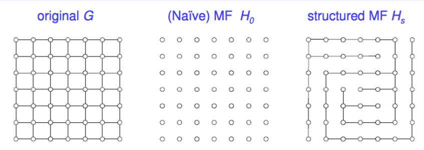
- Mean field for updating a node i with 2 neighbours (j and k).
  1. Compute $M_i(x_i) = \phi_i(x_i)\exp\left(\sum_{x_j} q_j(x_j)\log\phi_{ij}(x_i, x_j) + \sum_{x_k} q_k(x_k)\log\phi_{ik}(x_i, x_k)\right)$.
  2. Set $q_i(x_i)$ proportional to $M_i(x_i)$.

Structure Mean Field

- Common variant is structured mean field: q function includes some of the edges.



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



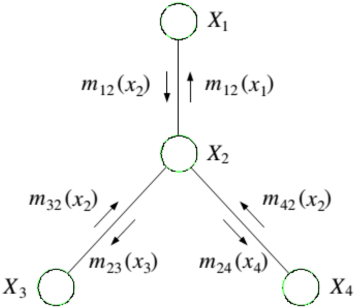
http://courses.cms.caltech.edu/cs155/slides/cs155-14-variational.pdf

- Original LDA article proposed a structured mean field approximation.

Previously: Belief Propagation

- We've discussed belief propagation for forest-structured UGMs.

(undirected graphs with no loops, which must be pairwise)



https://www.quora.com/

Probabilistic-graphical-models-what-are-the-relationships-between-sum-product-algorithm-belief-propagation-and-junction-tree-

- Defines "messages" that can be sent along each edge.
  - Generalizes forward-backward algorithm.

Loopy Belief Propagation

- In pairwise UGM, belief propagation "message" from parent p to child c is gven by

$$M_{pc}(x_c) \propto \sum_{x_p} \phi_i(x_p)\phi_{pc}(x_p, x_c)M_{jp}(x_p)M_{kp}(x_p),$$

  assuming that parent p has parents j and k.
  - We get marginals by multiplying all incoming messages with local potentials.

- Loopy belief propagation: a "hacker" approach to approximate marginals:
  - Choose an edge ic to update.
  - Update messages $M_{ic}(x_c)$ keeping all other messages fixed.
  - Repeat until "convergence".
    - We approximate marginals by multiplying all incoming messages with local potentials.

- Empirically much better than mean field, we've spent 20 years figuring out why.

Discussion of Loopy Belief Propagation

- Loopy BP decoding is used for "error correction" in WiFi and Skype.
  - Called "turbo codes" in information theory.
- Loopy BP is not optimizing an objective function.
  - Convergence of loopy BP is hard to characterize: does not converge in general.
- If it converges, loopy BP finds fixed point of "Bethe free energy":
  - Instead of "Gibbs mean-field free-energy" for mean field, which lower bounds Z.
  - Bethe typically gives better approximation than mean field, but not a bound.
- Recent works give convex variants that upper bound Z.
  - Tree-reweighted belief propagation.
  - Variations that are guaranteed to converge.
- Messages only have closed-form update for conjugate models.
  - Can approximate non-conjugate models using expectation propagation.

Variational vs. Monte Carlo

- Monte Carlo vs. variational methods:
  - Variational methods are typically more complicated.
  - Variational methods are not consistent.
    - q does not converge to p if we run the algorithm forever.
  - But variational methods often give better approximation for the same time.
    - Although MCMC is easier to parallelize.
  - Variational methods typically have similar cost to MAP.

- Combinations of variational inference and stochastic methods:
  - Stochastic variational inference (SVI): use SGD to speed up variational methods.
  - Variational MCMC: use Metropolis-Hastings where variational q can make proposals.

Convex Relaxations

- I've overviewed the "classic" view of variational methods that they minimize KL.
- Modern view: write exact inference as constrained convex optimization (bonus).
  - Based on convex conjugate, writing inference as maximizing entropy with constraints.
  - Different methods correspond to different function/constraints approximations.
  - There are also convex relaxations that approximate with linear programs.
- For an overview of this and all things variational, see:
  `people.eecs.berkeley.edu/~wainwrig/Papers/WaiJor08_FTML.pdf`

Summary

- Variational methods approximate p with a simpler distribution q.
- Mean field approximation minimizes reverse KL divergence with independent q.
- Loopy belief propagation is a heuristic that often works well.
- Next time: food-inspired models?

Variational Inference: Constrained Optimization View

- Modern view of variational inference:
  - Formulate inference problem as constrained optimization.
  - Approximate the function or constraints to make it easy.

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

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

- E.g., if we consider for logistic regression

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

we have 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, $\sup$ is ∞.

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.

Bonus slide: Maximum Likelihood and Maximum Entropy

- The maximum likelihood parameters w satisfy:

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

  which is $-\infty$ unless $F(D) = \mu$ (e.g., maximum likelihood w), 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.

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^*.
- Notatation trick: we put everything "inside" w to discuss general log-potentials.

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:

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

Mean Field as Non-Convex Lower Bound

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

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

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

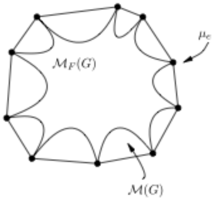
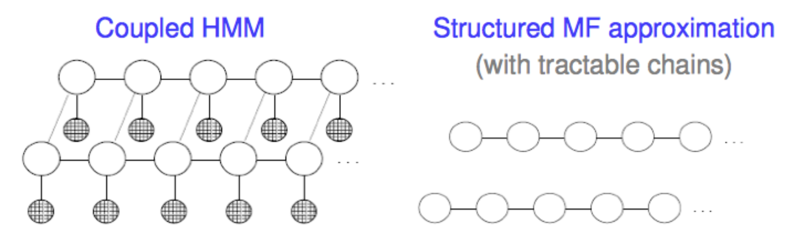


Fig. 5.3 Cartoon illustration of the set $\mathcal{M}_F(G)$ of mean parameters that arise from tractable distributions is a nonconvex inner bound on $\mathcal{M}(G)$. Illustrated here is the case of discrete random variables where $\mathcal{M}(G)$ is a polytope. The circles correspond to mean parameters 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 is coordinate descent on $w^T \mu + H(p_\mu)$ over $\mathcal{M}_F$.

Discussion of Mean Field and Structured MF

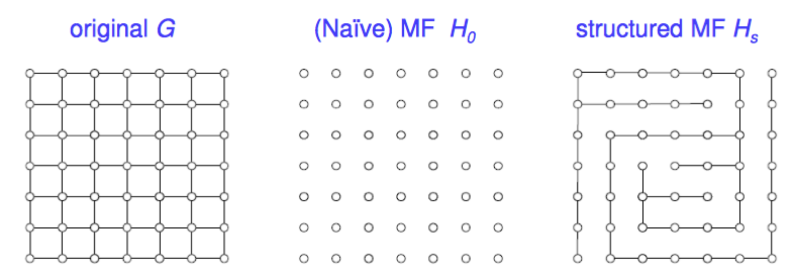
- Mean field is weird:
  - Non-convex approximation to a convex problem.
  - For learning, we want upper bounds on $\log(Z)$.
- Structured mean field:
  - Cost of computing entropy is similar to cost of inference.
  - Use a subgraph where we can perform exact inference.



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

- Fixed points of loopy correspond to using "Bethe" approximation of entropy and "local polytope" approximation of "marginal polytope".

- You can design better variational methods by constructing better approximations.