Bayesian Optimization of Hyperparameters when the Marginal likelihood is Estimated by MCMC

Oskar Gustafsson, **Mattias Villani** and Pär Stockhammar

Department of Statistics
Stockholm University

Department of Computer and Information Science
Linköping University

Overview

- Hyperparameter inference from the marginal likelihood
- Gaussian processes and Bayesian Optimization
- Bayesian Optimization with Optimized Precision (BOOP)
- Application to Bayesian VARs with steady-state priors.

- Slides: http://mattiasvillani.com/news.

Hyperparameter inference

- Parameter/Hyperparameter distinction:
  - **Parameters**, β, typically high-dim.
  - **Hyperparameters** θ, typically low-dim.
- **Deep neural networks**
  - β are the weights and biases
  - θ is the network architecture (no. of layers, learning rate etc).
- **Gaussian process models** (e.g. spatial)
  - β is the random field
  - θ are the kernel hyperparameters (length scale etc).

Hyperparameter inference

- DSGE models in econometrics
  - $\boldsymbol{\beta}$ are the persistence and variance of shocks etc
  - $\boldsymbol{\theta}$ are parameters in the steady state.

- Bayesian **vector autoregressive models** (**VAR**) models

$$y_t = \mu + \sum_{k=1}^{K} \Pi_k (y_{t-k} - \mu) + \varepsilon_t, \quad \varepsilon_t \overset{\text{iid}}{\sim} N(0, \Sigma)$$

  - $\boldsymbol{\beta} = (\mu, \Pi_1, \ldots, \Pi_K, \Sigma)$
  - $\boldsymbol{\theta} = (\lambda_1, \lambda_2, \lambda_3)$ are hyperparameters determining prior stdev:

$$\mathbb{S}(\pi_{ij}^{(k)}) = \begin{cases} \frac{\lambda_1}{k^{\lambda_3}} & \text{own lags } i = j \\ \frac{\lambda_1 \lambda_2}{k^{\lambda_3}} & \text{foreign lags } i \neq j \end{cases}$$

Hyperparameter inference

- **Bayesian inference** $p(\boldsymbol{\beta}, \boldsymbol{\theta}|\mathrm{Y}_{1:T})$ is simple in principle:
  - **Direct sampling** (rarely an option)
  - **Hamiltonian MC** on joint $\boldsymbol{\beta}, \boldsymbol{\theta}|\mathrm{Y}_{1:T}$
  - **Gibbs sampling** $\boldsymbol{\beta}|\boldsymbol{\theta}, \mathrm{Y}_{1:T}$ and $\boldsymbol{\theta}|\boldsymbol{\beta}, \mathrm{Y}_{1:T}$
- Practitioners prefer to fix $\boldsymbol{\theta}$ "once and for all". Move on to parameter inference, model checking, forecasting, policy etc
- Bayesian VARs: "we use the values from Doan et al (1984)".
- Empirical Bayes: **maximize marginal likelihood**

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ln p(\mathrm{Y}_{1:T}|\boldsymbol{\theta})$$

or marginal posterior $p(\boldsymbol{\theta}|\mathrm{Y}_{1:T})$.

Bayesian optimization

- Marginal likelihood often **intractable**:
  - analytical approximation (Laplace, INLA)
  - use HMC/MCMC simulation to compute $\ln p(Y_{1:T}|\boldsymbol{\theta})$.
- Typical hyperparameter estimation setup:
  - **costly** function evaluations
  - **noisy** function evaluations
  - function argument is **low-dimensional**.
- **Bayesian optimization** well suited for all three issues.

Gaussian processes

- **Gaussian process regression**

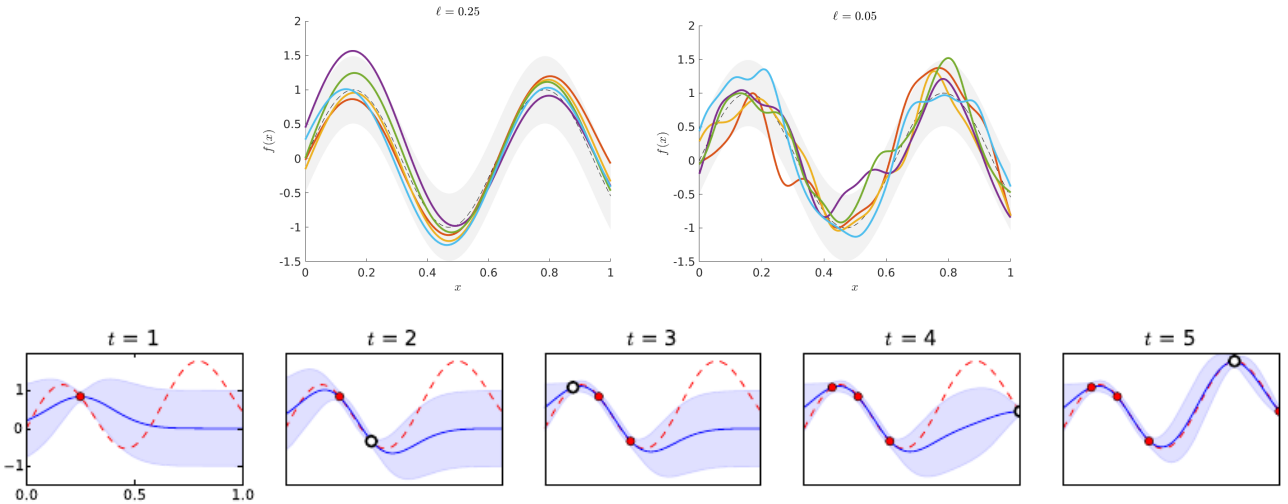
$$y_i = f(\mathrm{x}_i) + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma_n^2)$$

- **Gaussian process prior** over the space of functions

$$f(\mathrm{x}) \sim \mathcal{GP}\left(m(\mathrm{x}), k(\mathrm{x}, \mathrm{x}')\right)$$

with **squared exponential** covariance function

$$k(\mathrm{x}, \mathrm{x}') \equiv \mathrm{Cov}(f(\mathrm{x}), f(\mathrm{x}')) = \sigma_f^2 \exp\left(-\frac{\|\mathrm{x} - \mathrm{x}'\|^2}{2\ell^2}\right)$$

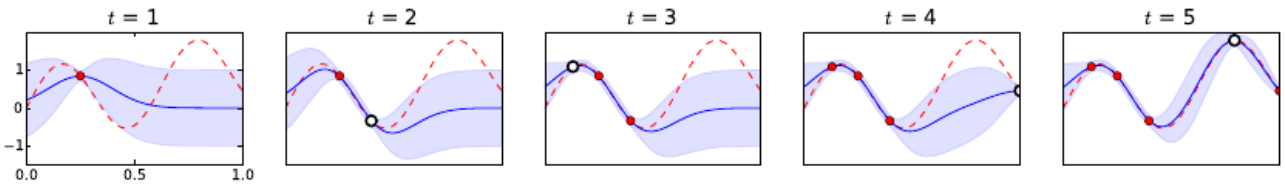


Bayesian optimization

- Aim: **maximization of expensive function**

$$\text{argmax}_{x \in \mathcal{X}} f(x)$$

- **Bayesian optimization**:
  - Assume $f \sim \mathcal{GP}$
  - Evaluate f at $x_1, x_2, ..., x_n$.
  - Update to posterior distribution $f|x_1, ..., x_n \sim \mathcal{GP}$.
  - Use posterior of f to find a new x_{n+1}. **Explore** vs **Exploit**.
  - Iterate until convergence.
- Optimal x_{n+1} through an **acquisition function**.
- Example: **upper confidence bound:**



Acquisition functions

- **Probability of Improvement (PI)**

$$a_{\mathrm{PI}}(\mathrm{x}) \equiv \Pr\left(f(\mathrm{x}) > f_{\mathrm{best}}\right) = 1 - \Phi\left(\frac{f_{\mathrm{best}} - \hat{m}\left(\mathrm{x}; \mathcal{D}_n\right)}{s\left(\mathrm{x}; \mathcal{D}_n\right)}\right)$$

  - $\mathcal{D}_n = \{f(\mathrm{x}_1), \ldots, f(\mathrm{x}_n)\}$ are past evaluations
  - f_{best} is the smallest function value so far
  - $\hat{m}\left(\mathrm{x}; \mathcal{D}_n\right)$ is posterior mean of $f(\mathrm{x})$
  - $s\left(\mathrm{x}; \mathcal{D}_n\right)$ is posterior standard deviation of $f(\mathrm{x})$.

- **Expected Improvement (EI)** takes also into account the **size of the improvement**.

- **Expected Improvement per Second** - takes a **known function cost** into account.

- Non-convex acquisition function optimization, but **deterministic** and **cheaper** than original problem.

Marginal likelihood estimated from sampling

- **Marginal likelihood** $f(\theta) \equiv \ln p(Y_{1:T}|\theta)$ often estimated by **sampling:**
  - Chib (Gibbs) and Chib-Jeliazkov (MH)
  - Importance sampling,
  - etc etc
- **Noisy** evaluations $\hat{f}(\theta)$.
- **Precision** of $\hat{f}(\theta)$ controlled via **number of samples** G.
- **MCMC efficiency** and therefore $\mathbb{V}(\hat{f}(\theta))$ **varies over** θ**-space**, particularly when θ contains prior/regularization hyperparameters.
- **Stopping early** when probability of improvement (PI) is low.

Bayesian Optimization with Optimized Precision

- **BOOP**:
  - **Early stopping of evaluation** when **PI**$< \alpha$.
  - G **random** - we don't know G until we visit $\boldsymbol{\theta}$.
  - **EI per second**, but with G **predicted** for every $\boldsymbol{\theta}$.
- Early stopping affects the **planning of future computations**.
- BOOP can try $\boldsymbol{\theta}$ with low EI, if expected to be cheap enough.
- **Heteroscedastic GP regression model** for the estimates

$$\hat{f}(\boldsymbol{\theta}_i) = f(\boldsymbol{\theta}_i) + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2(G_i)\right)$$

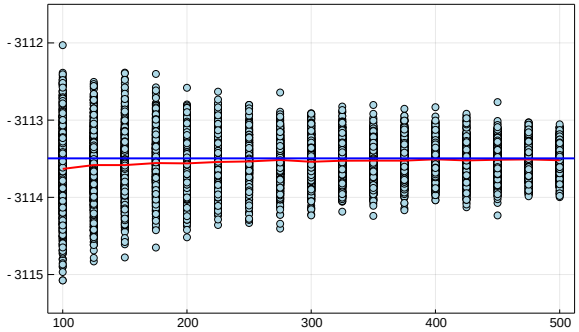
- GP for **predicting the number of samples** G:

$$\ln G_i = h(\mathbf{z}_i) + \varepsilon_i \qquad \varepsilon_i \overset{\text{iid}}{\sim} N\left(0, \psi^2\right),$$

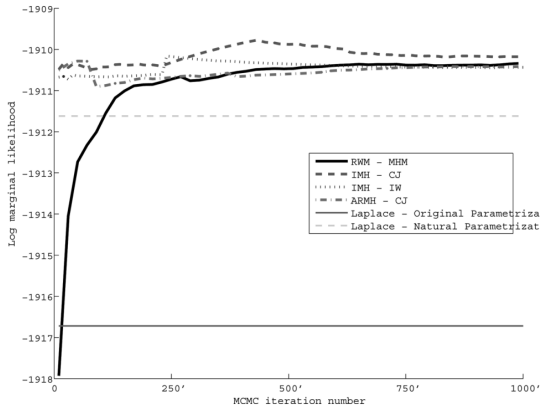
where z are variables with predictive power for G, e.g. the hyperparameter values themselves, or $\hat{m}(\boldsymbol{\theta}) - f_{\max}$.

Bayesian Optimization with Optimized Precision

- BOOP **assumes (approx) unbiasedness** $\mathbb{E}\hat{f}(\boldsymbol{\theta}_i) = f(\boldsymbol{\theta}_i)$.
- **Sampling distribution of Chib's estimator** for SSBVAR.



- Unbiasedness depends on the Sampler-Estimator combination.
- Log marginal likelihood estimates in large-scale DSGE model:



The BOOP algorithm

for j until convergence do:

a) **Fit the heteroscedastic GP** for f based on past evaluations

$$\hat{f}(\mathrm{x}_{1:(j-1)}) = f(\mathrm{x}_{1:(j-1)}) + \epsilon, \quad \epsilon \sim N(0, \Sigma_{1:(j-1)})$$
$$f(\mathrm{x}) \sim \mathcal{GP}(m(\mathrm{x}), k(\mathrm{x}, \mathrm{x}')),$$

where $\Sigma_{1:(j-1)} \equiv \mathrm{Diag}(\sigma^2(G_1), \ldots, \sigma^2(G_{j-1}))$.

b) **Fit the GP for** $\log G$ based on past evaluations

$$\log G_{1:(j-1)} = h(\mathrm{z}_{1:(j-1)}) + \varepsilon, \quad \varepsilon \sim N(0, \psi^2 \mathrm{I})$$
$$h(\mathrm{z}) \sim \mathcal{GP}(m_G(\mathrm{z}), k_G(\mathrm{z}, \mathrm{z}')),$$

Return (median) point prediction $\hat{G}_\alpha(\mathrm{x})$.

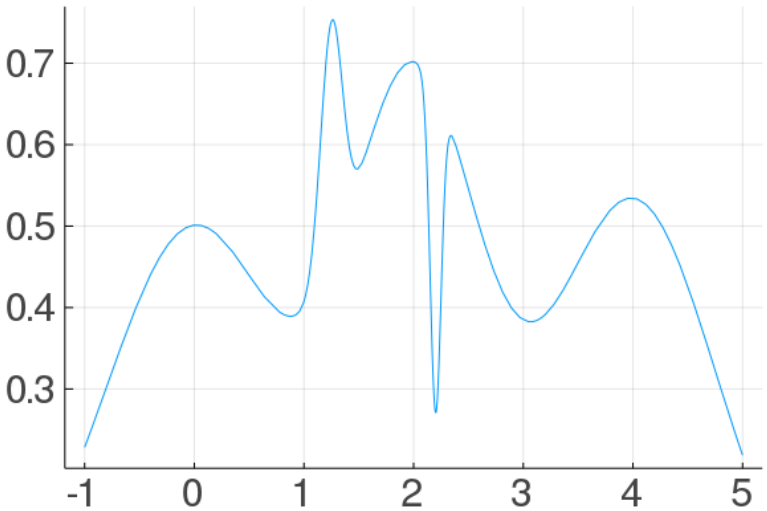
c) **Select the next point** x_j by maximizing $\tilde{a}_\alpha(\mathrm{x}) = a(\mathrm{x}) / \hat{G}_\alpha(\mathrm{x})$.

d) Compute $\hat{f}(\mathrm{x}_j)$ and $\sigma^2(G_j)$ by early stopping at thresholding probability α.

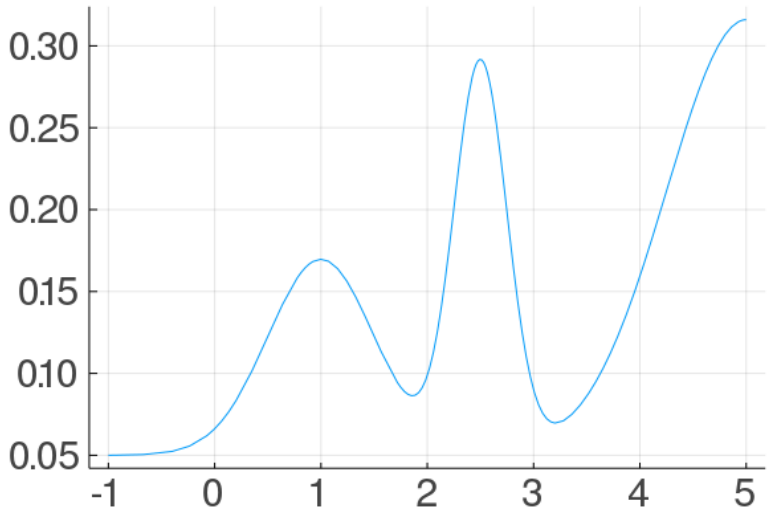
e) Update the datasets in a) with $(\mathrm{x}_j, \hat{f}(\mathrm{x}_j), \sigma^2(G_i))$ and in b) with $(\mathrm{z}_j, \log G_j)$.

BOOP in action - simulated example

True $f(x)$



True $S(\hat{f}(x))$



BOOP in action - simulated example

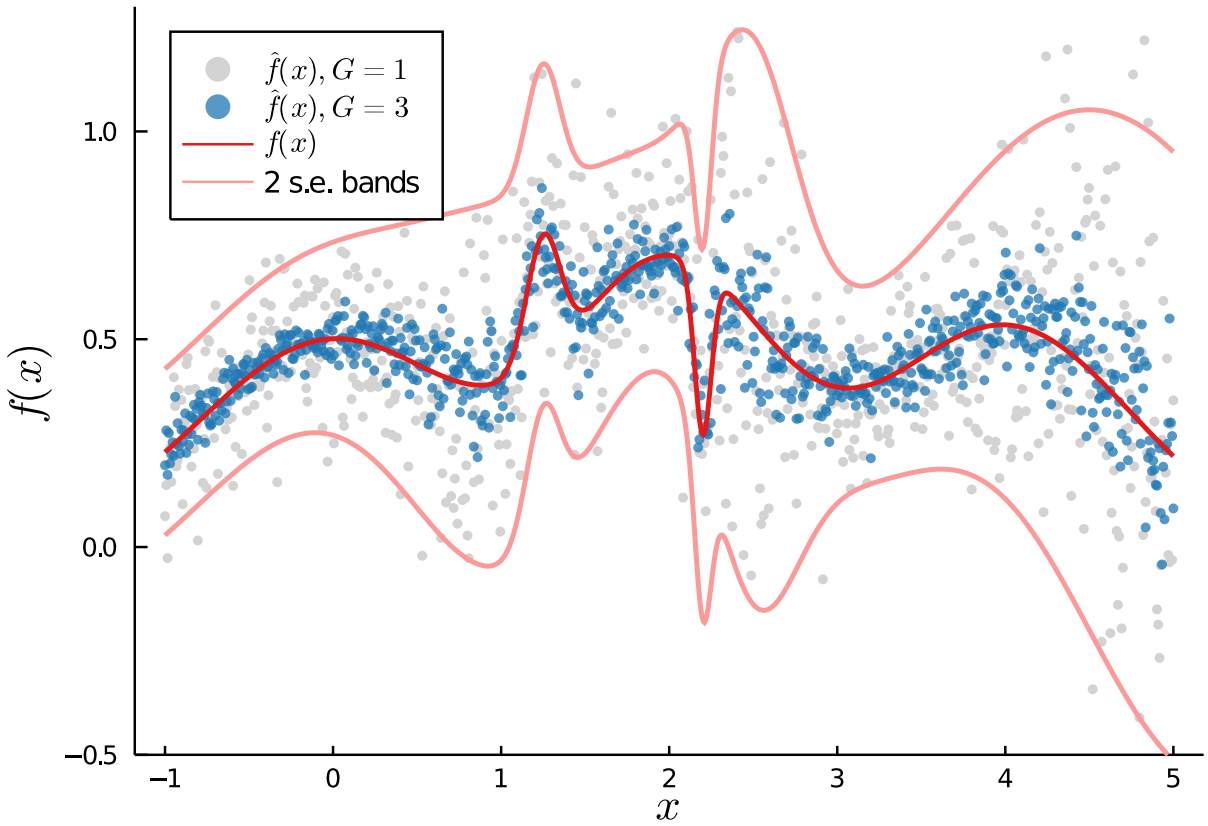
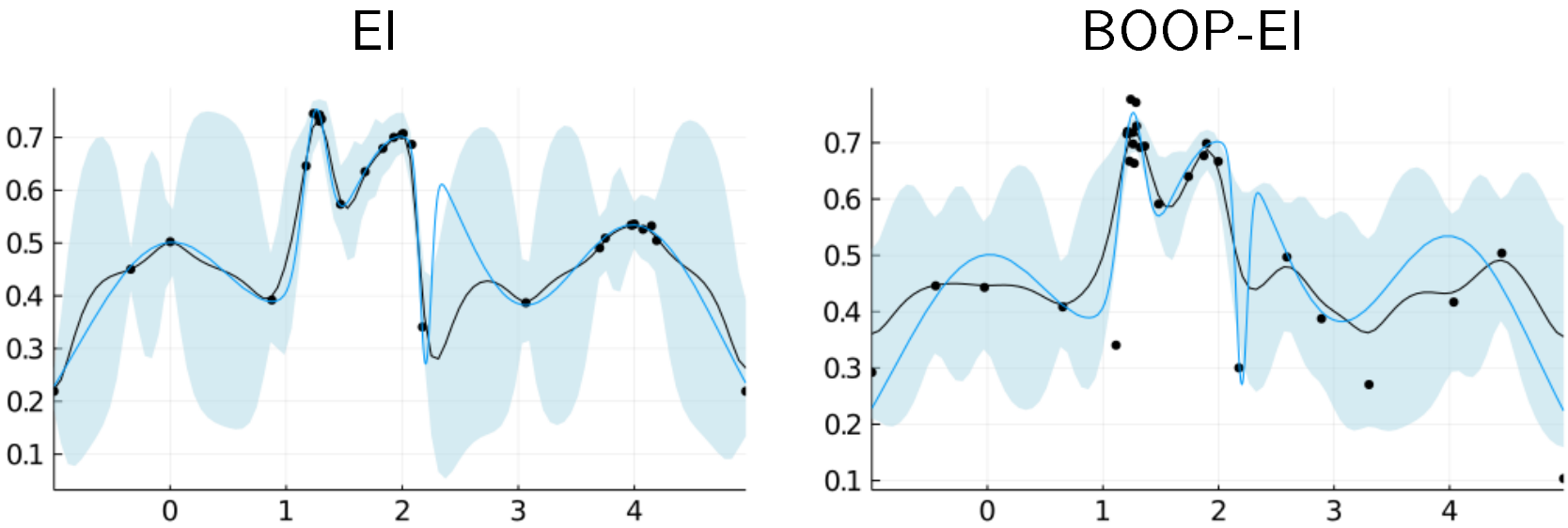
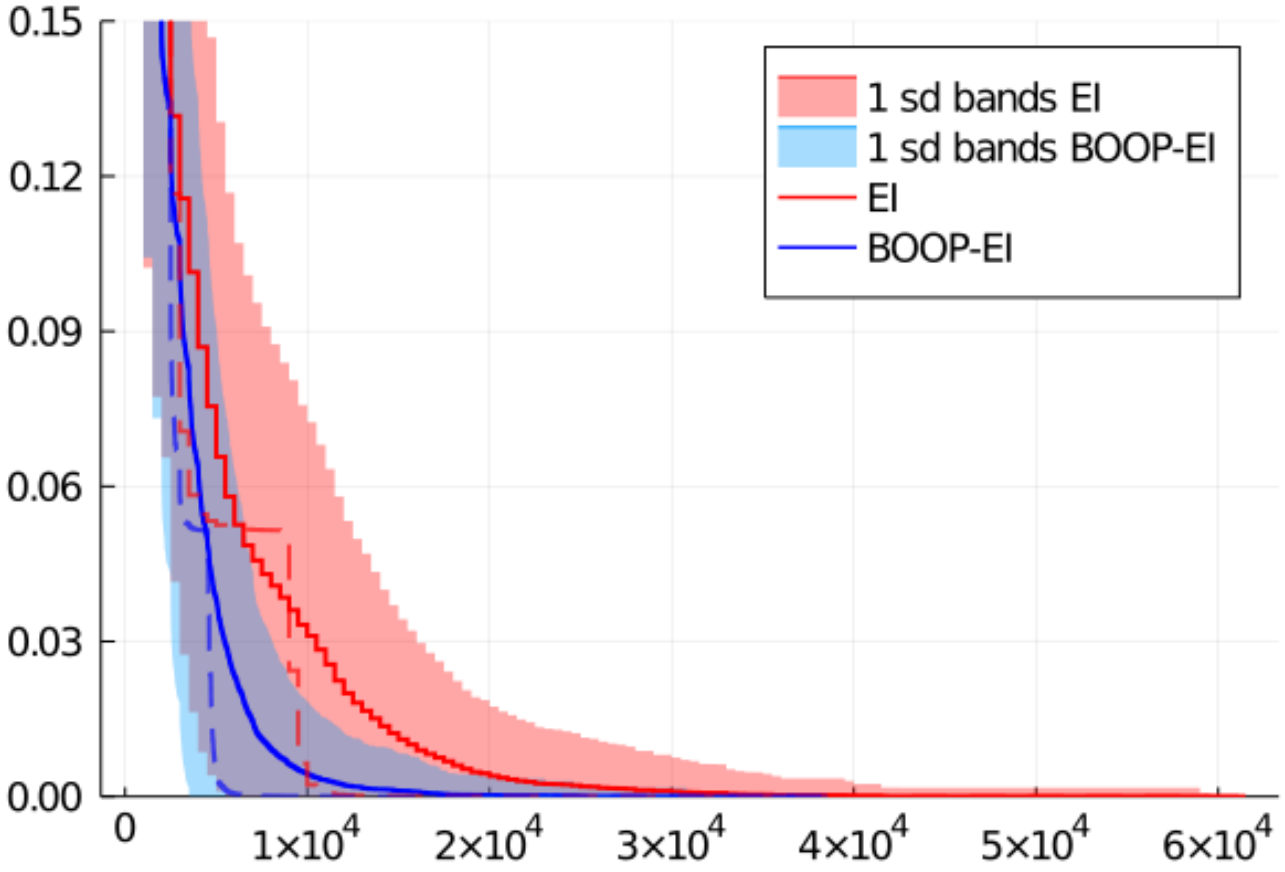


Figure: Estimates for $G = 1$ samples (grey dots), $G = 3$ samples (blue dots), the mean function (red line) and 2 standard deviation error bands (pink lines).

BOOP in action - one example run

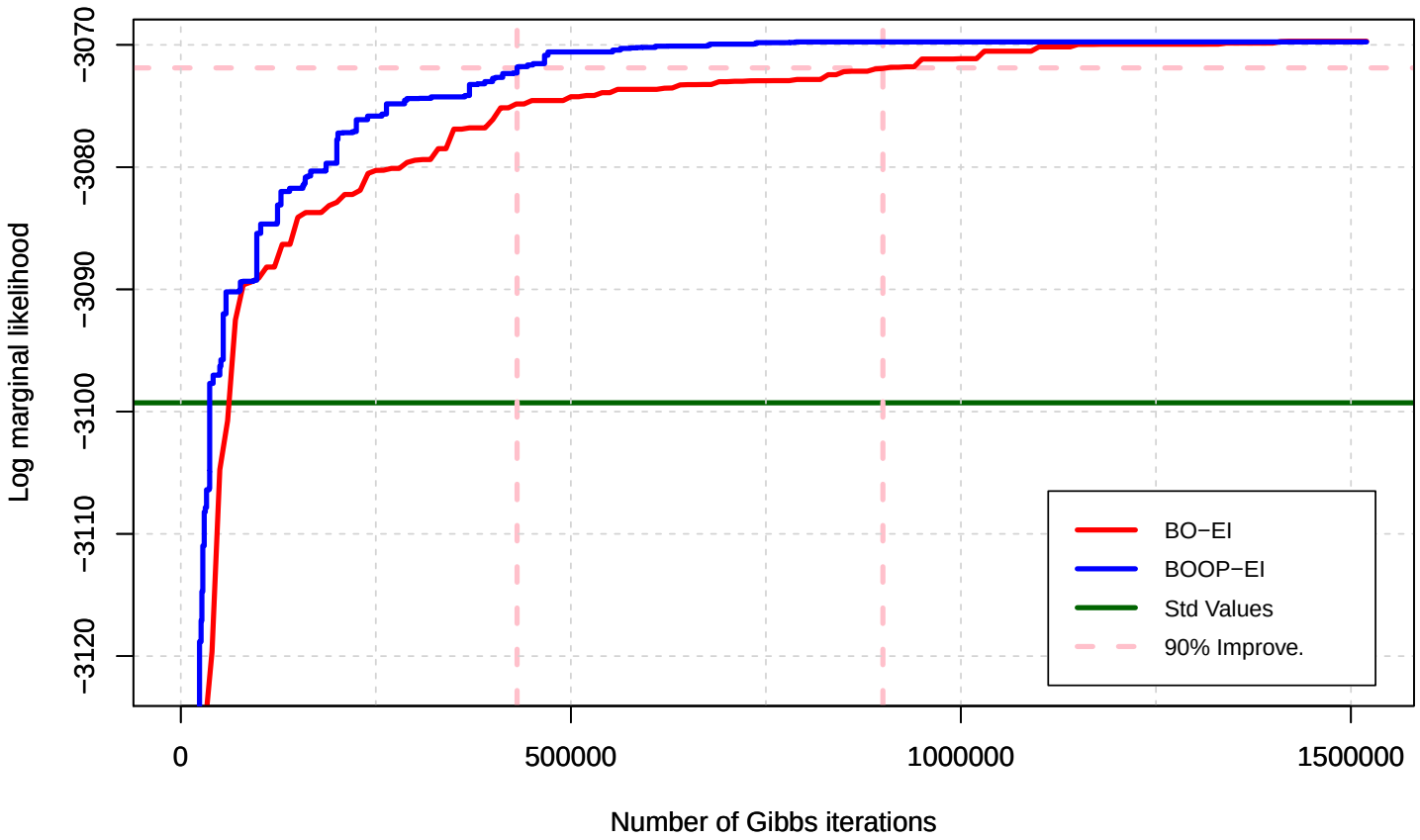


BOOP in action - repeated runs



7-variable BVAR - BOOP finds estimates quicker

- 7 variable **steady-state BVAR** on US data.
- Gibbs sampling with **Chib's marginal likelihood estimator**.
- BO to find optimal prior hyperparameters $\boldsymbol{\theta} = (\lambda_1, \lambda_2, \lambda_3)$.



22-variable steady-state BVAR

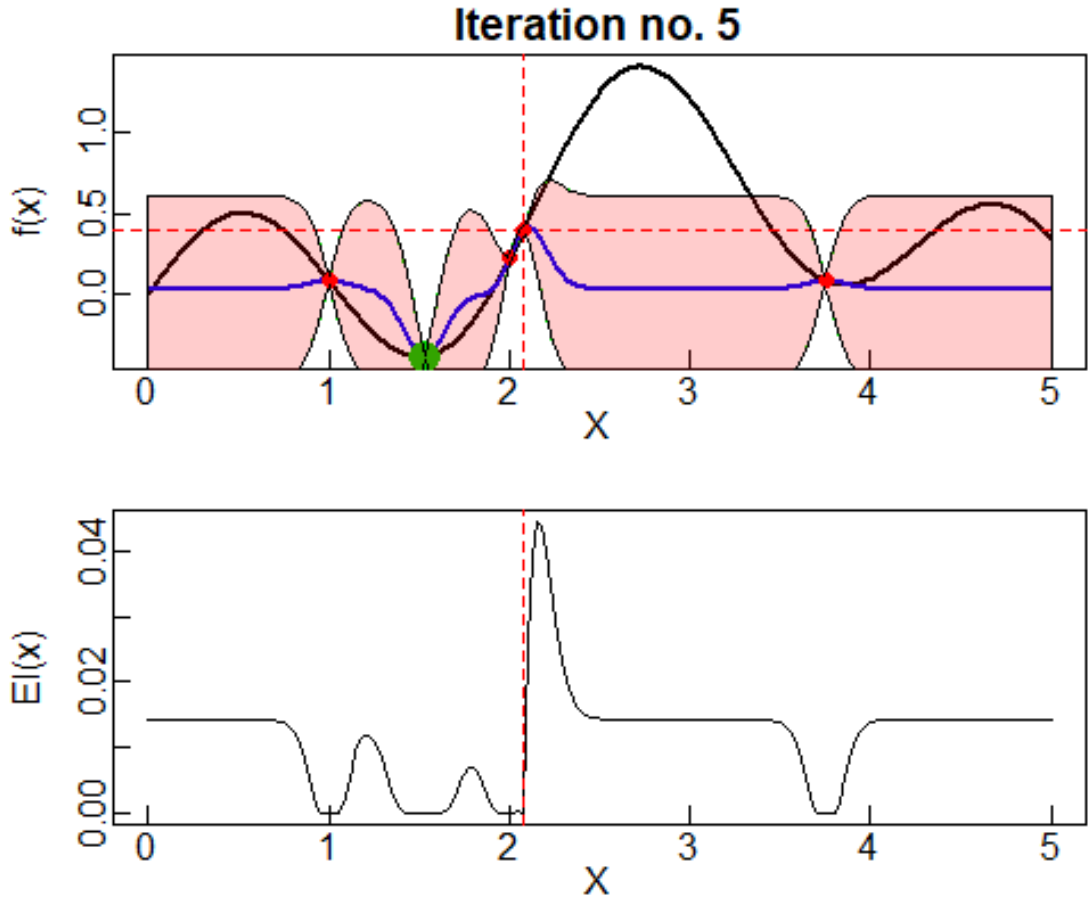
	Standard	BO-EI	BOOP-EI	Grid (Medium)
Log ML	-7576.31	-7458.61	-7458.19	-7566.86
Sd log ML	0.54	0.28	0.42	0.53
λ_1	0.1	0.52	0.58	0.3
λ_2	0.5	0.11	0.08	0.4
λ_3	1	1.79	1.72	0.9

Table: Optimization Results Large Steady-State BVAR.

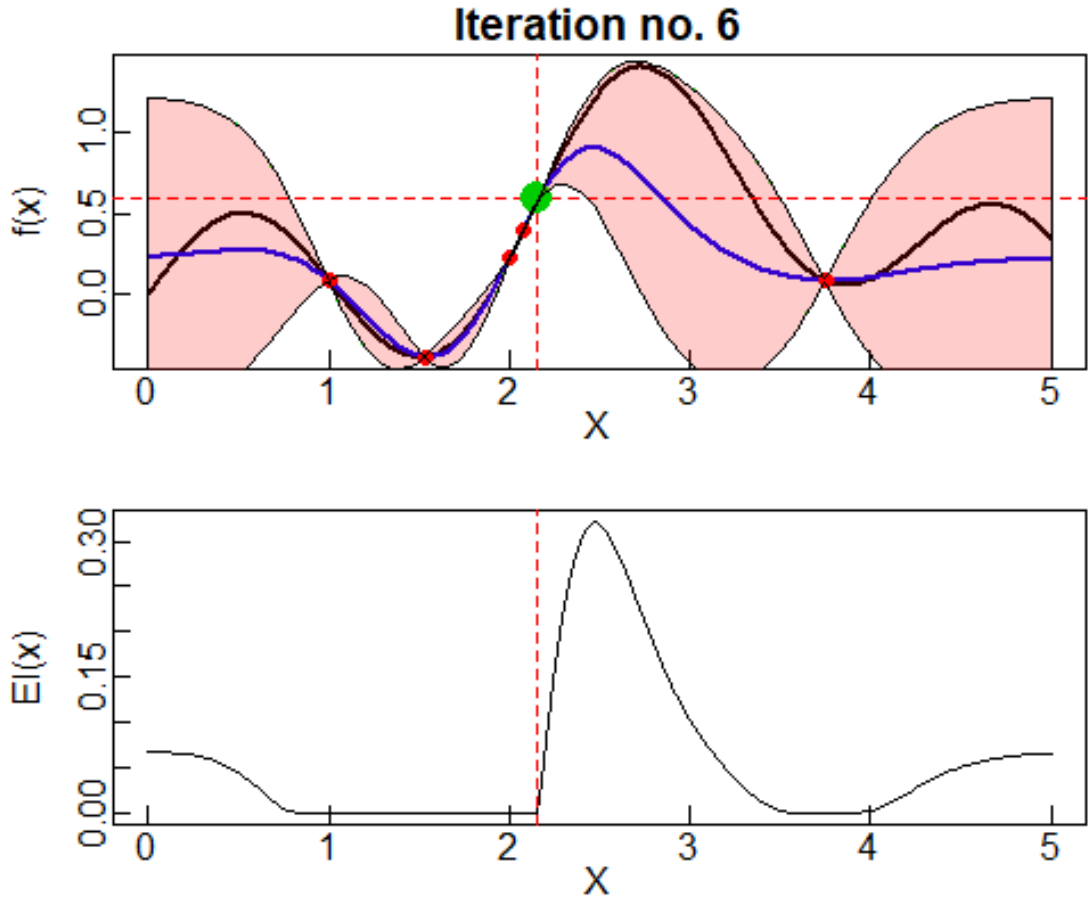
Conclusions

- **Bayesian optimization** is an attractive method for **costly**, **noisy**, **low-dimensional functions**.
- Hyperparameter optimization using **marginal likelihood estimated from MC sampling**.
- We extend BO to exploit that **the user controls the precision of the evaluations** via the number of samples.
- Experiments on simulated data and steady-state BVAR on US macro data shows that **BOOP find the optimal hyperparameters faster**.
- **Future work**: other models, marginal likelihood estimators. Handling biased estimators.

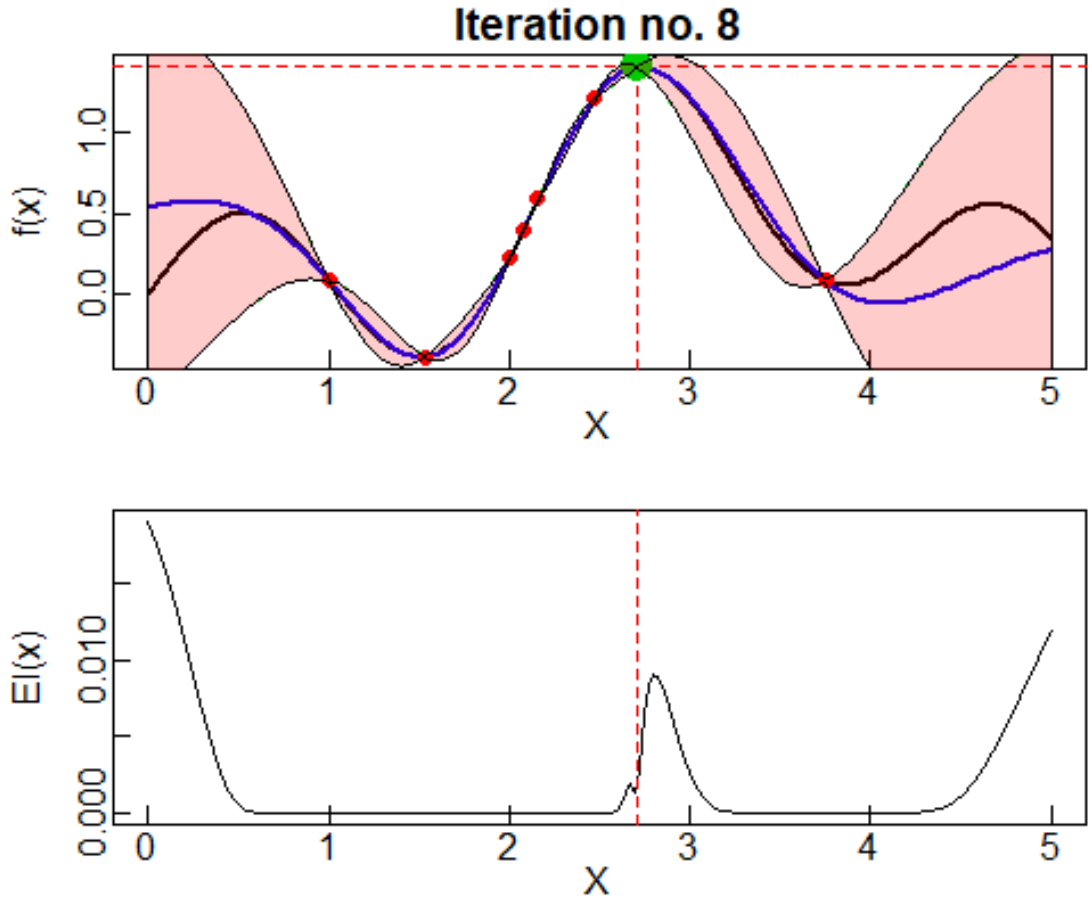
Bayesian optimization in action



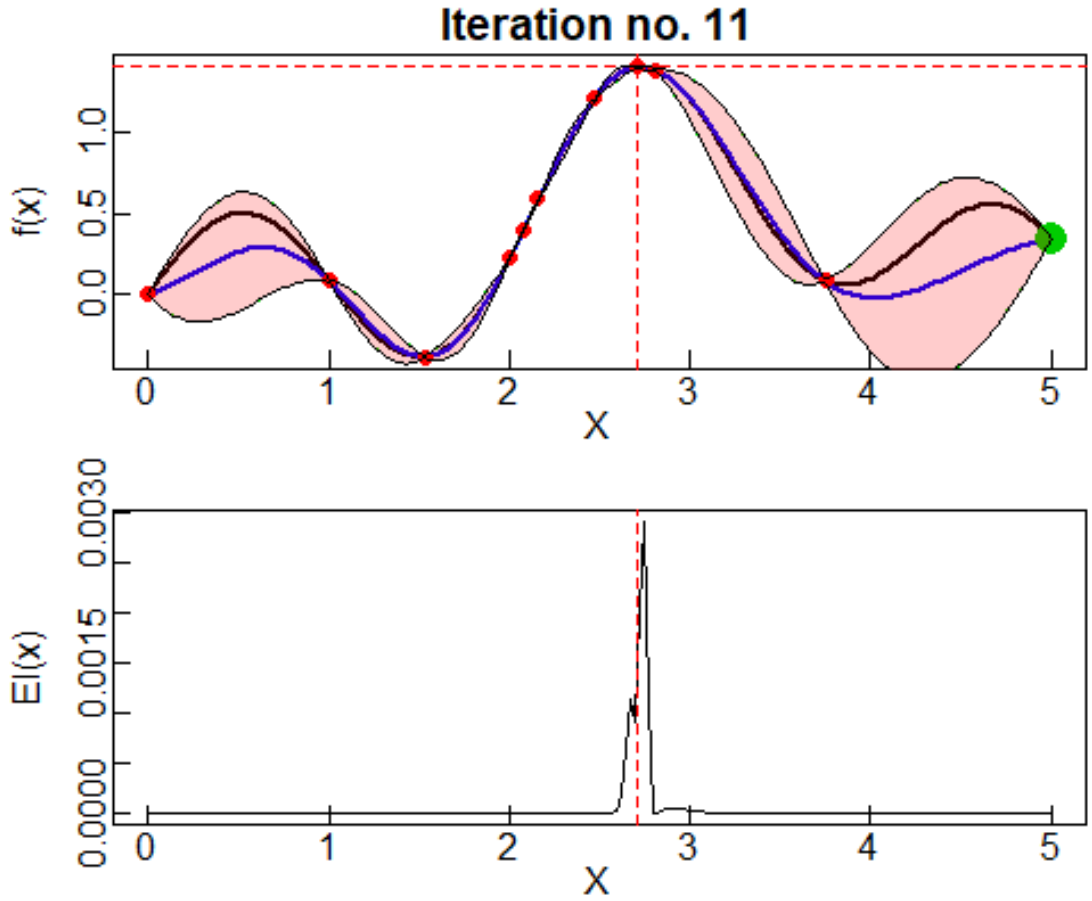
Bayesian optimization in action



Bayesian optimization in action



Bayesian optimization in action



7-variable BVAR - BOOP finds estimates quicker

	Standard	BO-EI	BOOP-EI	Grid
log ML	-3099.28	-3069.01	-3068.84	-3069.03
Gibbs iterations		$1.5 \cdot 10^6$	788138	10^9
Avg. iter to 90%		$9 \cdot 10^5$	431061	
Model evaluations		150	150	10^5
λ_1	0.1	0.30	0.27	0.3
λ_2	0.5	0.38	0.43	0.4
λ_3	1	0.69	0.76	0.9

Table: Optimization Results Medium Steady-State BVAR.

How well do GPs estimate the $\ell(\theta)$ surface?