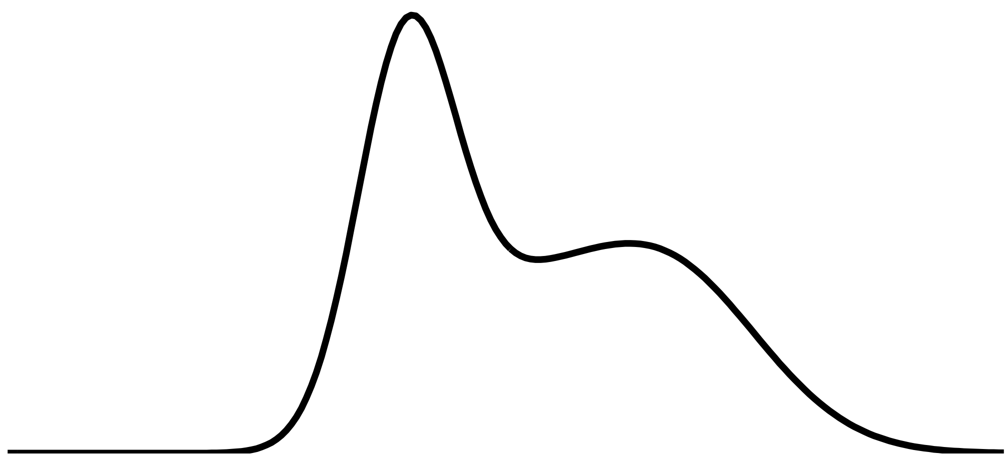
Two problems with variational expectation maximisation for time-series models

Richard Turner, Pietro Berkes and Maneesh Sahani

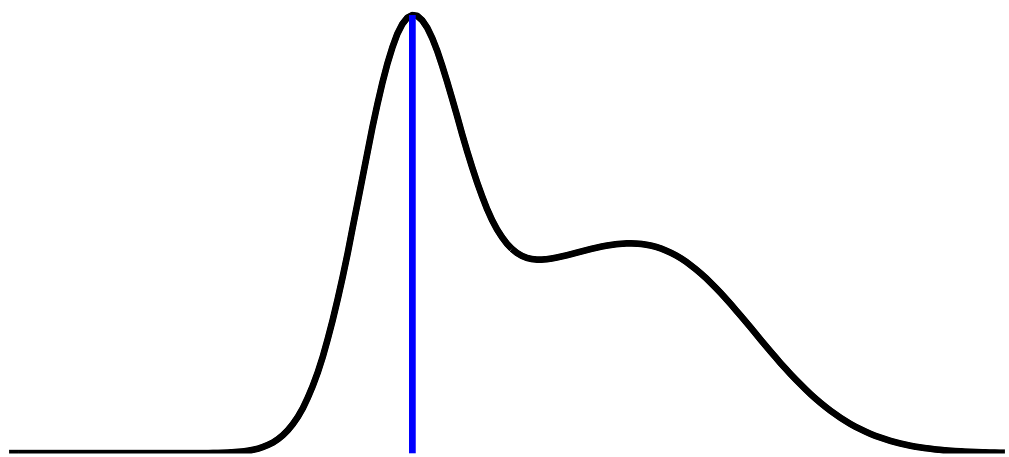
({turner, berkes, maneesh}@gatsby.ucl.ac.uk)

Gatsby Computational Neuroscience Unit, UCL, London

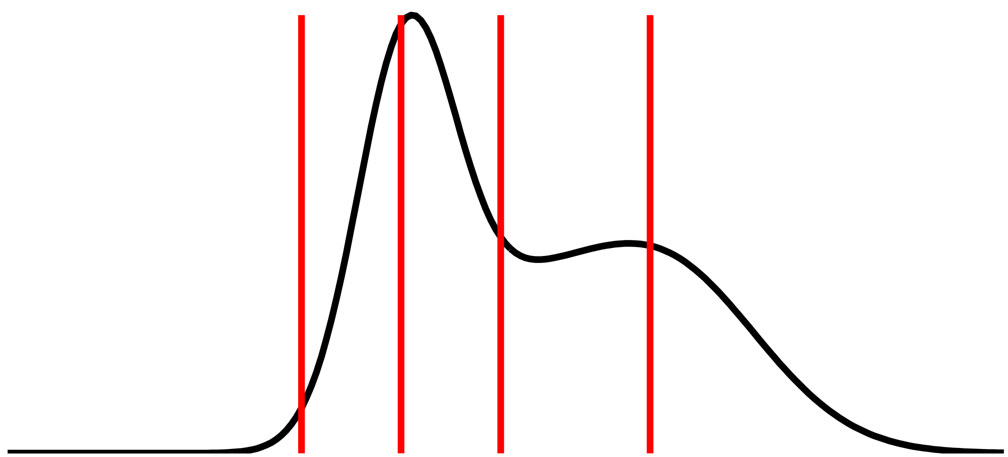
Motivation



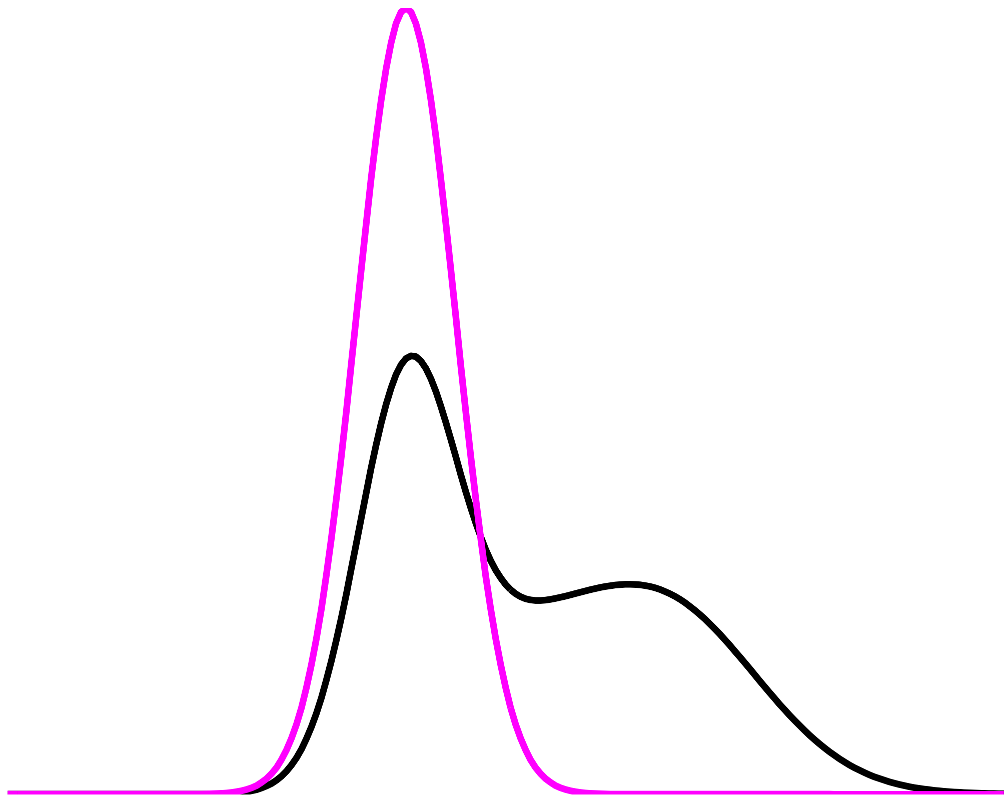
Motivation



Motivation



Motivation



Introduction

- Variational methods allow you to **side-step intractabilities in inference**
- The methods are justified as:
  - **fast** (compared to MCMC)
  - give back **uncertainty estimates** (unlike MAP)
  - extendable to **learning** (variational EM and variational Bayes)
  - optimises a **lower bound on the likelihood** (evidence)
- Today: Investigate the properties of variational algorithms for **Toy Gaussian Liner Dynamical Systems**

Take home message

1. Variational methods are **compact** (Well known: Mackay, 2003)
   - Mean-field **fails to propogate uncertainty information between time steps**
   - Can reduce mean field to an iterative MAP-like algorithm for finding the mean
   - Factored variational methods fall-over in the worst possible way: **When the approximation is a terrible one they become uber confident**

2. Variational methods are **biased**
   - Parameter estimates are often **very different from the maximum-likelihood solution**
   - The tightest approximation is not always the best for learning

All the theory you need to understand this talk

$\log p(Y|\theta)$

All the theory you need to understand this talk

$$\log p(Y|\theta) = \log \int dX p(Y, X|\theta)$$

All the theory you need to understand this talk

$$\log p(Y|\theta) = \log \int dX p(Y, X|\theta) \frac{q(X)}{q(X)},$$

All the theory you need to understand this talk

$$\log p(Y|\theta) = \log \int dX p(Y, X|\theta) \frac{q(X)}{q(X)},$$
$$\geq \int dX q(X) \log \frac{p(Y, X|\theta)}{q(X)}$$

All the theory you need to understand this talk

$$\log p(Y|\theta) = \log \int dX p(Y,X|\theta) \frac{q(X)}{q(X)},$$

$$\geq \int dX q(X) \log \frac{p(Y,X|\theta)}{q(X)} \qquad = F(q(X),\theta).$$

All the theory you need to understand this talk

$$\log p(Y|\theta) = \log \int dX p(Y, X|\theta) \frac{q(X)}{q(X)},$$

$$\geq \int dX q(X) \log \frac{p(Y|\theta)p(X|Y,\theta)}{q(X)} = F(q(X), \theta).$$

All the theory you need to understand this talk

$$\log p(Y|\theta) = \log \int dX p(Y,X|\theta) \frac{q(X)}{q(X)},$$

$$\geq \int dX q(X) \log \frac{p(Y|\theta) p(X|Y,\theta)}{q(X)} = F(q(X),\theta).$$

$$F(q(X),\theta) = \log p(Y|\theta) - \mathrm{KL}(q(X)||p(X|Y,\theta)),$$

All the theory you need to understand this talk

$$\log p(Y|\theta) = \log \int dX p(Y,X|\theta)\frac{q(X)}{q(X)},$$
$$\geq \int dX q(X) \log \frac{p(Y|\theta)p(X|Y,\theta)}{q(X)} = F(q(X),\theta).$$

$$F(q(X),\theta) = \log p(Y|\theta) - \mathrm{KL}(q(X)||p(X|Y,\theta)),$$

$$q(X) = \prod_{i=1}^{I} q_i(x_{C_i})$$

All the theory you need to understand this talk

$$\log p(Y|\theta) = \log \int dX p(Y,X|\theta) \frac{q(X)}{q(X)},$$
$$\geq \int dX q(X) \log \frac{p(Y|\theta)p(X|Y,\theta)}{q(X)} = F(q(X),\theta).$$

$$F(q(X),\theta) = \log p(Y|\theta) - \mathrm{KL}(q(X)||p(X|Y,\theta)),$$

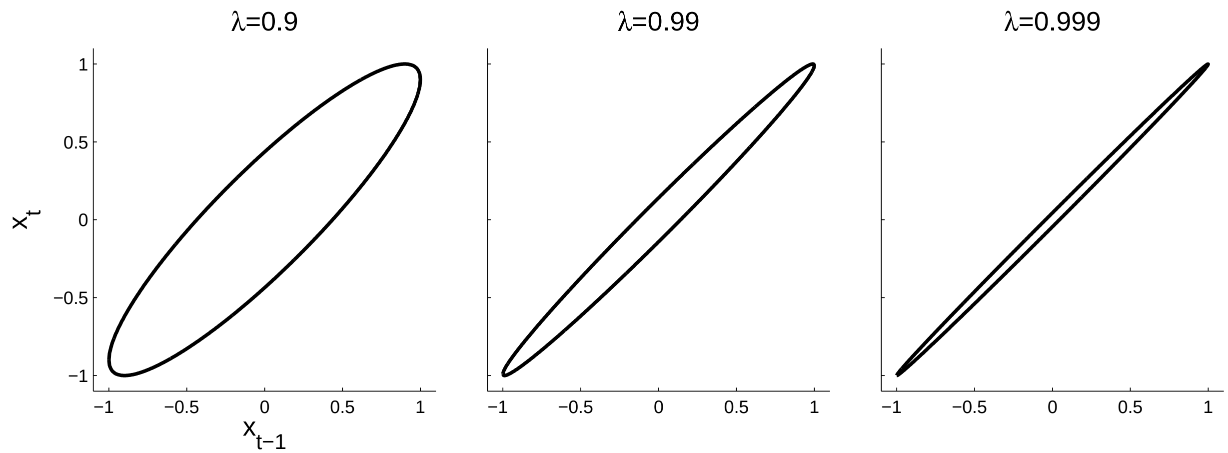
$$q(X) = \prod_{i=1}^{I} q_i(x_{C_i})$$
$$q(x_{C_i}) = \frac{1}{Z_i} \exp\left(\langle \log p(Y,X|\theta) \rangle_{\prod_{j\neq x_i} q(x_{C_j})} \right)$$

Example 1: Mean-field for inference in time-series models

Consider an AR(1) prior over latent variables and an arbitrary likelihood function

$$p(x_t|x_{t-1}) = \mathrm{Norm}(\lambda x_{t-1}, \sigma^2_{\mathrm{COND}})$$

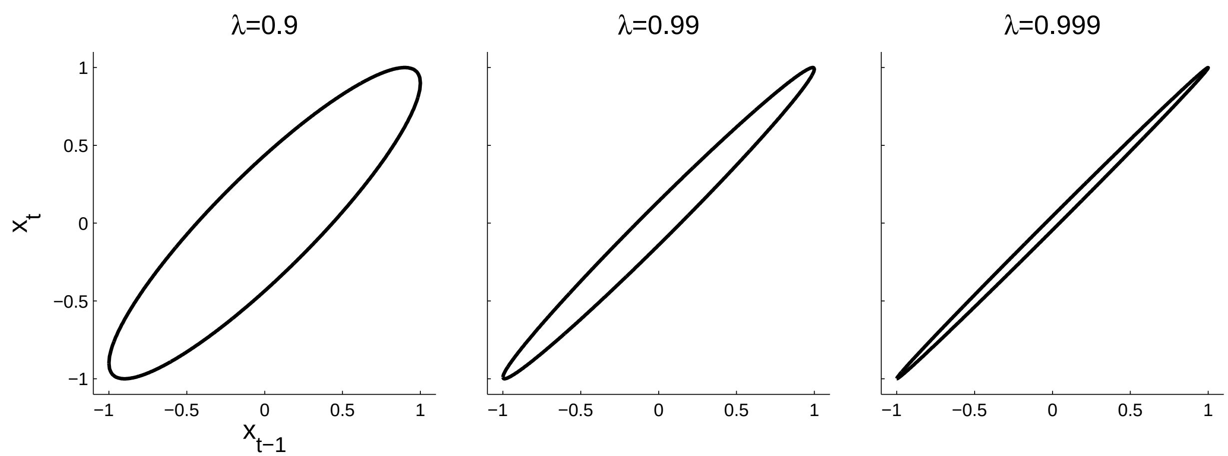
Point of time-series models is strong correlations; $\lambda \approx 1$ and $\sigma^2_{\mathrm{COND}} \approx 0$.



Marginal variance is $\frac{\sigma^2_{\mathrm{COND}}}{1-\lambda^2}$

Example 1: Mean-field for inference in time-series models

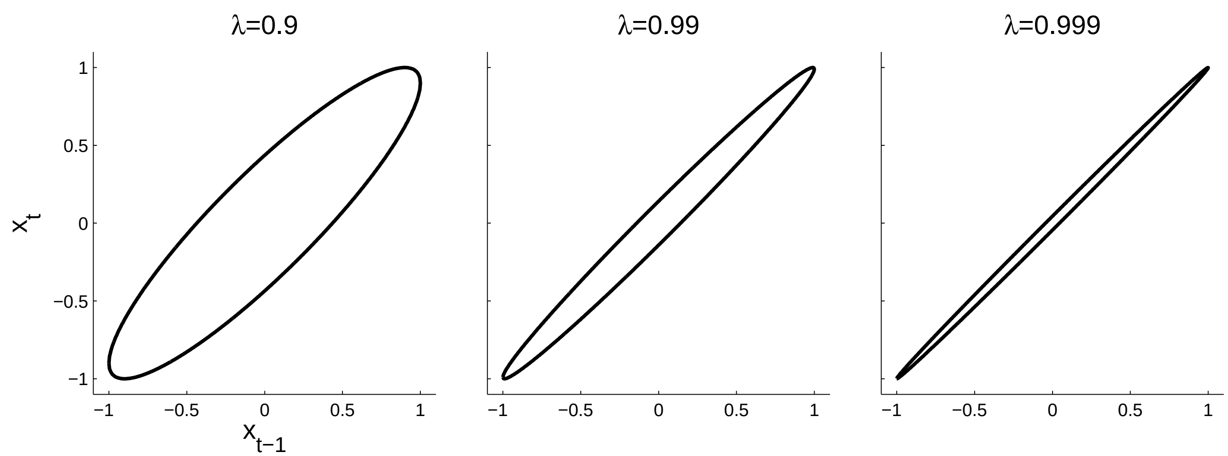
$$q(x_t) = \frac{1}{Z} p(y_t|x_t) \exp(\langle \log p(x_t|x_{t-1}) p(x_{t+1}|x_t) \rangle_{q(x_{t+1})q(x_{t-1})}),$$



Example 1: Mean-field for inference in time-series models

$$q(x_t) = \frac{1}{Z} p(y_t|x_t) \exp(\langle \log p(x_t|x_{t-1}) p(x_{t+1}|x_t) \rangle_{q(x_{t+1})q(x_{t-1})}),$$

$$= \frac{1}{Z} p(y_t|x_t) \mathrm{Norm}\left(\frac{\lambda}{1+\lambda^2} \left(\langle x_{t-1} \rangle_{q(x_{t-1})} + \langle x_{t+1} \rangle_{q(x_{t+1})} \right), \frac{\sigma^2_{\mathrm{COND}}}{1+\lambda^2} \right),$$

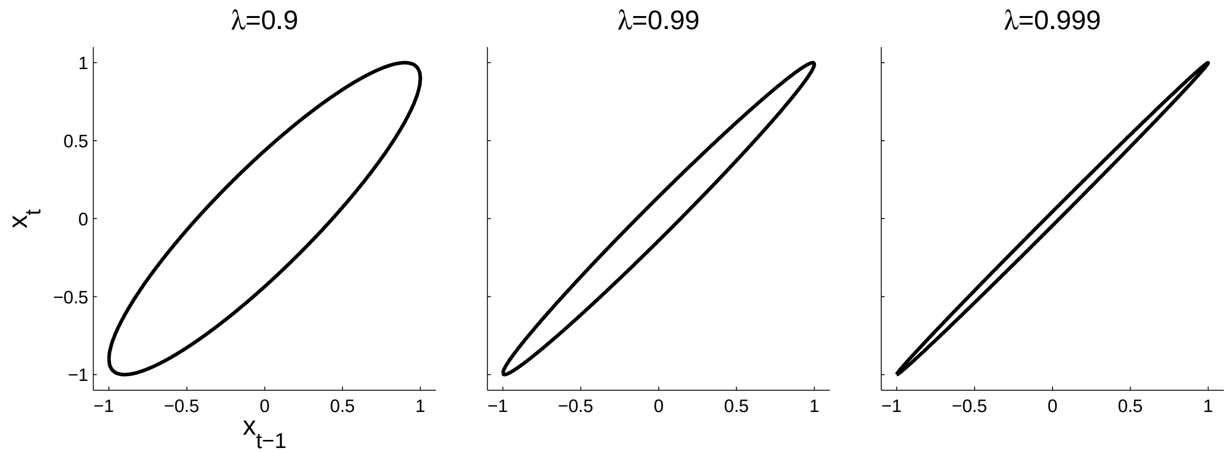


Example 1: Mean-field for inference in time-series models

$$q(x_t) = \frac{1}{Z} p(y_t|x_t) \exp(\langle \log p(x_t|x_{t-1}) p(x_{t+1}|x_t) \rangle_{q(x_{t+1})q(x_{t-1})}),$$

$$= \frac{1}{Z} p(y_t|x_t) \mathrm{Norm}\left(\frac{\lambda}{1+\lambda^2} \left(\langle x_{t-1} \rangle_{q(x_{t-1})} + \langle x_{t+1} \rangle_{q(x_{t+1})} \right), \frac{\sigma^2_{\mathrm{COND}}}{1+\lambda^2} \right),$$

$$= \frac{1}{Z} p(y_t|x_t) q_{\mathrm{prior}}(x_t)$$

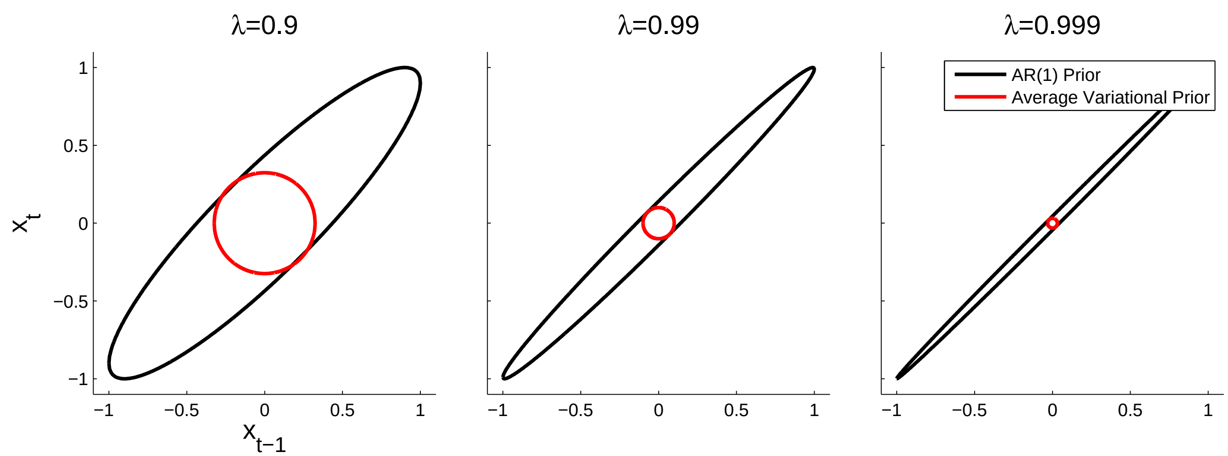


Example 1: Mean-field for inference in time-series models

$$q(x_t) = \frac{1}{Z} p(y_t|x_t) \exp(\langle \log p(x_t|x_{t-1}) p(x_{t+1}|x_t) \rangle_{q(x_{t+1})q(x_{t-1})}),$$

$$= \frac{1}{Z} p(y_t|x_t) \mathrm{Norm}\left(\frac{\lambda}{1+\lambda^2} \left(\langle x_{t-1} \rangle_{q(x_{t-1})} + \langle x_{t+1} \rangle_{q(x_{t+1})} \right), \frac{\sigma^2_{\mathrm{COND}}}{1+\lambda^2} \right),$$

$$= \frac{1}{Z} p(y_t|x_t) q_{\mathrm{prior}}(x_t)$$



Point of time-series models: Large observation noise (wide $p(y_t|x_t)$)

Example 1: Summary

- Mean-field = the field you would see if all the other variables took their mean values.
- Variational prior is identical to the inference we'd make if we **knew the adjacent latent variables**: the uncertainty in them is not folded in
- Temporally factored variational approximations for time series are **narrower than the conditional** (which is very concentrated)
- Uncertainties are **meaningless** (compared to the true marginals)
- When variational approximations are **least accurate**, they are at their **most confident**

Learning: What do you want out of a variational method?

- **Learning parameters is important** e.g. in scientific enquiry (how sparse are sounds, how slow are natural scenes)

- What makes for a good variational approximation in this case?

$$F(q(X), \theta) = \log p(Y|\theta) - \mathrm{KL}(q(X)||p(X|Y, \theta))$$

- Instant reaction: Want the KL to be as **tight as possible, everywhere**.

- Not necessarily the case: Better to be **equally tight everywhere**.

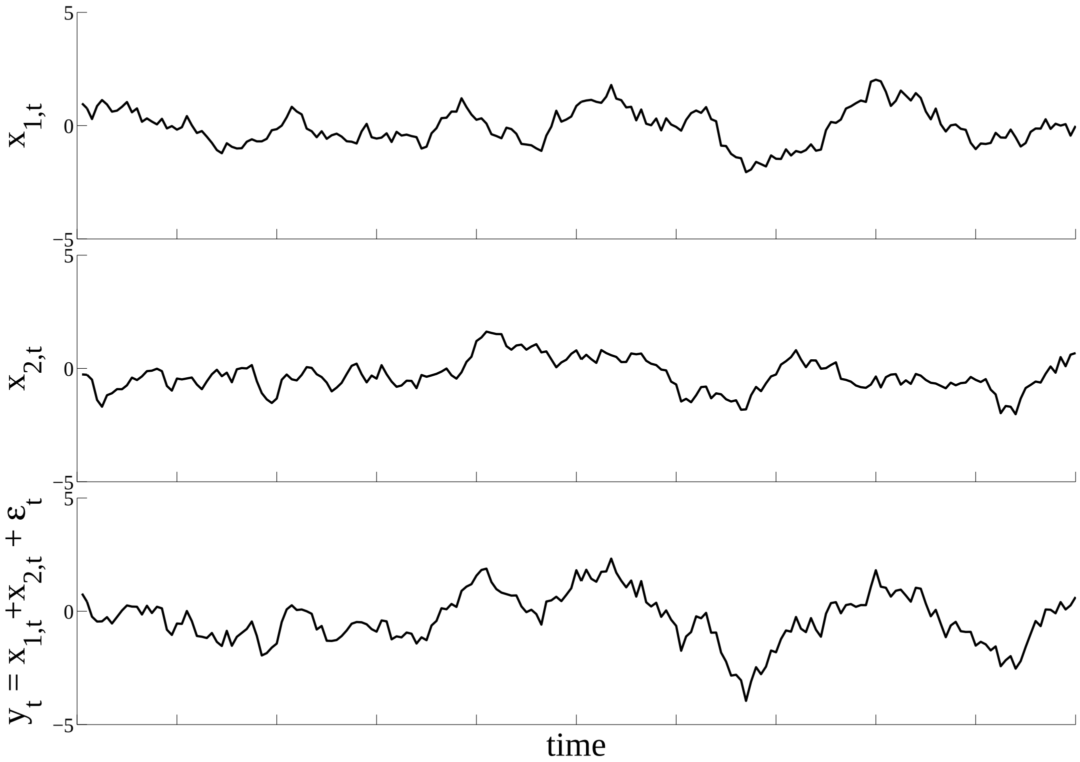
- We show:
  - The KL can be strongly parameter dependent and bias learning to regions where the bound is tight, rather than the likelihood large.
  - Mean-field can out-perform more structured approximations as its bound is less parameter dependent

Example 2: Structured approximations for time-series

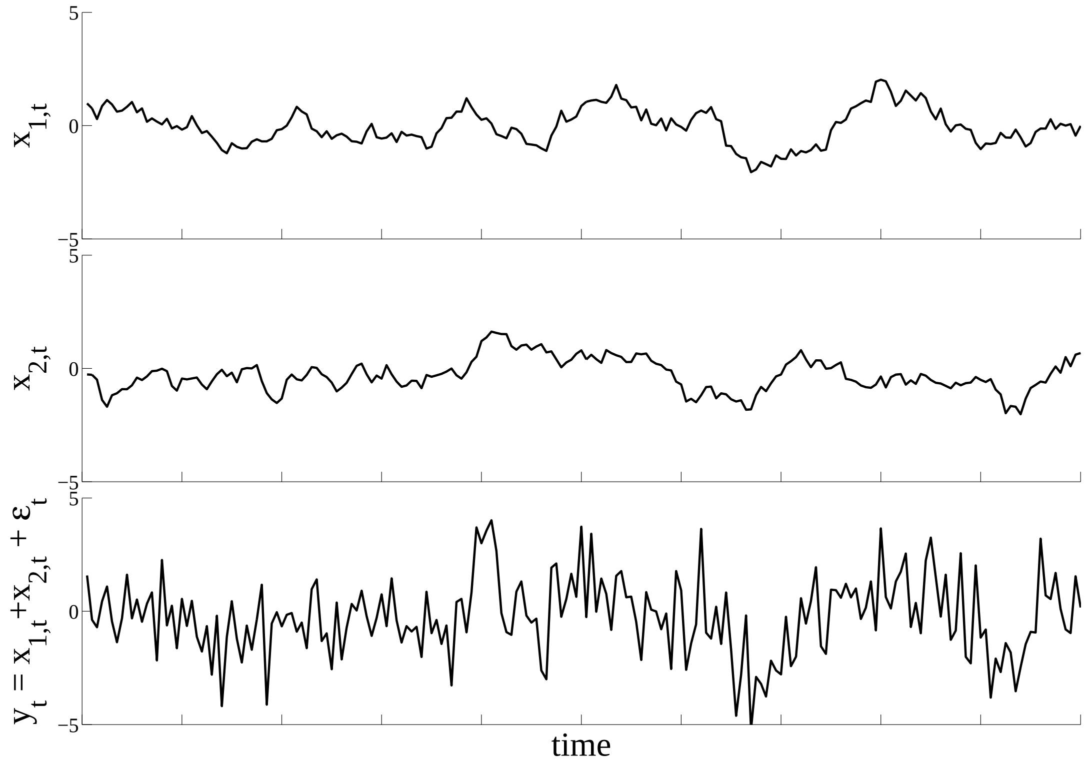
- Simplest possible **linear Gaussian state-space model** (two latent chains and two-time steps)

$$
\begin{aligned}
p(x_{k,1}) &= \mathrm{Norm}\left(0, \frac{\sigma_x^2}{1-\lambda^2}\right)\\
p(x_{k,2}|x_{k,1}) &= \mathrm{Norm}\left(\lambda x_{k,1}, \sigma_x^2\right)\\
p(y_t|x_{1,t}, x_{2,t}) &= \mathrm{Norm}(x_{1t} + x_{2t}, \sigma_y^2)
\end{aligned}
$$

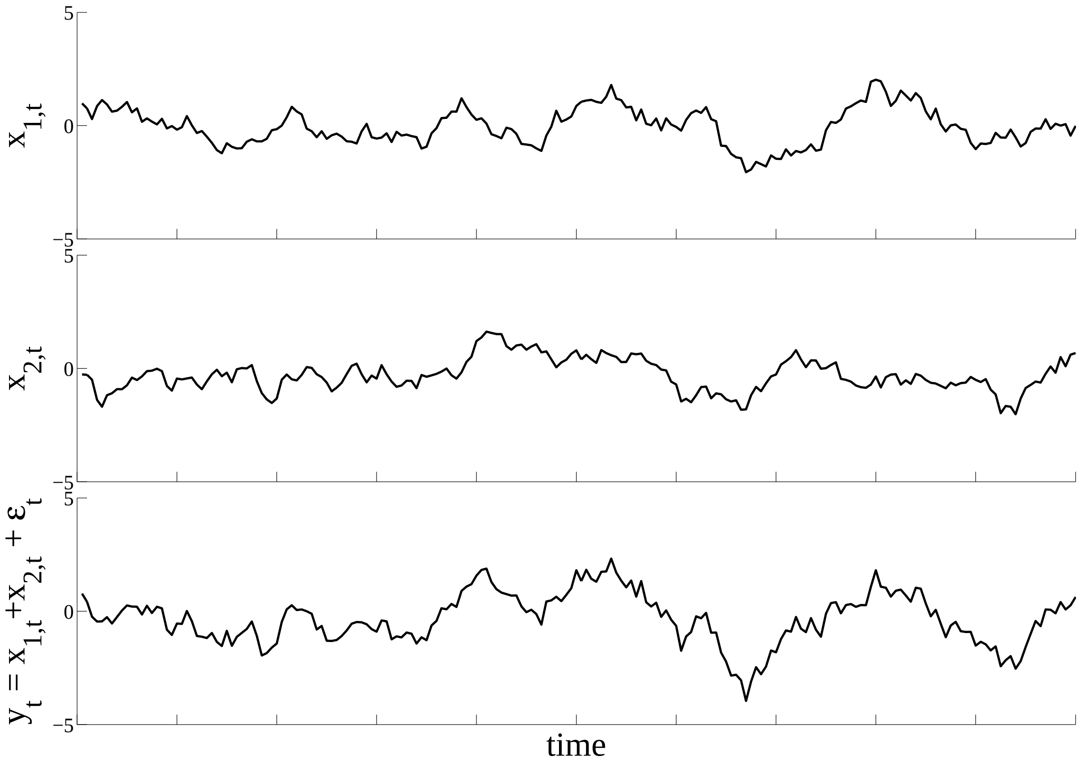
Example 2: Sample from the model



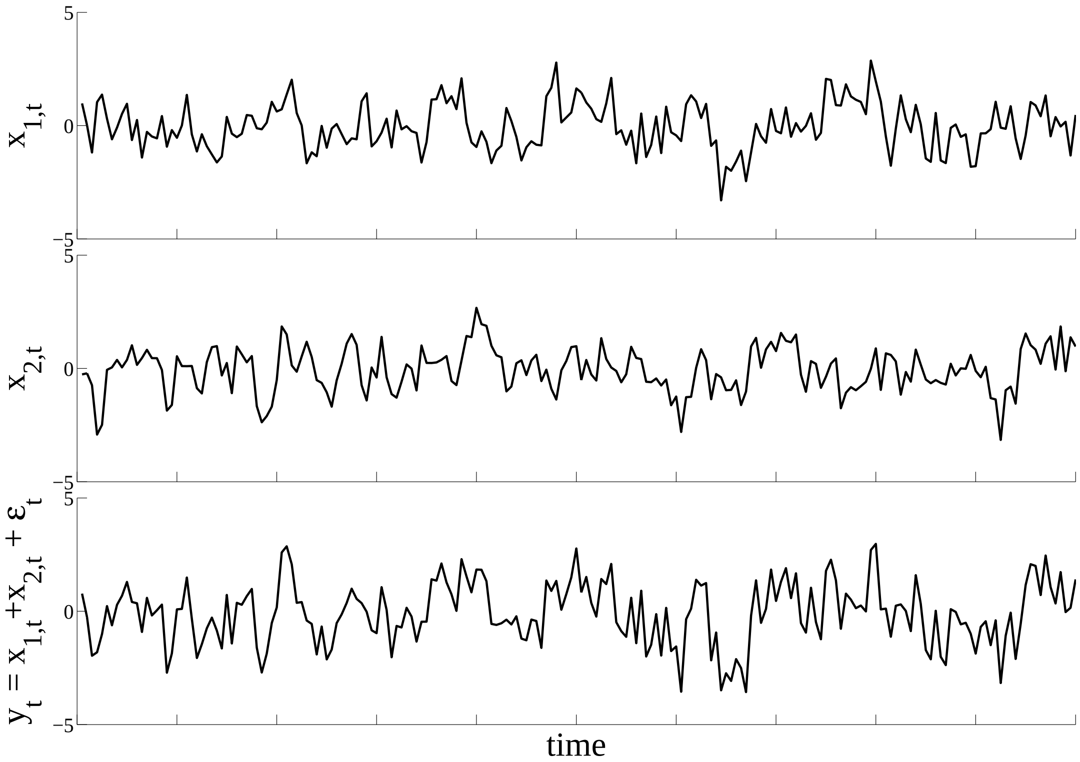
Example 2: Larger observation noise



Example 2: Original



Example 2: Faster dynamics



Example 2: Structured approximations for time-series

- Simplest possible **linear Gaussian state-space model** (two latent chains and two-time steps)

$$p(x_{k,1}) = \mathrm{Norm}\left(0, \frac{\sigma_x^2}{1-\lambda^2}\right)$$
$$p(x_{k,2}|x_{k,1}) = \mathrm{Norm}\left(\lambda x_{k,1}, \sigma_x^2\right)$$
$$p(y_t|x_{1,t}, x_{2,t}) = \mathrm{Norm}(x_{1t} + x_{2t}, \sigma_y^2)$$

- Joint distribution $p(X, Y)$, probability of the data $p(Y)$ and posterior distribution over hidden variables $p(X|Y)$ are all Gaussian.

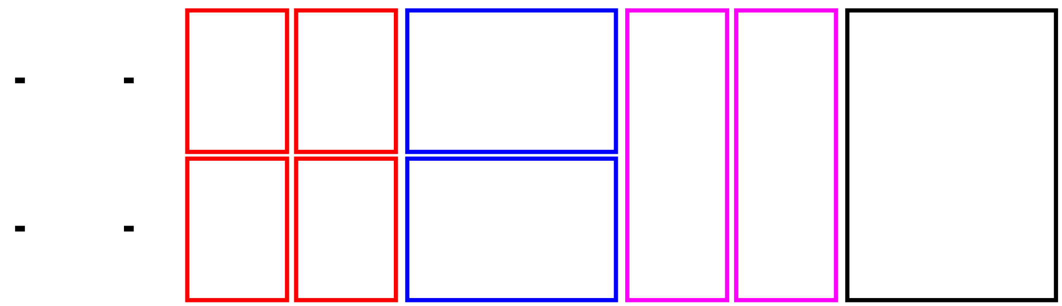
- Posterior is correlated
  - through time (due to the slowness-prior; more so as $|\lambda| \to 1$ and $\sigma_x^2 \to 0$),
  - across chains (explaining away; more so as $\sigma_y^2 \to 0$)

Example 2: Structured approximations for time-series

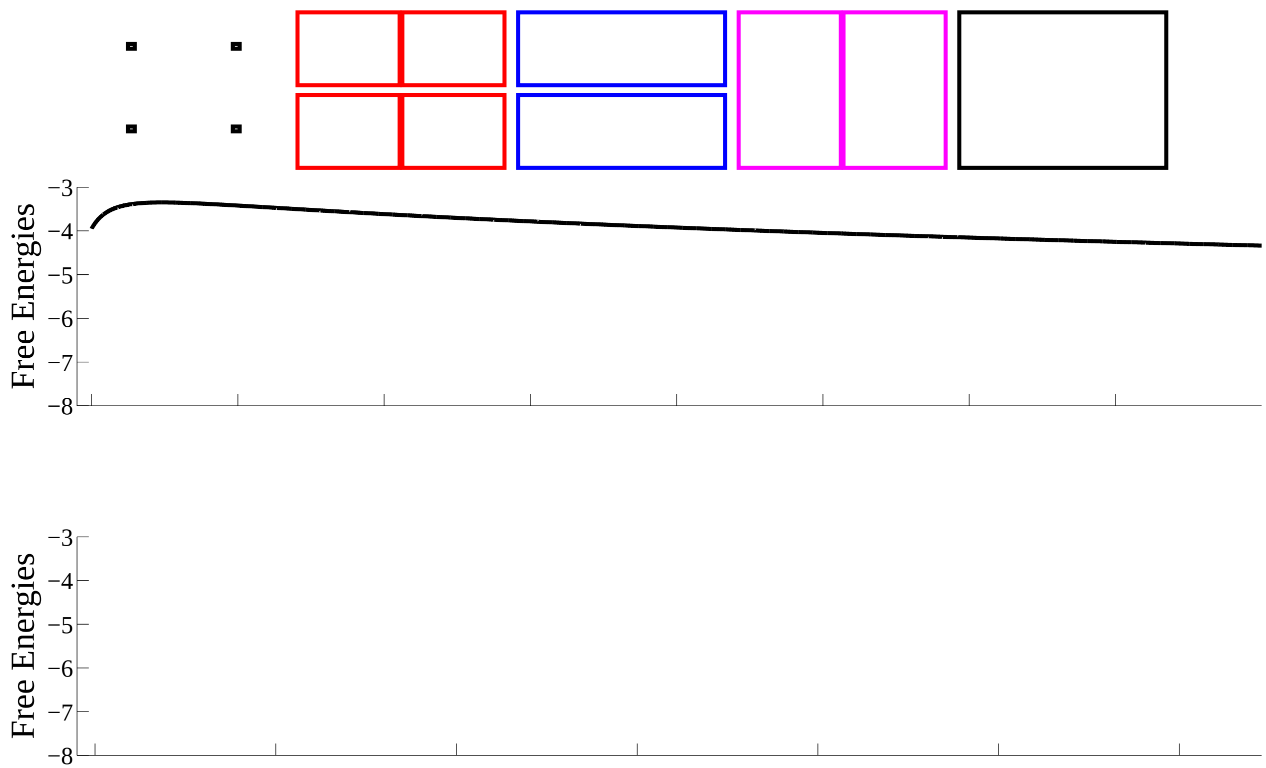
Four approx schemes: mean-field, chain-factored, temporally-factored, and MAP

	unfactored over chains	factored over chains
unfactored over time	$p(\mathbf{x}\vert\mathrm{y}) = q(x_{11}, x_{12}, x_{21}, x_{22})$	$q_2(\mathbf{x}) = q_{21}(x_{11}, x_{12})q_{22}(x_{21}, x_{22})$
factored over time	$q_3(\mathbf{x}) = q_{31}(x_{11}, x_{21})q_{32}(x_{12}, x_{22})$	$q_1(\mathbf{x}) = q_{11}(x_1)q_{12}(x_2)q_{13}(x_3)q_{14}(x_4)$

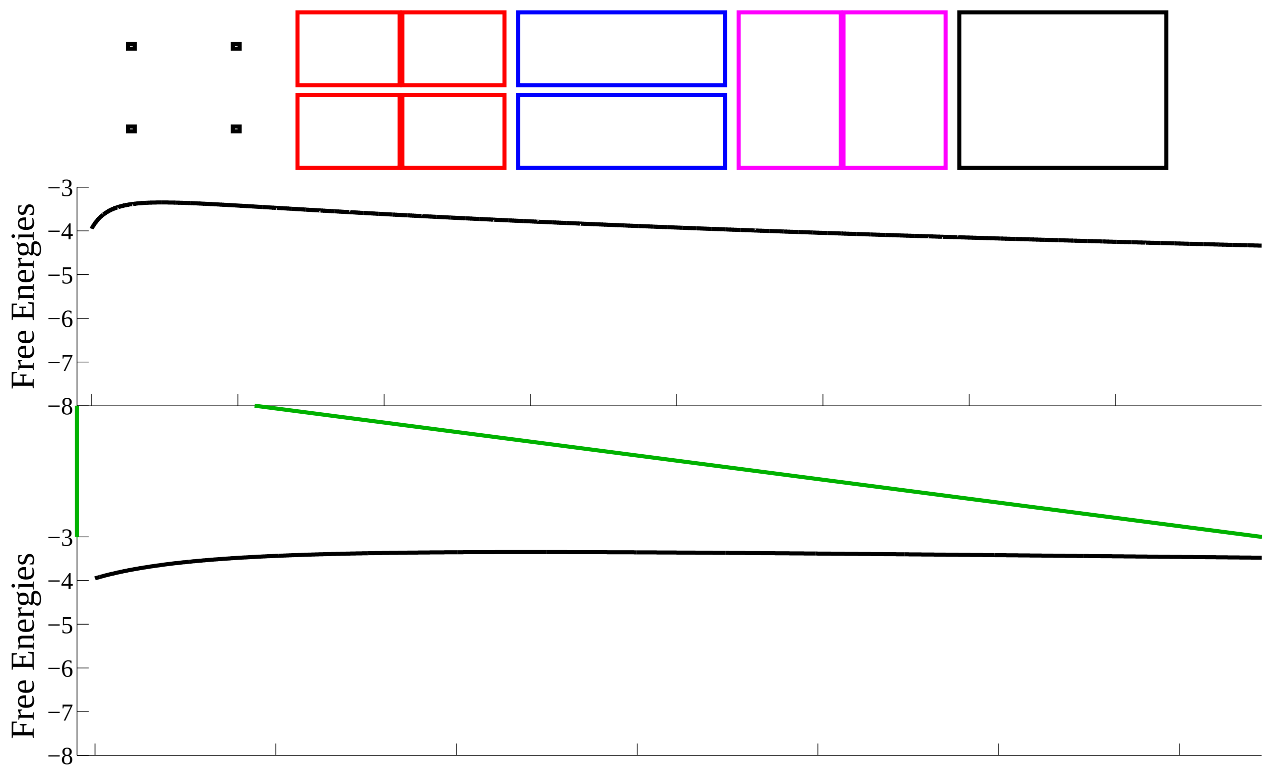
q_{ij} Gaussian with a mean and precision matching elements in $\mu_{\mathbf{x}|\mathrm{y}}$ and $\Sigma^{-1}_{\mathbf{x}|\mathrm{y}}$.



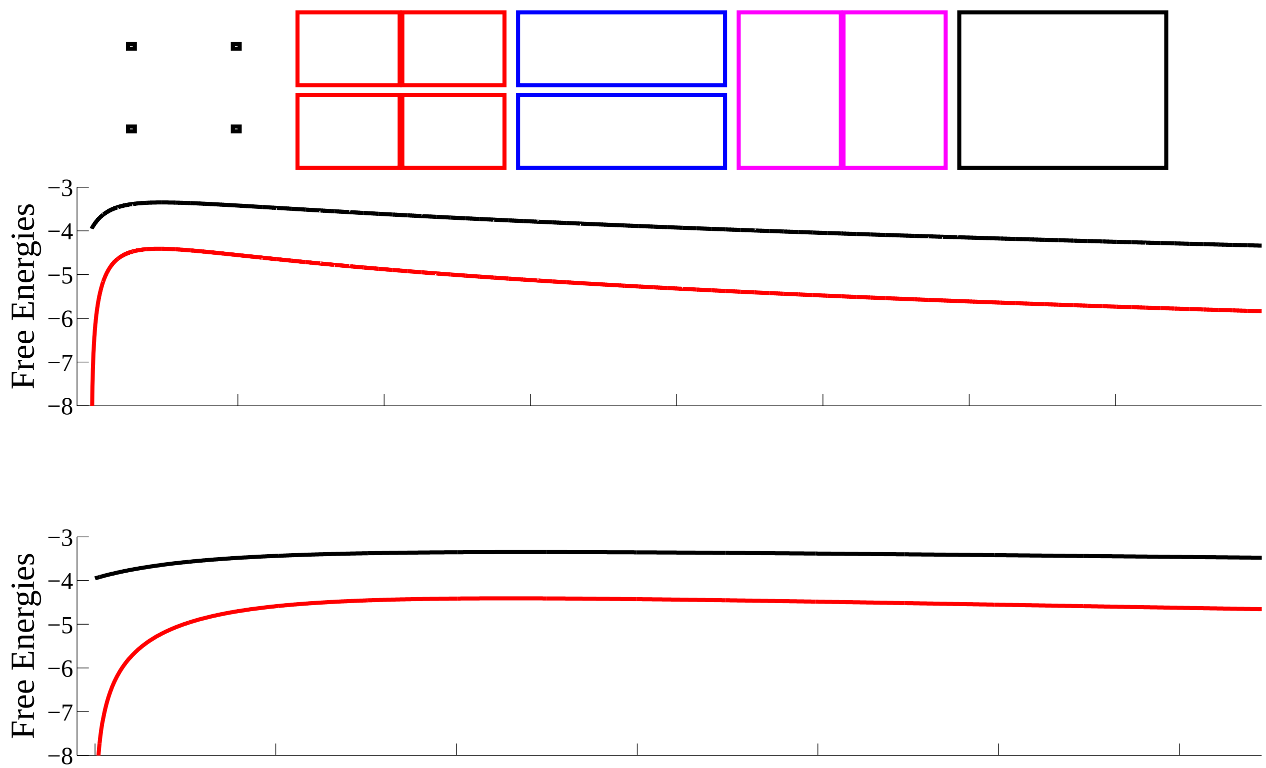
Example 2: Learning σ_y^2, Likelihood



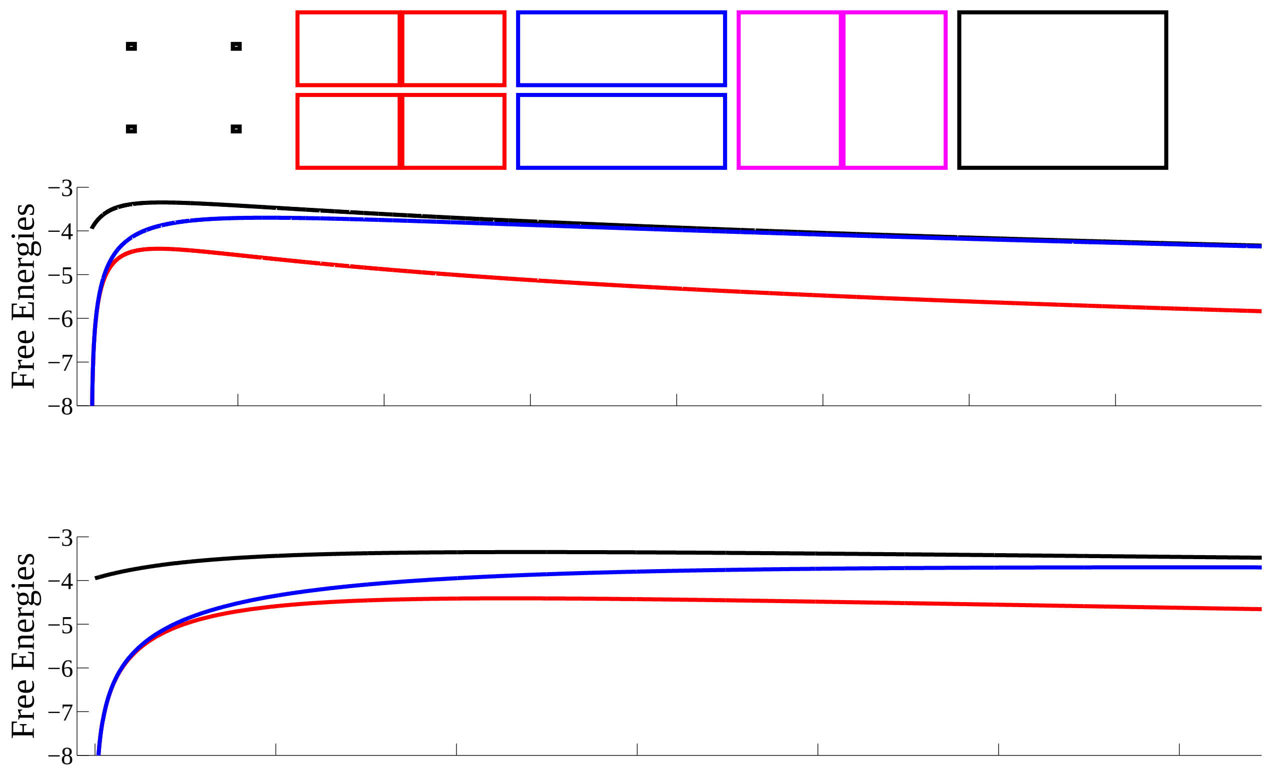
Example 2: Learning σ_y^2, Likelihood



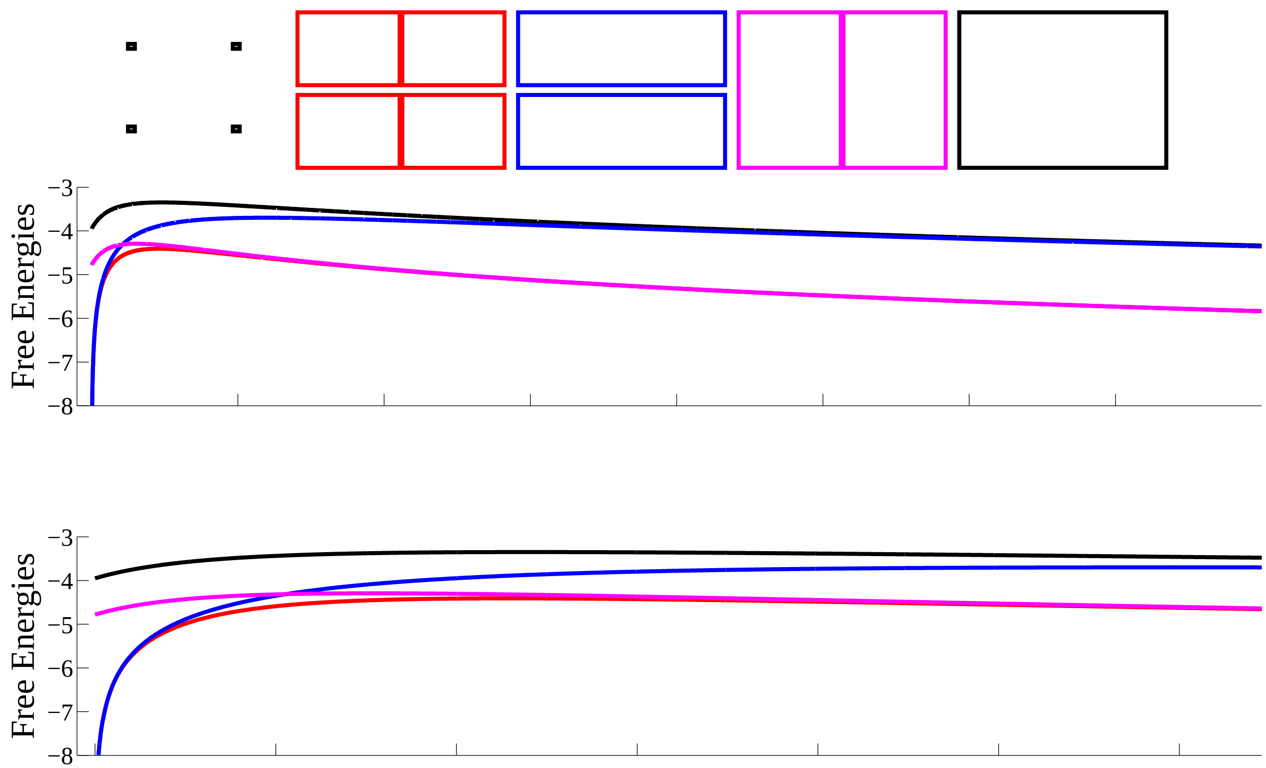
Example 2: Learning σ_y^2, Mean-field



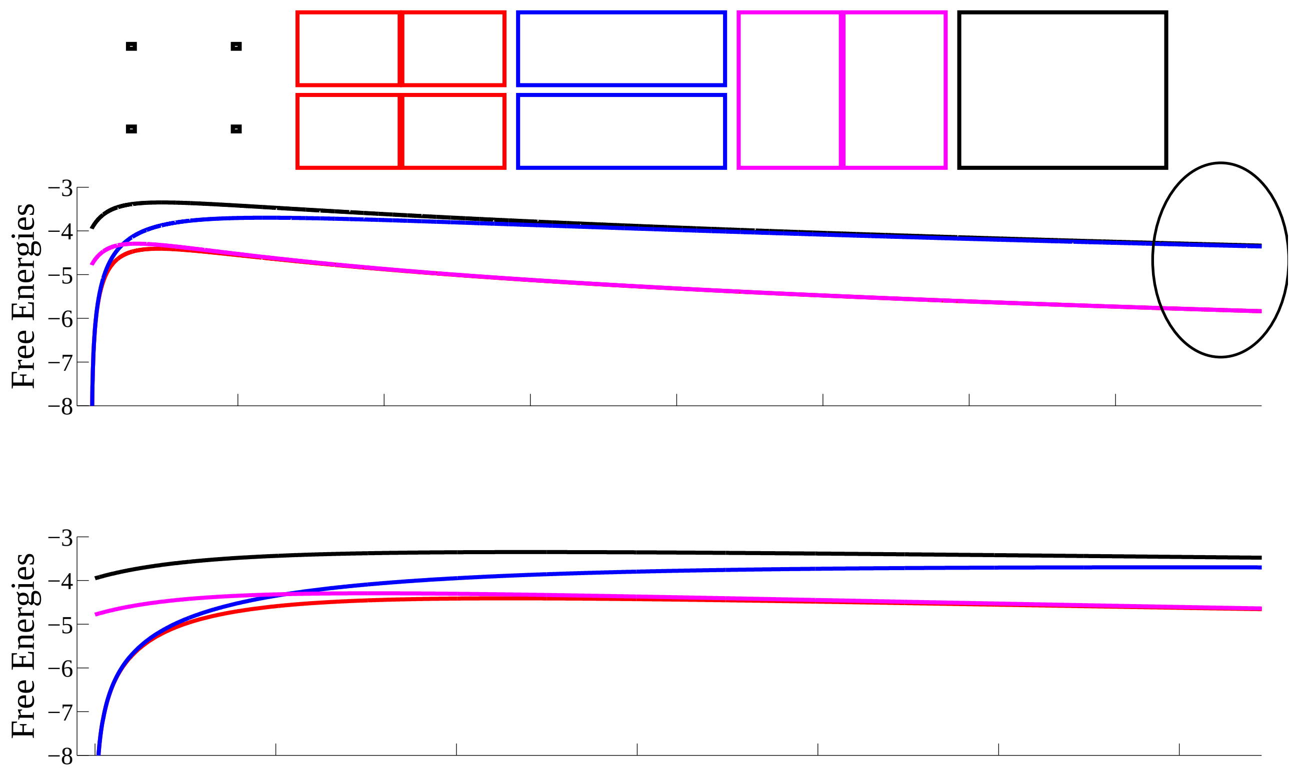
Example 2: Learning σ_y^2, unfactored over time



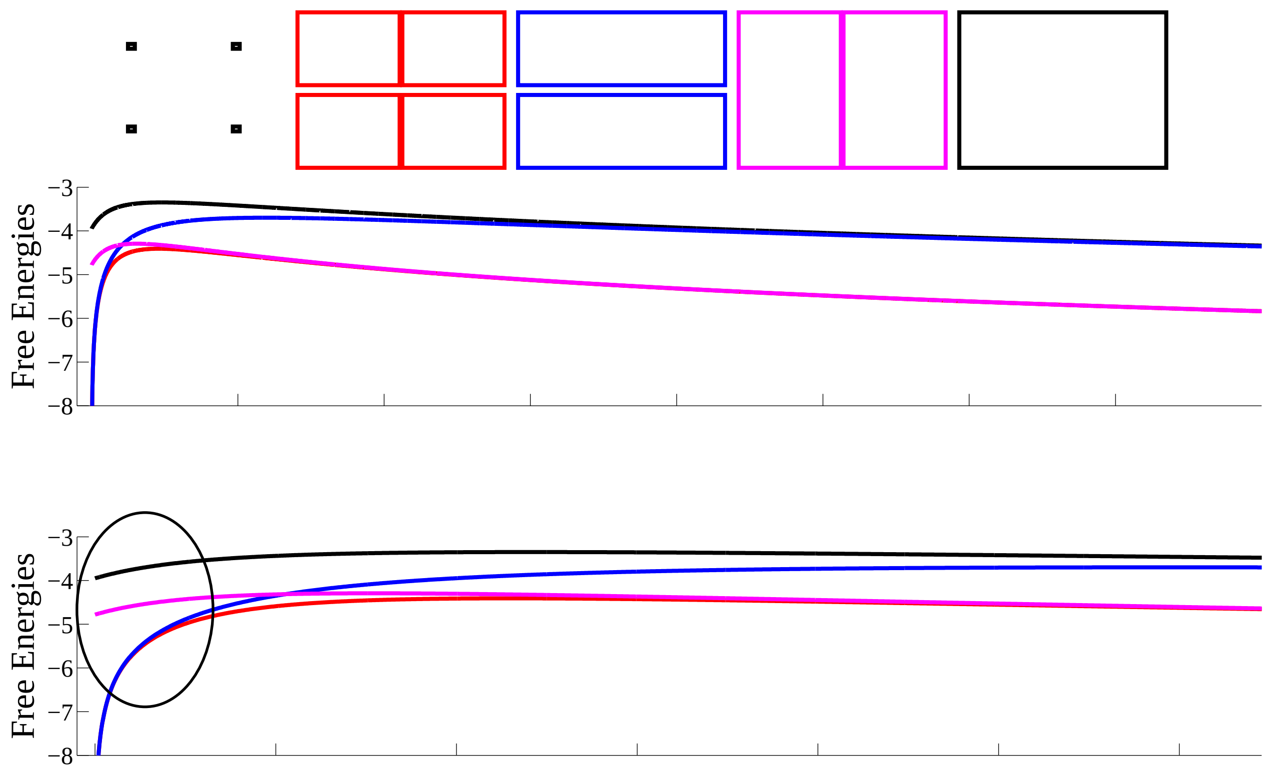
Example 2: Learning σ_y^2, unfactored over chains



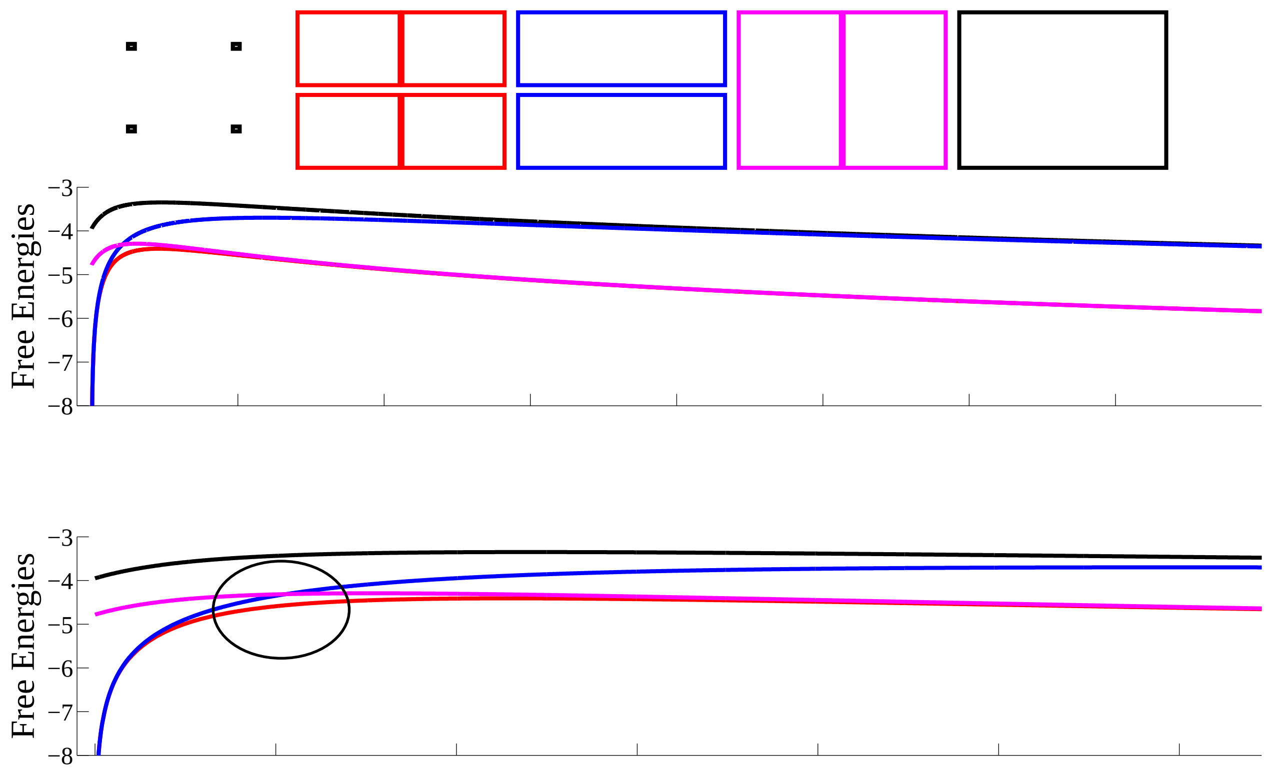
Example 2: Learning σ_y^2, high noise behaviour



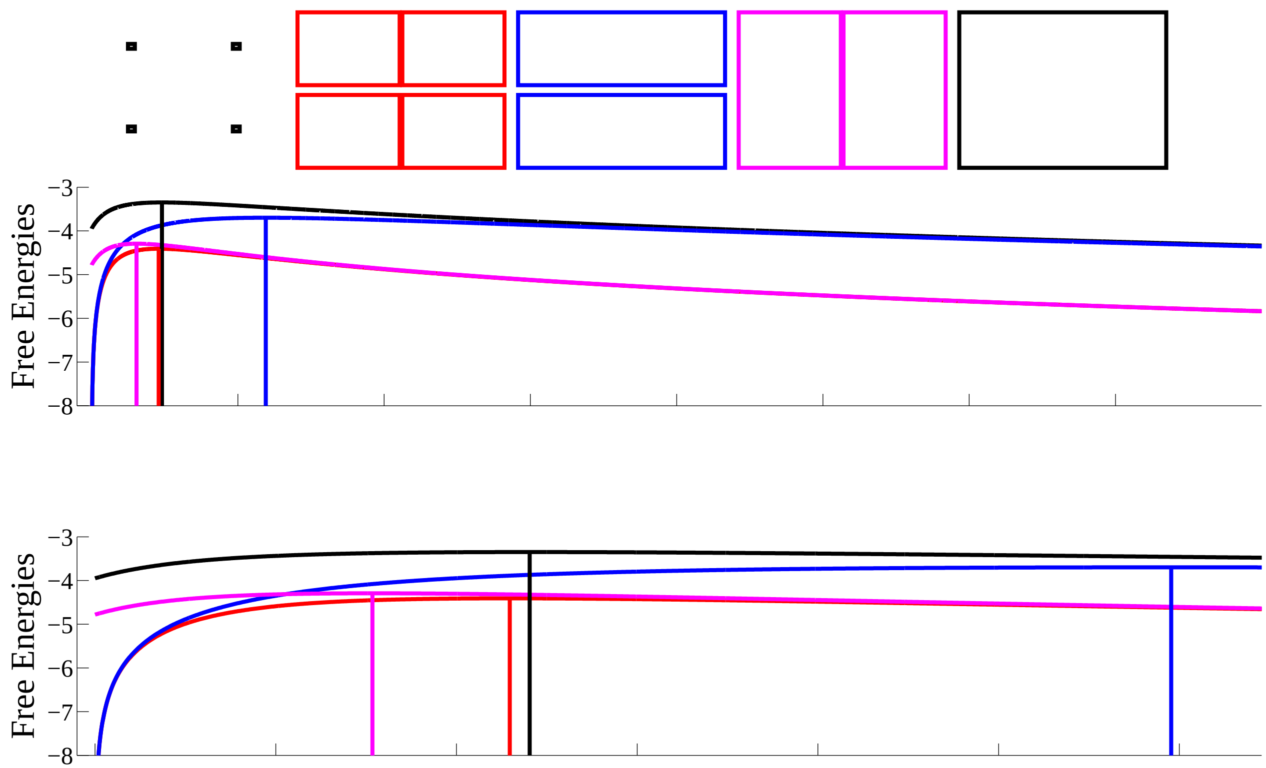
Example 2: Learning σ_y^2, low noise behaviour



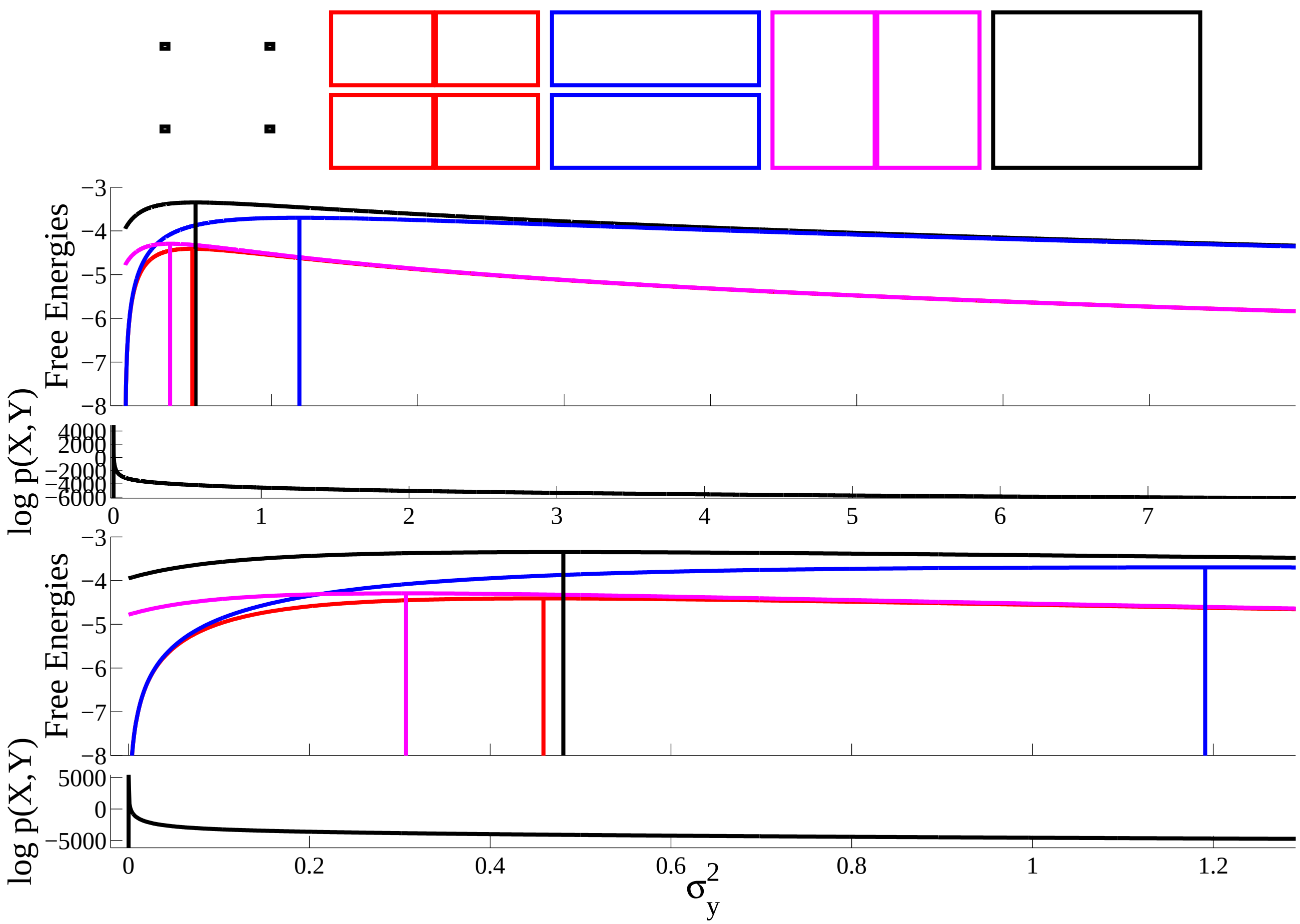
Example 2: Crossing point $\sigma_y^2 = \sigma_x^2/|\lambda|$



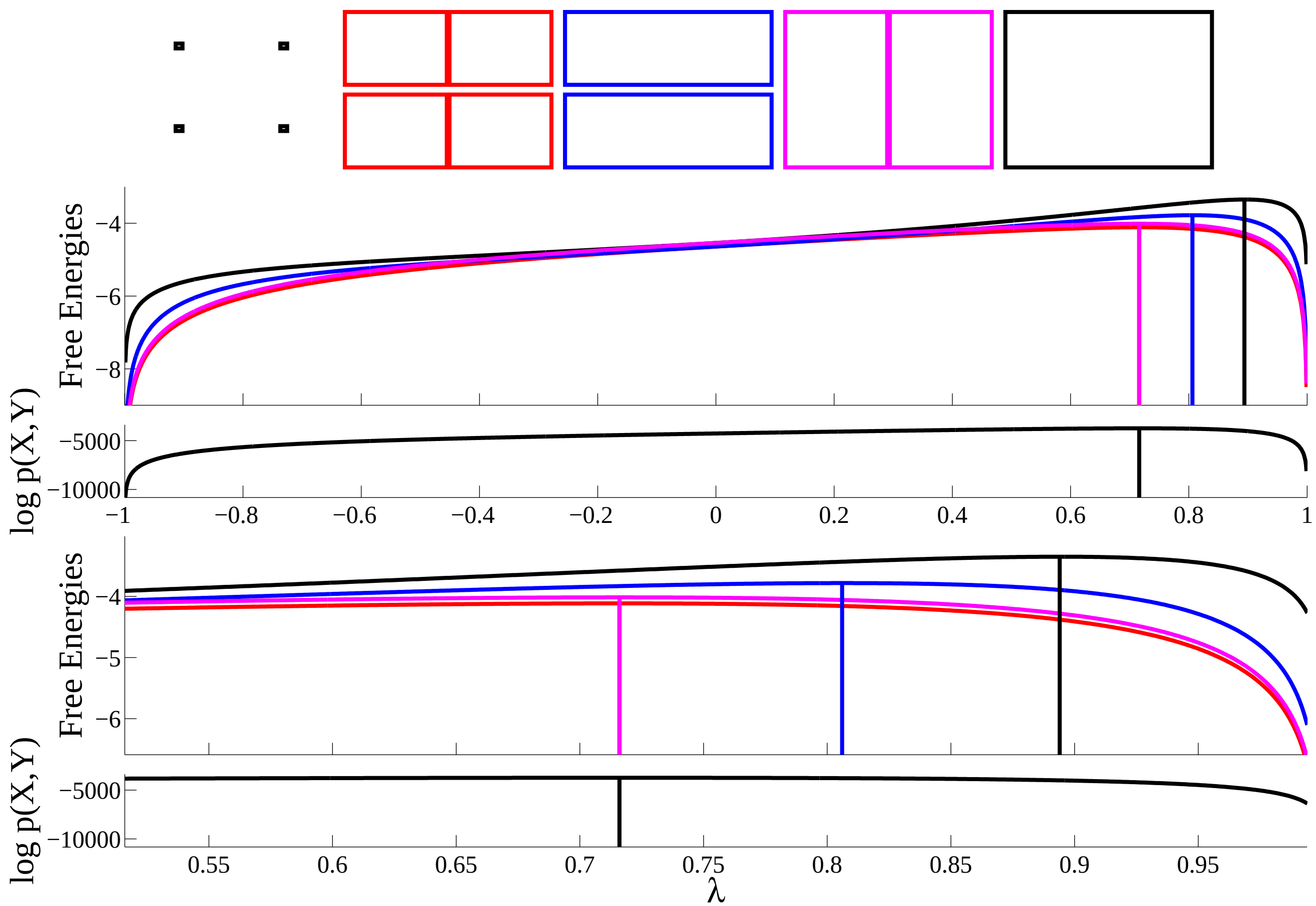
Example 2: Learning σ_y^2, Maxima



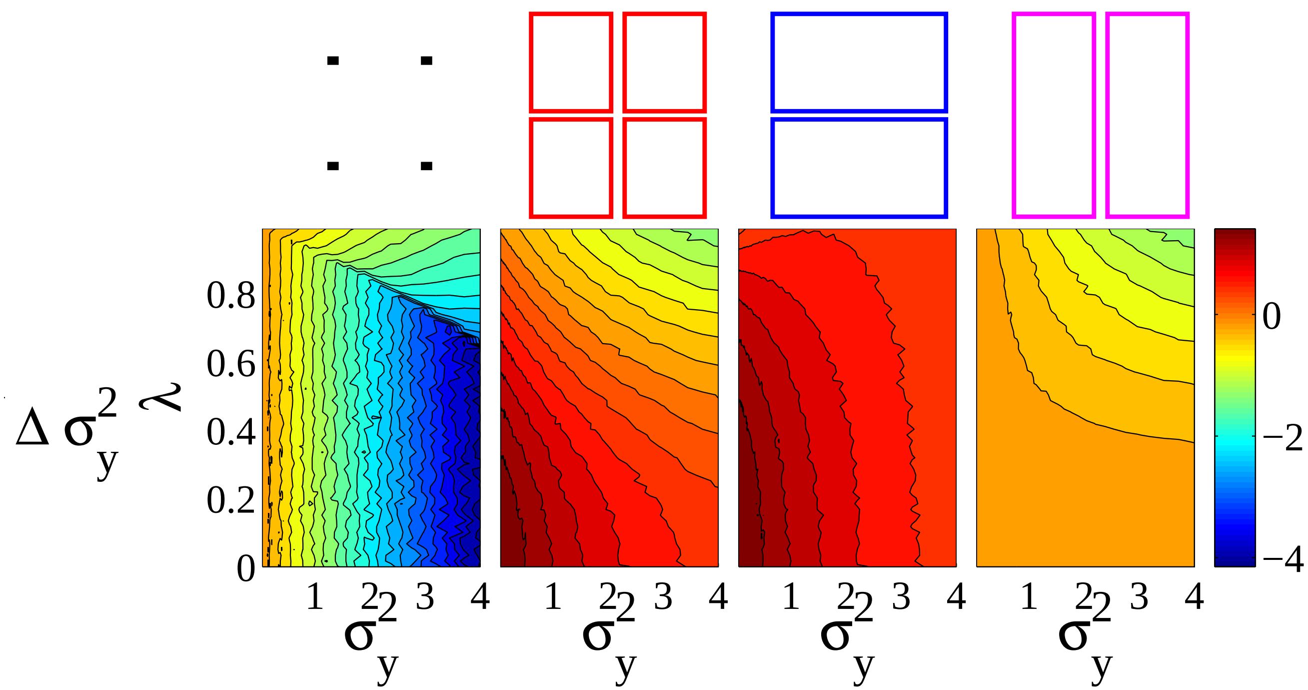
Example 2: Learning σ_y^2, MAP solution



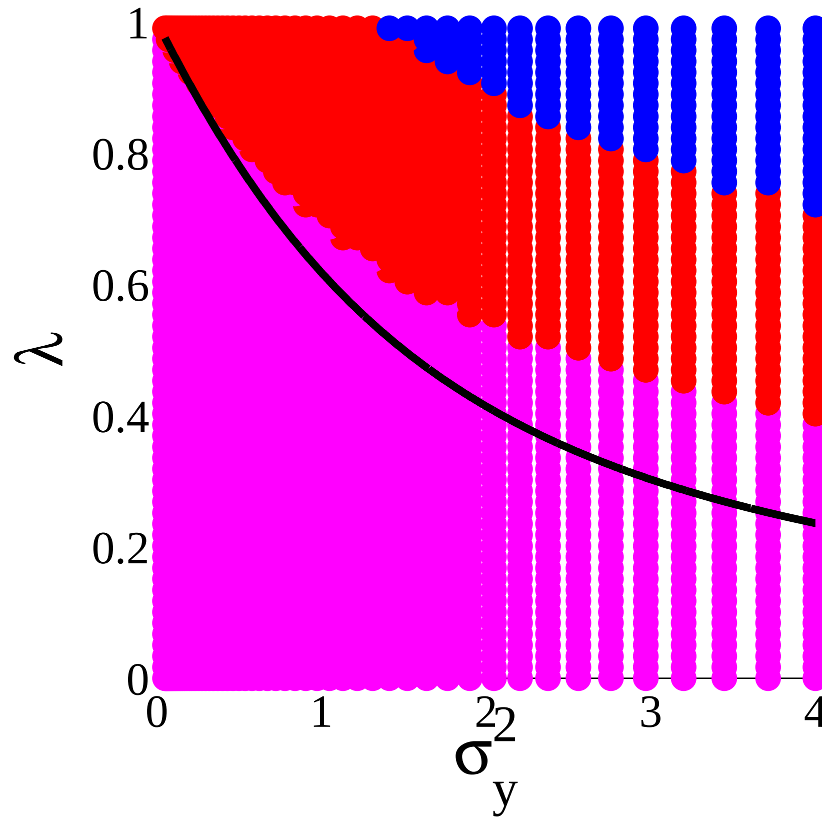
Example 2: Learning λ



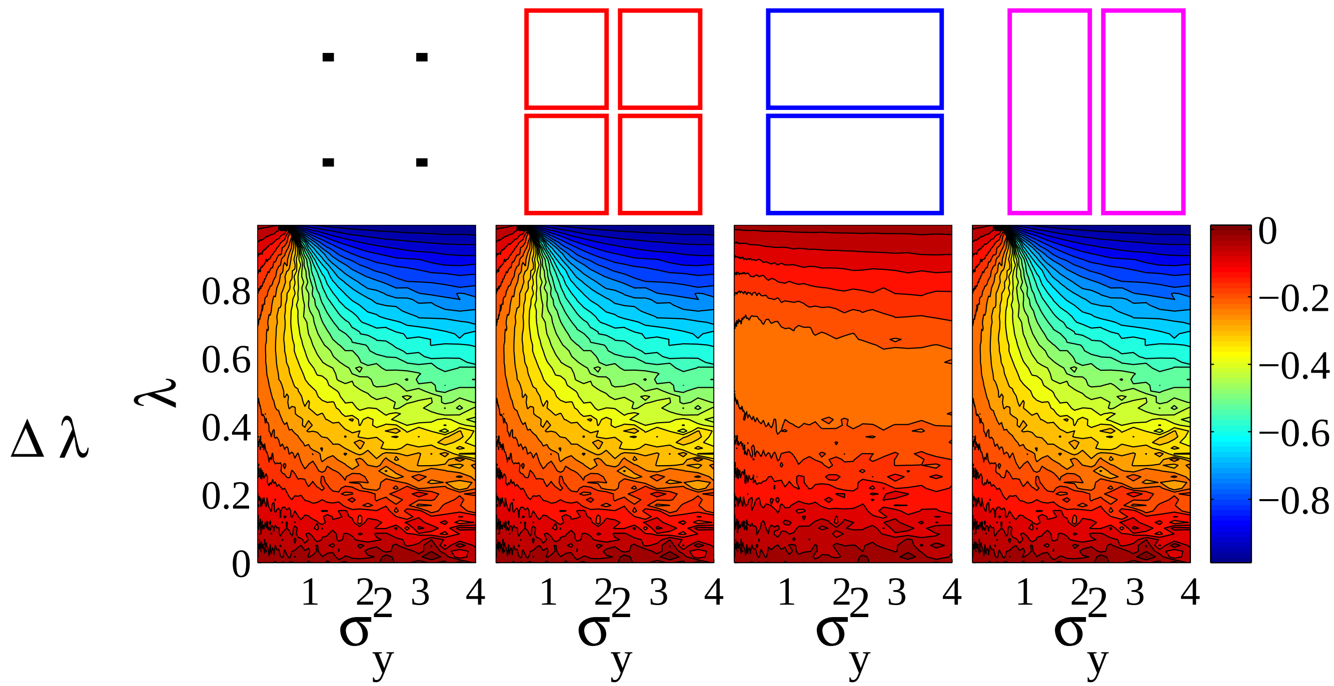
Example 2: Biases σ_y^2



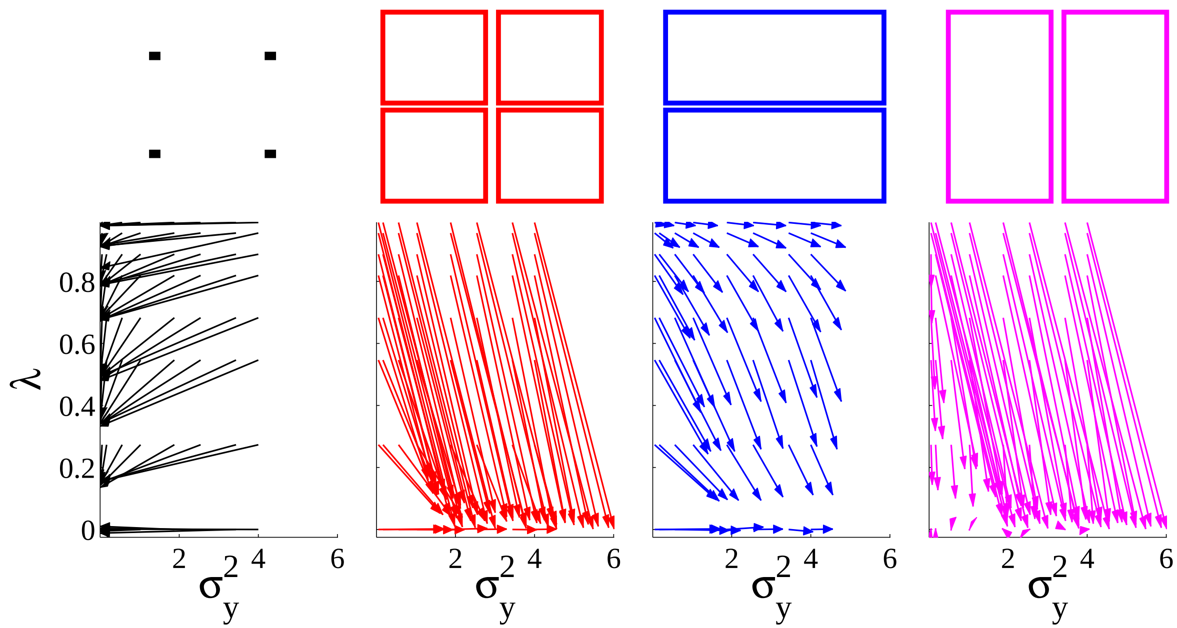
Example 2: Best Approximation



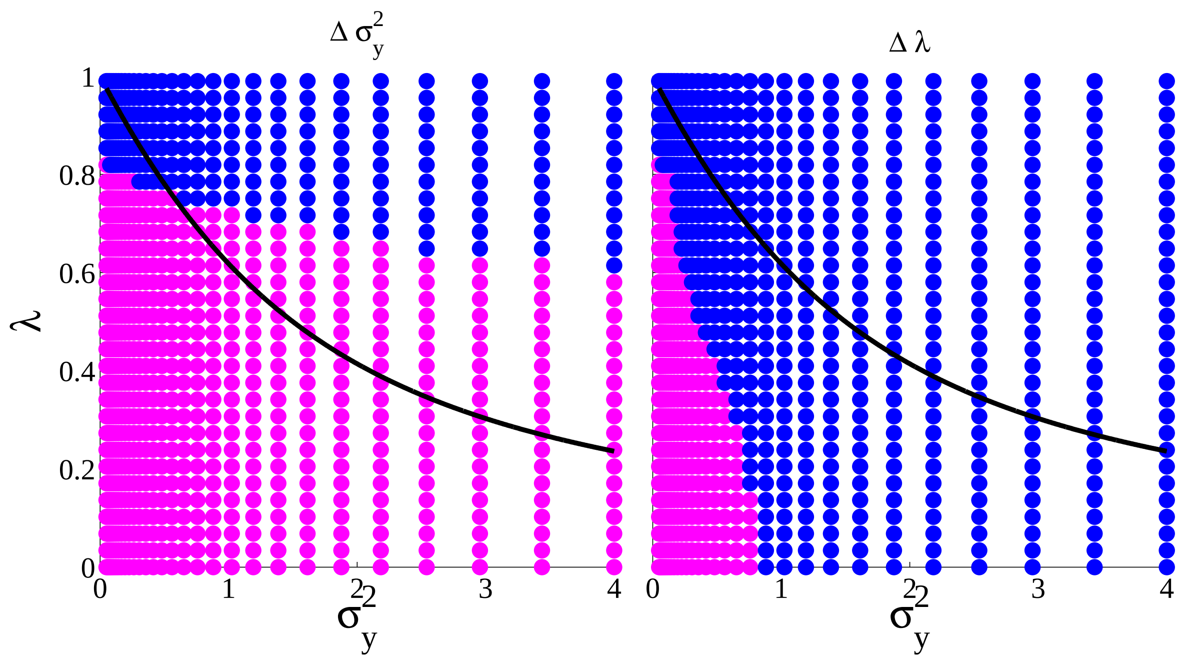
Example 2: Biases λ



Example 2: Inferring two parameters



Example 2: Best Approximation



Example 2: Summary

- The **parameter dependent bias is often very large**.
- Mean field methods can out-perform more structured approximations
- MAP methods can out-perform variational methods
- **Different structural approximations are better at determining different parameters and tightness is not a brilliant indicator**

Take home message

1. Variational methods are **compact** (Well known: Mackay, 2003)

   - Mean-field **fails to propogate uncertainty information between time steps**
   - Can reduce mean field to an iterative MAP-like algorithm for finding the mean
   - Factored variational methods fall-over in the worst possible way: **When the approximation is a terrible one they become uber confident**

2. Variational methods are **biased**

   - Parameter estimates are often **very different from the maximum-likelihood solution**
   - The tightest approximation is not always the best for learning