Nonparametric Variational Inference

Sam Gershman, **Matt Hoffman**, David Blei
Princeton University, Adobe Creative Technologies Lab

Approximate inference

- We want to approximate a distribution p(θ), but we can only compute it up to a constant.
  - E.g., we're interested in p(θ | y), but can only compute p(y, θ).

Variational inference

- Variational inference approximates $p(\theta \mid y)$ with some tractable distribution $q(\theta)$ by solving an optimization problem.

Variational inference: the agony and the ecstasy

- Variational methods often converge much faster than Markov chain Monte Carlo (MCMC) methods. But they suffer from two major drawbacks:

1. **Model expressivity:** updates and objective functions are usually restricted to conditionally conjugate models paired with simple approximating distributions.

2. **User-friendliness:** deriving variational updates involves a fair amount of tedious math.

Nonparametric variational inference

- We derive a variational inference algorithm that
  1. is applicable to models without conditional conjugacy and
  2. only requires the ability to evaluate the log-posterior (up to a constant), its gradient, and optionally the diagonal of its Hessian.

Our approach

- We restrict q to be a mixture of Gaussians (cf. the mixture mean-field approach of Lawrence, Jaakola, et al.):

  $q(\theta) = (1/N) \Sigma_n N(\theta; \mu_n, \sigma_n^2)$

- Can be interpreted as kernel density estimation of the posterior $p(\theta \mid y)$.

Our approach

- The standard variational objective ("evidence lower bound", or ELBO) is

  $F(q) = E_q[\log p(y, \theta)] - E_q[\log q(\theta)]$

  where y is a set of observed variables, θ is a set of latent variables, and q is the approximating distribution.

- We derive an approximate ELBO that can be easily optimized using gradient methods (e.g. LBFGS).

The basic idea

$$F(q) = E_q[\log p(y, \theta)] - E_q[\log q(\theta)]$$

Approximate using Taylor series expansion around the mean of each Gaussian component

Lower-bound entropy using Jensen's inequality and by exploiting properties of Gaussian mixtures

Entropy bound

$$H(q) = -\int_\theta q(\theta) \log q(\theta)\, d\theta$$

$$= -\int_\theta q(\theta) \log (1/N) \Sigma_n N(\theta; \mu_n, \sigma_n^2)\, d\theta$$

$$\geq -(1/N) \Sigma_n \log \int_\theta q(\theta) N(\theta; \mu_n, \sigma_n^2)\, d\theta$$

$$\geq -(1/N) \Sigma_n \log \Sigma_j N(\mu_n; \mu_j, \sigma_n^2 + \sigma_j^2)$$

Log-joint bound

2nd-order Taylor expansion (multivariate delta method for moments) yields

$E_q[\log p(y, \theta)] \approx (1/N)\ \Sigma_n \log p(y, \mu_n) + (\sigma_n^2/2)\ \mathrm{Tr}(H_n)$

Only requires diagonal of Hessian H_n evaluated at μ_n.

Approximate ELBO

Encourages each μ_n to be in a high-density region

Discourages overly broad Gaussians

$$(1/N)\ \Sigma_n \log p(y, \mu_n) + (\sigma_n^2/2)\ \mathrm{Tr}(H_n)$$
$$- \log \Sigma_j N(\mu_n; \mu_j, \sigma_n^2 + \sigma_j^2)$$

Encourages means to spread out

Encourages Gaussians to be broader

Optimizing the approximate ELBO

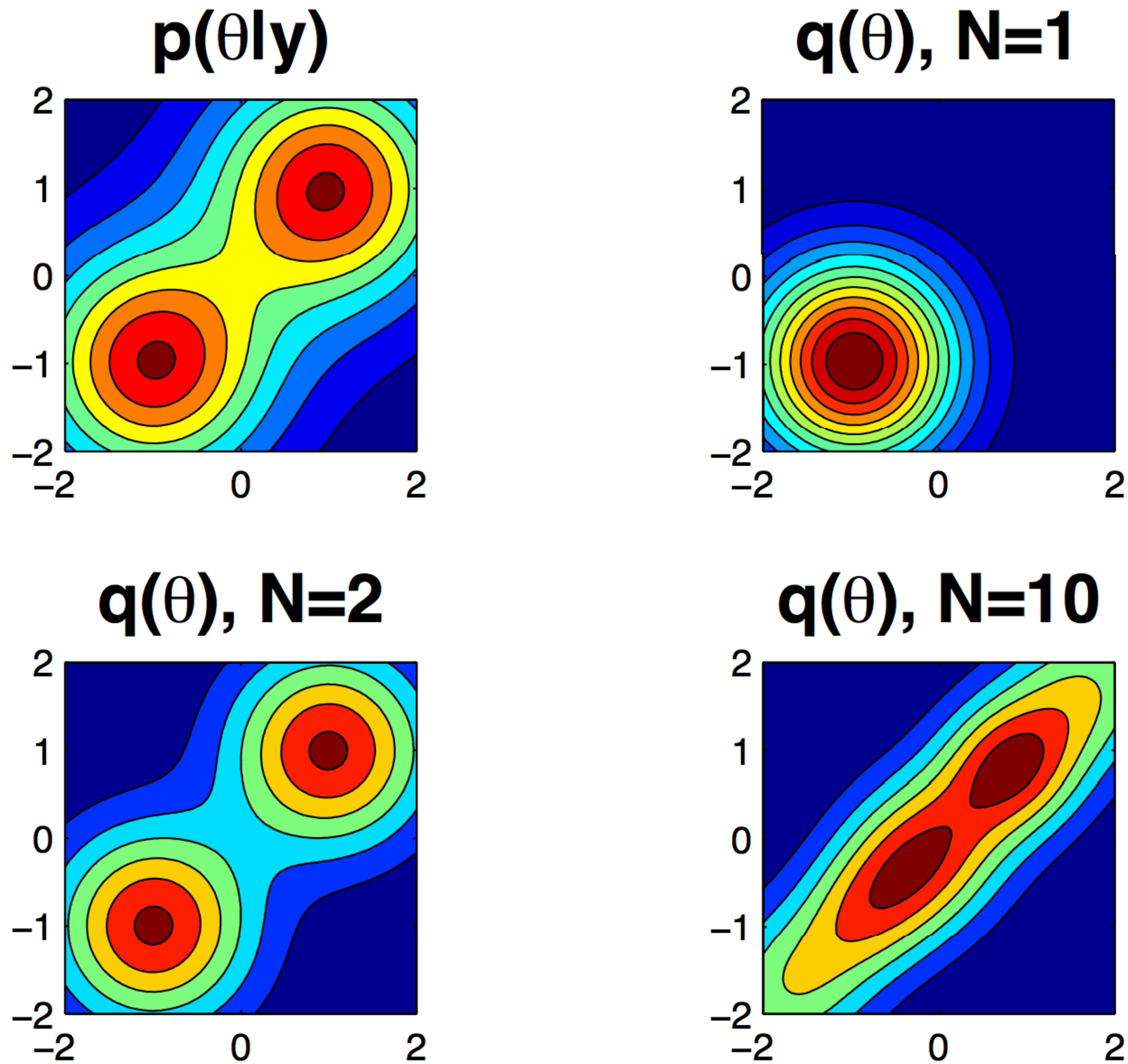
$(1/N)\ \Sigma_n \log p(y, \mu_n) + (\sigma_n^2/2)\ \mathrm{Tr}(H_n)$
$\quad - \log \Sigma_j N(\mu_n; \mu_j, \sigma_n^2 + \sigma_j^2)$

1. Optimize each μ_n holding others fixed, ignoring Hessian trace term.
   - Avoids computing N^2 third derivatives.
   - Avoids possible degeneracies with non-log-concave posteriors.
2. Optimize σ vector holding μ fixed.

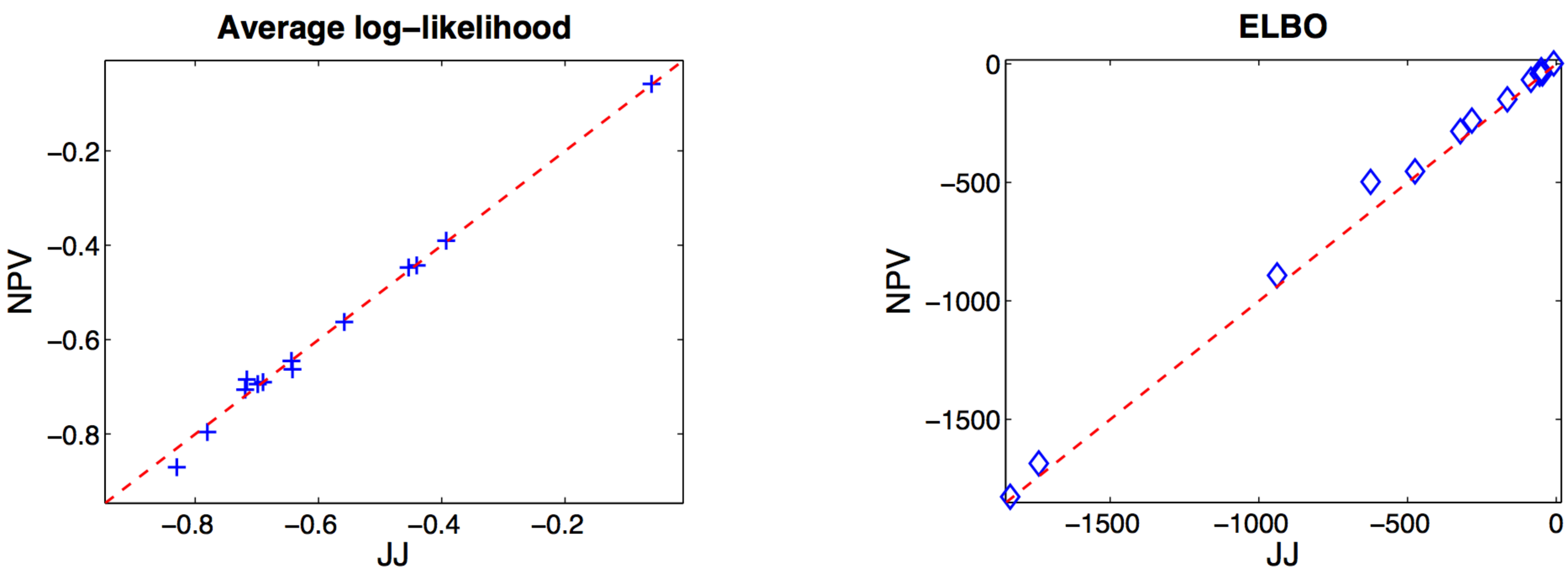
Relationships to other algorithms

- $N = 1$, $\sigma \to 0$: maximum a posteriori (MAP).
- $N = 1$, σ variable: diagonalized Laplace approximation.
- $N > 1$, $\sigma \to 0$: quasi-Monte Carlo.
- $N > 1$, σ variable: a form of mixture mean-field (Jaakkola & Jordan, 1998; Lawrence, 2000).
    - Analogous to KDE.

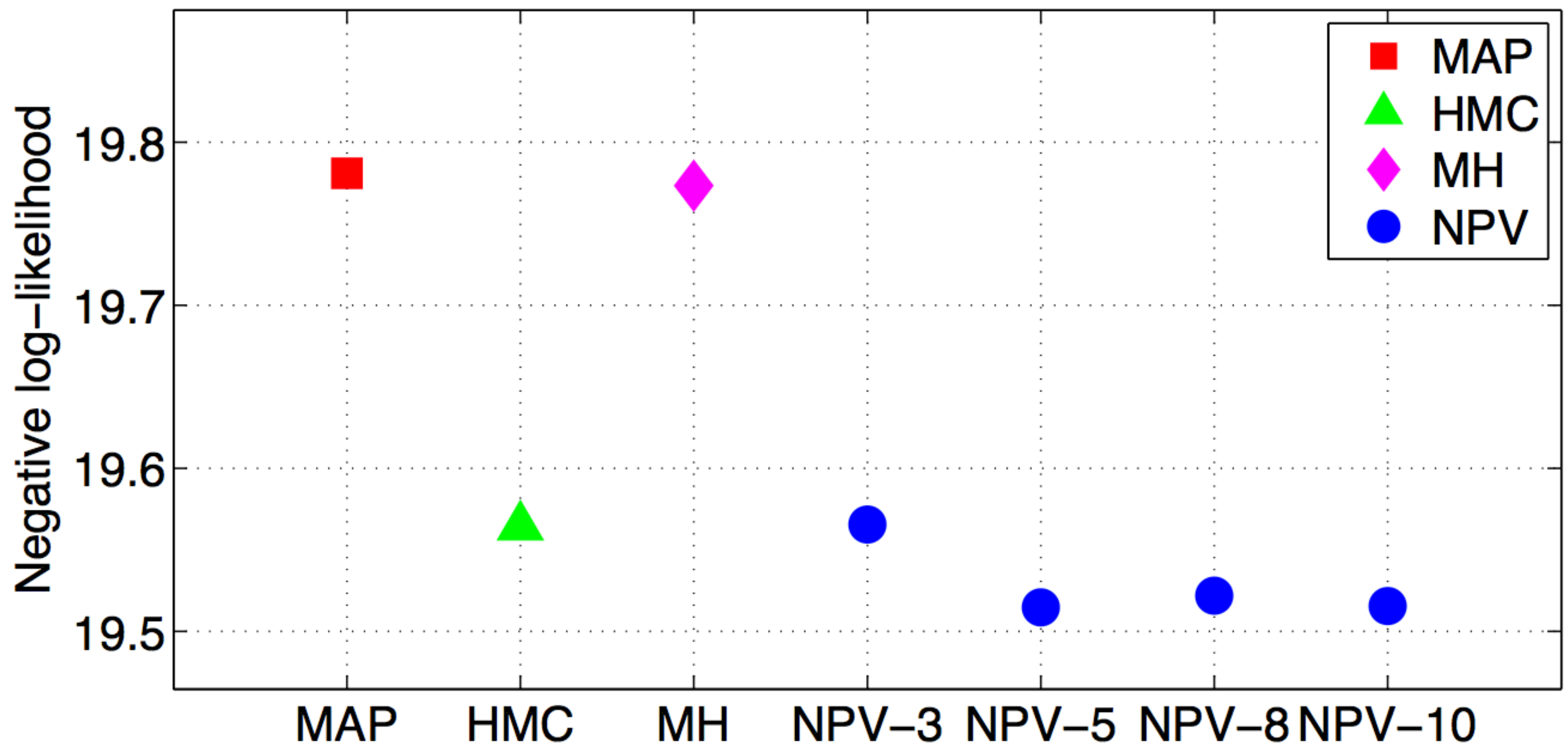
Synthetic example



Logistic regression: NPV vs. Jordan & Jaakkola



Topographic latent source analysis: NPV vs. MAP and MCMC



Summary

- Nonparametric variational inference
  1. circumvents conjugacy restrictions and
  2. allows for more expressive variational distributions than mean-field.
- Can be used for arbitrary graphical models.

Future work

- Consider more flexible classes of approximating distributions
  - Non-isotropic Gaussians
  - Nonuniform mixture weights
- Extend to models with discrete random variables
  - Continuous relaxations?
- Implement in Stan (mc-stan.org)