HERIOT WATT UNIVERSITY

On approximate message passing and Expectation-Propagation for scalable inference

LEADERS IN IDEAS AND SOLUTIONS

Yoann Altmann

Joint work with A. Perelli, D. Yao, S. McLaughlin and M. Davies

RAEng Research Fellow
School of Engineering and Physical Sciences
Heriot-Watt University

Edinbburgh, February, 20th 2019

ROYAL ACADEMY OF ENGINEERING

UDRC

HERIOT WATT UNIVERSITY

Motivations

- Bayesian estimation for complex models
  - Hierarchical
  - Discrete/continuous variables
  - Large scale (imaging)
  - Model selection

- Simulation methods
  - Flexible but not (yet) scalable
- More efficient methods needed

HERIOT WATT UNIVERSITY

Potential solutions

- Point estimation
  - MAP: convex/non-convex optimization
  - Limited uncertainty quantification

- Approximate methods
  - Proximal MCMC
  - Variational Bayes (VB) methods
  - Approximate message passing (AMP/EP)

HERIOT WATT UNIVERSITY

Bayesian modeling

- Likelihood (observation model)

$$f(\boldsymbol{y}|\boldsymbol{x})$$

- (Hierarchical) prior model

$$f(\boldsymbol{x}, \theta) = f(\boldsymbol{x}|\theta) f(\theta)$$

HERIOT WATT UNIVERSITY

Bayesian modeling

- Exact model

$$f(\boldsymbol{x}, \theta|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{x})f(\boldsymbol{x}|\theta)f(\theta)}{f(\boldsymbol{y})}$$

- Approximating distribution

$$f(\boldsymbol{x}, \theta|\boldsymbol{y}) \approx q(\boldsymbol{x}, \theta)$$

  – Proximal MCMC: Moreau approx.

  – VB/EP: Divergence-based

- $f(\boldsymbol{y})$: evidence (model selection)

HERIOT WATT UNIVERSITY

Approximate methods

- Variational Bayes

Often $f(\boldsymbol{x}, \theta|\boldsymbol{y}) \approx q(\boldsymbol{x}, \theta) = q(\boldsymbol{x})q(\theta)$

$$\min_{q(\boldsymbol{x}),q(\theta)} KL\big(q(\boldsymbol{x})q(\theta) \,||\, f(\boldsymbol{x}, \theta|\boldsymbol{y})\big)$$

- Expectation Propagation

$f(\boldsymbol{y}, \boldsymbol{x}, \theta) = f(\boldsymbol{y}|\boldsymbol{x})f(\boldsymbol{x}|\theta)f(\theta)$
$q(\boldsymbol{x}, \theta) \propto q_0(\boldsymbol{x})q_1(\boldsymbol{x}, \theta)q_0(\theta)$

$$\min_{Z,\, q(\boldsymbol{x},\theta)} KL(f(\boldsymbol{y}, \boldsymbol{x}, \theta) \,||\, Zq(\boldsymbol{x}, \theta))$$

$$Z \approx f(y)$$

EP: preserves better marginals

HERIOT WATT UNIVERSITY

Iterative minimization

$f(\boldsymbol{y}, \boldsymbol{x}, \theta) = f(\boldsymbol{y}|\boldsymbol{x}) f(\boldsymbol{x}|\theta) f(\theta)$
$q(\boldsymbol{x}, \theta) \propto q_0(\boldsymbol{x}) q_1(\boldsymbol{x}, \theta) q_0(\theta)$

- $\min_{q_0(\boldsymbol{x})} \; KL(f(y|\boldsymbol{x}) q_1(\boldsymbol{x}, \theta) q_0(\theta) || q_0(\boldsymbol{x}) q_1(\boldsymbol{x}, \theta) q_0(\theta))$
- $\min_{q_1(\boldsymbol{x}, \theta)} KL(q_0(\boldsymbol{x}) f(\boldsymbol{x}|\theta) q_0(\theta) || q_0(\boldsymbol{x}) q_1(\boldsymbol{x}, \theta) q_0(\theta))$
- $\min_{q_0(\theta)} KL(q_0(\boldsymbol{x}) q_1(\boldsymbol{x}, \theta) f(\theta) || q_0(\boldsymbol{x}) q_1(\boldsymbol{x}, \theta) q_0(\theta))$

… until convergence

HERIOT WATT UNIVERSITY

Divergence minimization

- Difficult in general
  - Gaussian approx.: moment matching
  - Exponential families: natural parameters
- More difficult with multivariate distributions
- Approximating graph/distributions
  - Tradeoff accuracy/complexity
  - Sequential updates
  - Possibility to add constraints

HERIOT WATT UNIVERSITY

Example

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e} \Rightarrow f(\boldsymbol{y}|\boldsymbol{A}\boldsymbol{x}) = \prod_{m=1}^{M} f(y_m|\boldsymbol{a}_m^T x)$$

$$f(\boldsymbol{x}) = \prod_{n=1}^{N} f_n(x_n)$$

- $\boldsymbol{A}$: blur, dictionary,…
- Arbitrary observation noise (Gaussian, Poisson,…)
- Separable prior model

HERIOT WATT UNIVERSITY

Classical EP scheme

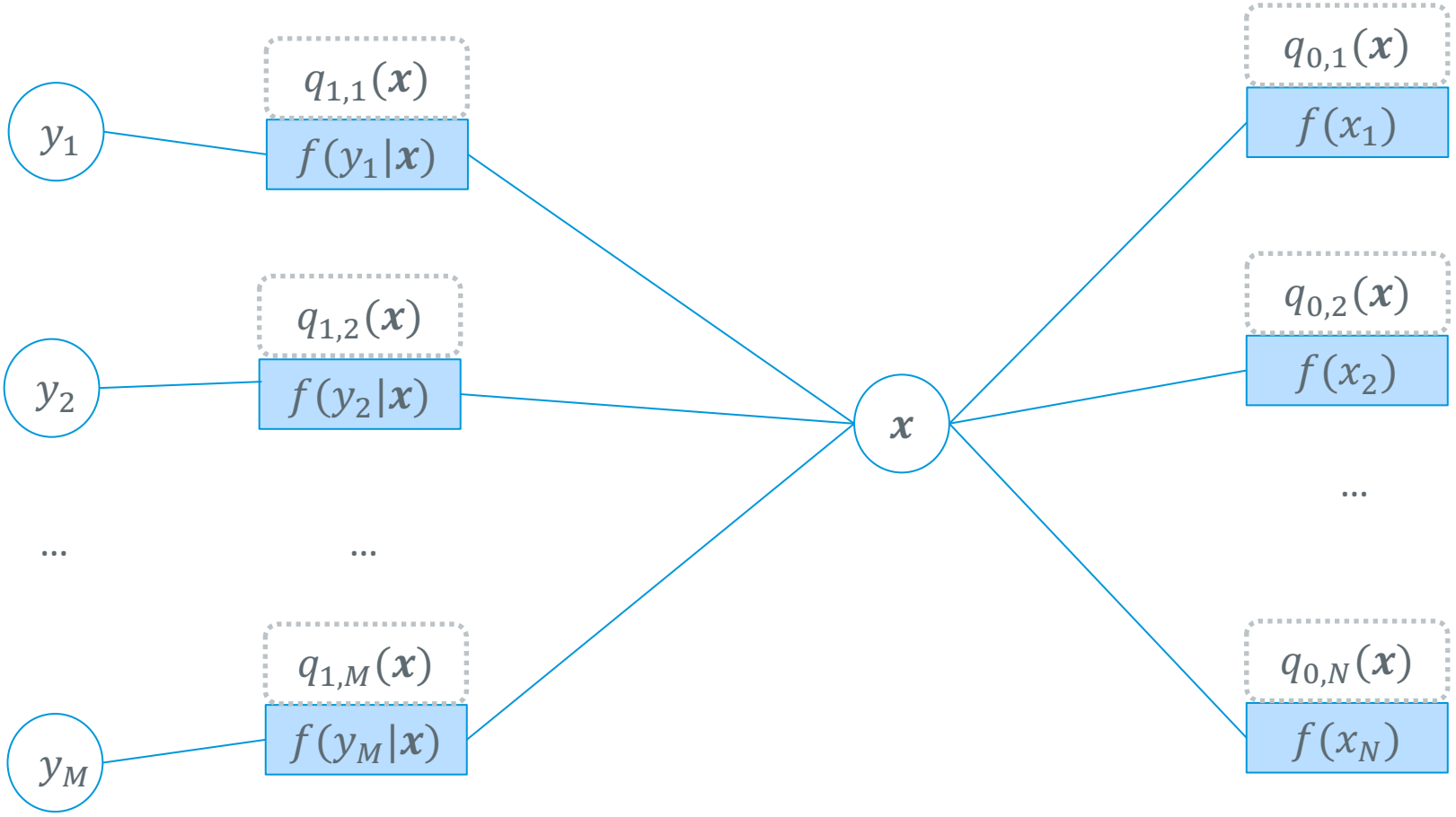
$$f(\boldsymbol{y}, \boldsymbol{x}) = \left[\prod_{m=1}^{M} f(y_m | \boldsymbol{a}_m^T x)\right]\left[\prod_{n=1}^{N} f_n(x_n)\right]$$

$$q(\boldsymbol{x}) \propto \left[\prod_{m=1}^{M} q_{1,m}(\boldsymbol{x})\right]\left[\prod_{n=1}^{N} q_{0,n}(\boldsymbol{x})\right]$$

$N + M$ factors in the approximation

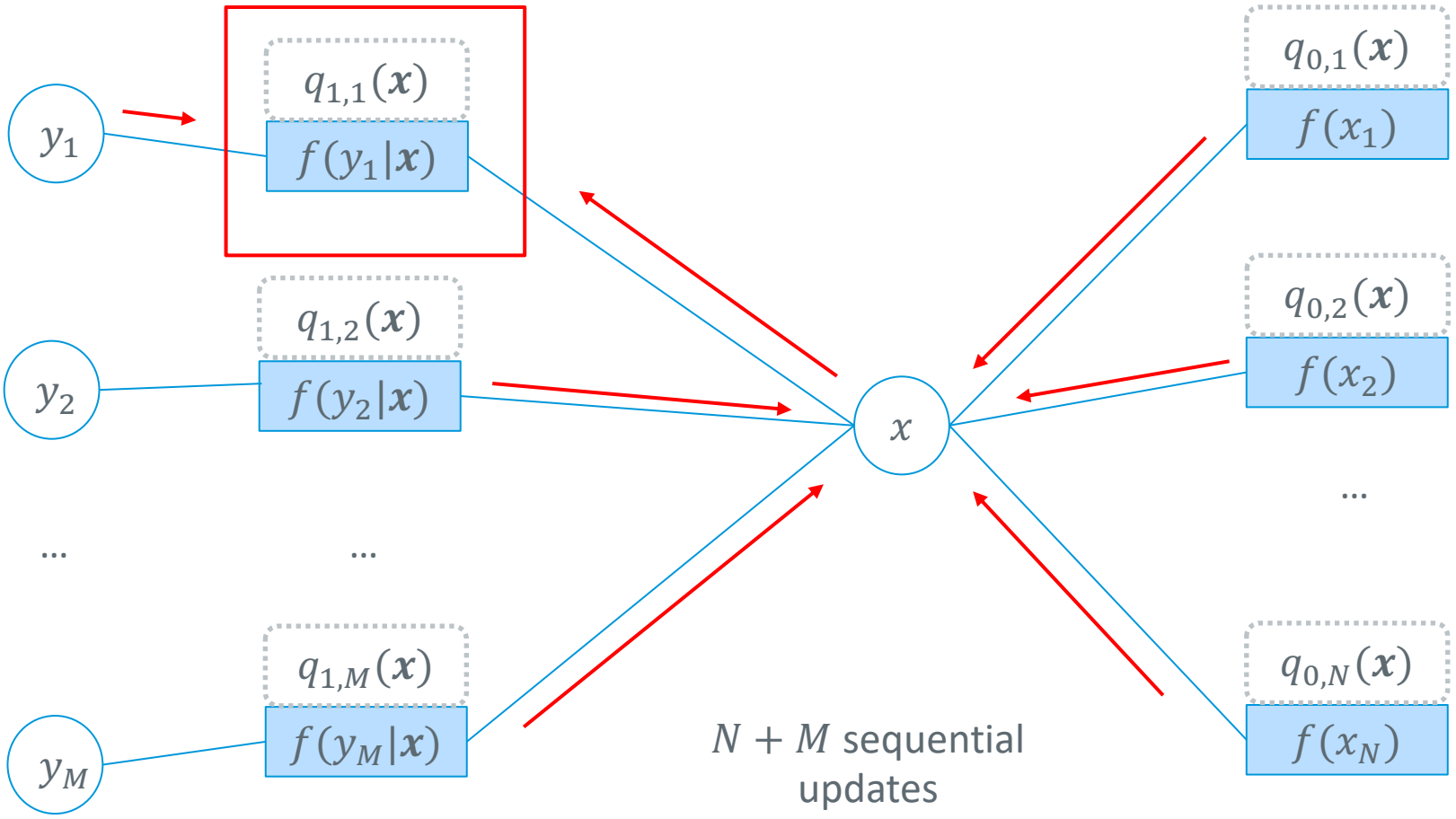
HERIOT WATT UNIVERSITY

Classical Bayesian graph



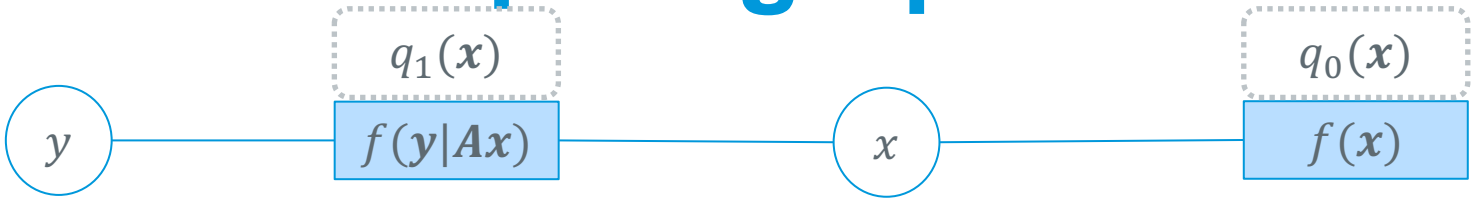
HERIOT WATT UNIVERSITY

EP as message passing



HERIOT WATT UNIVERSITY

EP with compact graph



- Simple models: 2 updates

$(q_1(\boldsymbol{x}), q_1(\boldsymbol{x}))$ isotropic cov.: AMP

$(q_1(\boldsymbol{x}), q_1(\boldsymbol{x}))$ diagonal cov.: VAMP

$f(\boldsymbol{x})$ implicit: D-AMP

- Messages potentially difficult to compute…

HERIOT WATT UNIVERSITY

Extended model

$$f(\boldsymbol{y}|\boldsymbol{u}) = \prod_{m=1}^{M} f(y_m|u_m)$$

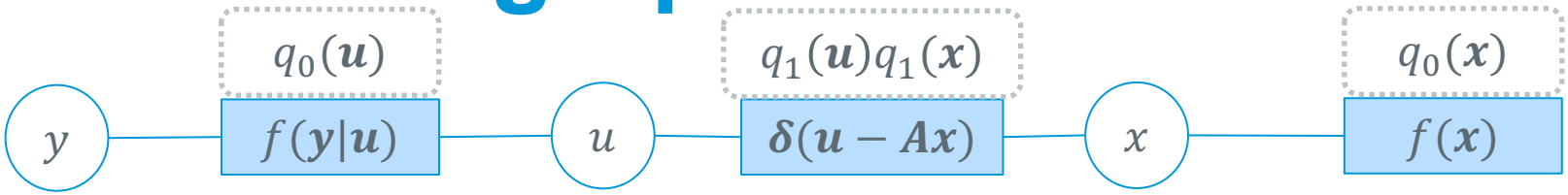
$$f(\boldsymbol{u}|\boldsymbol{x}) = \prod_{m=1}^{M} f(u_m|\boldsymbol{x}) = \prod_{m=1}^{M} \delta(u_m - a_m^T \boldsymbol{x})$$

$$f(\boldsymbol{x}) = \prod_{n=1}^{N} f_n(x_n)$$

$$f(\boldsymbol{y}|\boldsymbol{x}) = \int f(\boldsymbol{y}|\boldsymbol{u}) f(\boldsymbol{u}|\boldsymbol{x}) du$$

HERIOT WATT UNIVERSITY

Extended graph

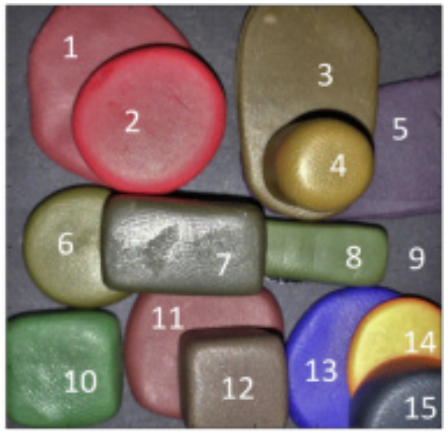


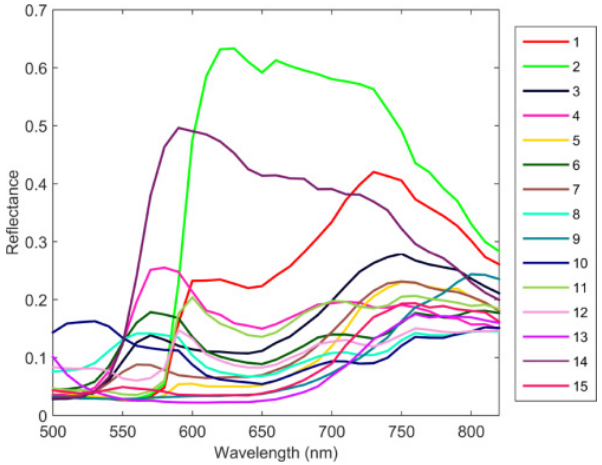
- Flexibility in choosing $q_0(\boldsymbol{u}), q_0(\boldsymbol{x})$ and $q_1(\boldsymbol{u},\boldsymbol{x})$
  - Partial correlation
- Here $(\boldsymbol{u},\boldsymbol{x})$ a posteriori independent
  - But covariance for $\boldsymbol{x}$ and/or $\boldsymbol{u}$
- $N \ll M$: regression ; $\mathrm{M} \ll N$: CS
- Isotropic/diagonal messages: AMP/VAMP

HERIOT WATT UNIVERSITY

Example: regression with Poisson noise

- Spectral unmixing in photon-starved regime
- $M = 33$ spectral bands
- $N = 15$ sources (polymer clay)

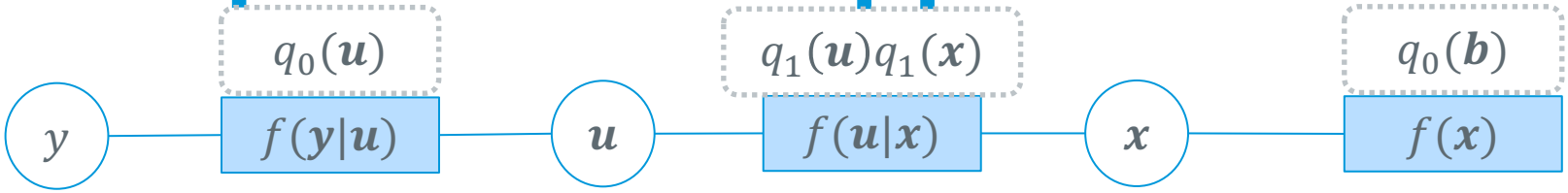




Altmann et al. ICASSP 2019, to appear

HERIOT WATT UNIVERSITY

Comparison EP approx.

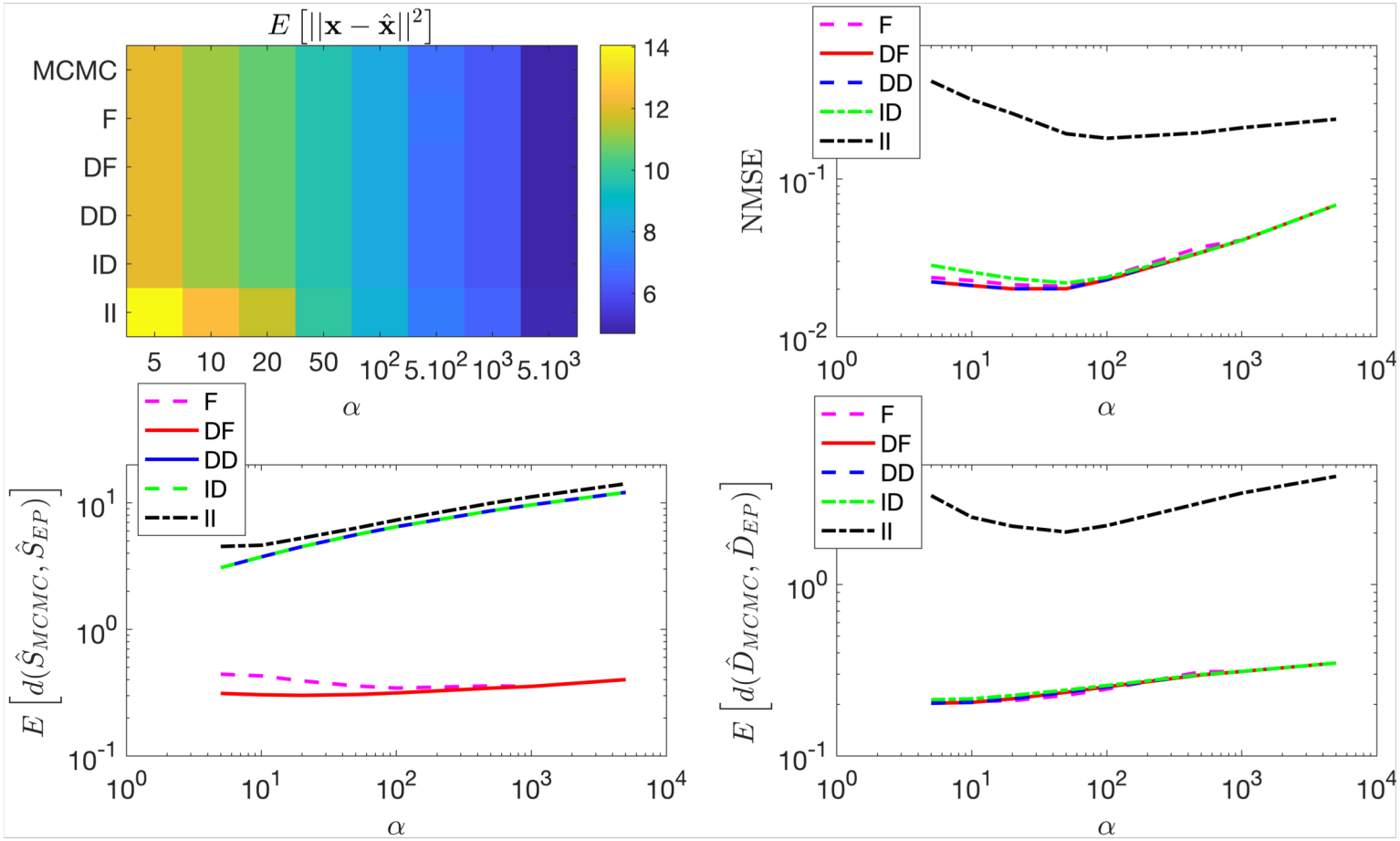


- Exponential priors $f(x)$
- Gaussian approx. only
- Different choices of covariance matrices
- $q_0(\boldsymbol{u})$: diag. (D) $q_1(\boldsymbol{u})$: iso./diag.
- $q_0(\boldsymbol{x})$: diag. $q_1(\boldsymbol{x})$: iso./diag./full (I/D/F)

- Positivity not enforced in $q(\boldsymbol{x})$

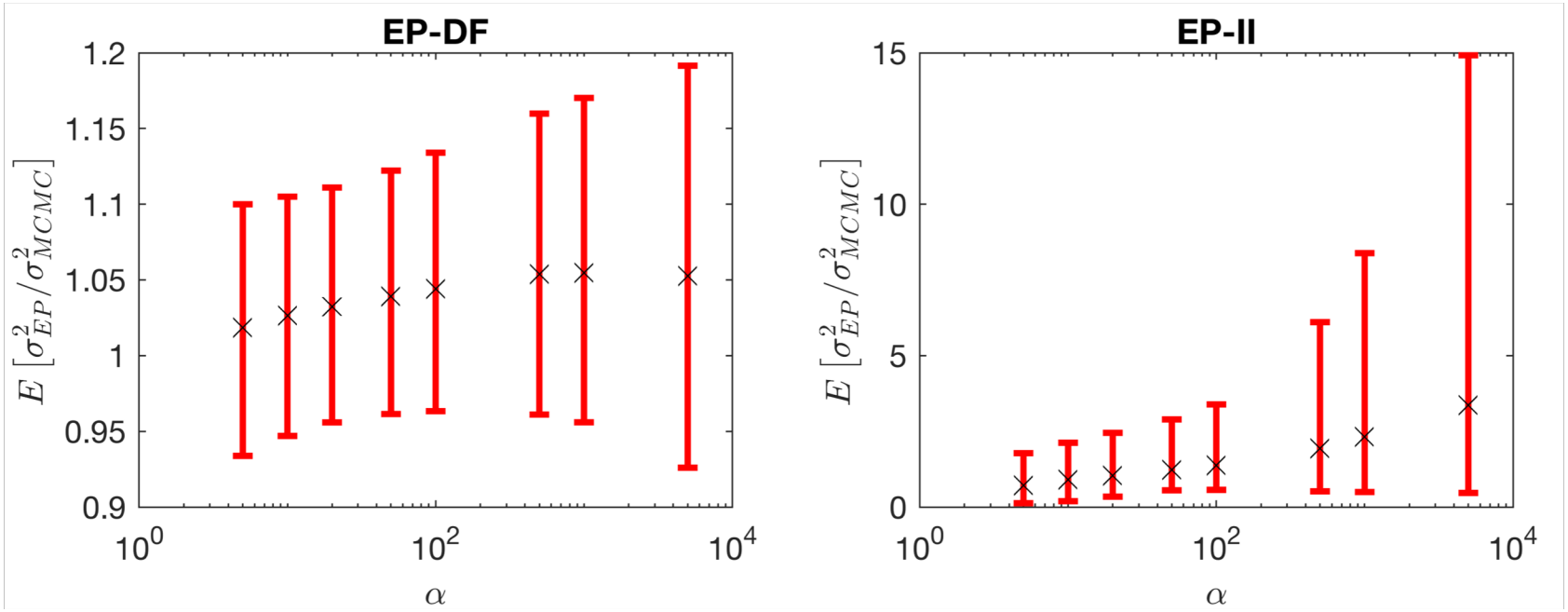
HERIOT WATT UNIVERSITY

Results



HERIOT WATT UNIVERSITY

Results



EP-DF only slightly overestimates the marginal variances

HERIOT WATT UNIVERSITY

Conclusion

- Approximate message passing
  - Scalable/distributed
  - Flexible
  - Fast convergence
- Good estimation performance
- Approx. evidence: by-product
- But can be difficult to implement

HERIOT WATT UNIVERSITY

Ongoing work

- Convergence properties of EP
  - Constraints seem to help
- Approximations
  - for non-Gaussian noise
  - Hard constraints (e.g., positivity)
- EP within simulation schemes
- Application to imaging (Denoising-AMP)