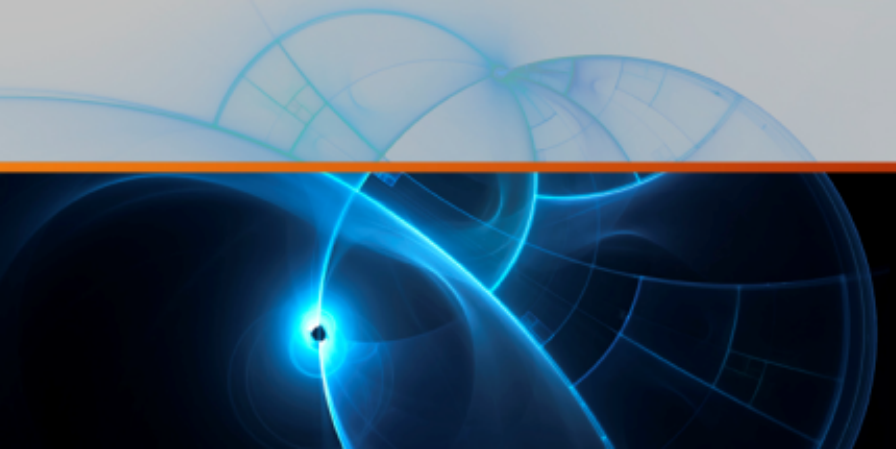
Towards Large-Scale Surrogate-Based Optimization

F.V. Gubarev, A.I. Pospelov



DATADVANCE
AN AIRBUS GROUP COMPANY

SBO methods fit nicely into **Eng**ineering **Opt**imization framework:

- Evaluation budget is easily controlled and is minimal [in the majority of cases]
- Robust wrt undefined designs and noisy responses
- Search is globalized with easily regulated globalization degree

SBO usage is truly justified when:

- Underlying model evaluation is time-expensive compared to the time-cost of internal optimizer activities:

$$T_{external} \gg T_{internal}$$

We assume that without externally imposed budget the number of sampled designs remains relatively small

$$N_{sample} \ll 2^D$$

(e.g., underlying model multi-modality is only moderate)

Generic SBO scheme:

1. Generate DoE-based sample
2. Construct the surrogate model(s)
3. Globally optimize model-based criterion to get evaluation candidates
4. Evaluate underlying model at predicted designs
5. Augment current sample, goto 2

For concreteness we'll discuss **G**aussian **P**rocesses based models with stationary correlations.

What are the bottlenecks of the above scheme?

Major bottlenecks:

1. **GP Model Construction (Training)**

   Conventional training becomes technically impossible at $N_{sample} \sim O(10^3)$. Using an estimate $N_{sample} \sim D^2$ (quadratic RSM) we obtain:

   $$D_{max} \sim O(10)$$

2. **Optimization of model-derived criterion**

   It is virtually impossible to predict GP model changes upon sample augmentation. Hence, computationally expensive **global** optimization of model-derived criterion is to be performed anew.

   In fact, the above issues might easily lead to

   $$T_{external} \ll T_{internal}$$

   in many practically relevant applications

Ultimately, one wants to:

1. Significantly reduce the cost of GP model construction
2. Boost the maximal available design space dimensionality
3. Get the control over model changes upon sample augmentation

Let's estimate:

- GP-type predictions require to invert correlation matrix at least once. Hence for direct matrix algorithms:

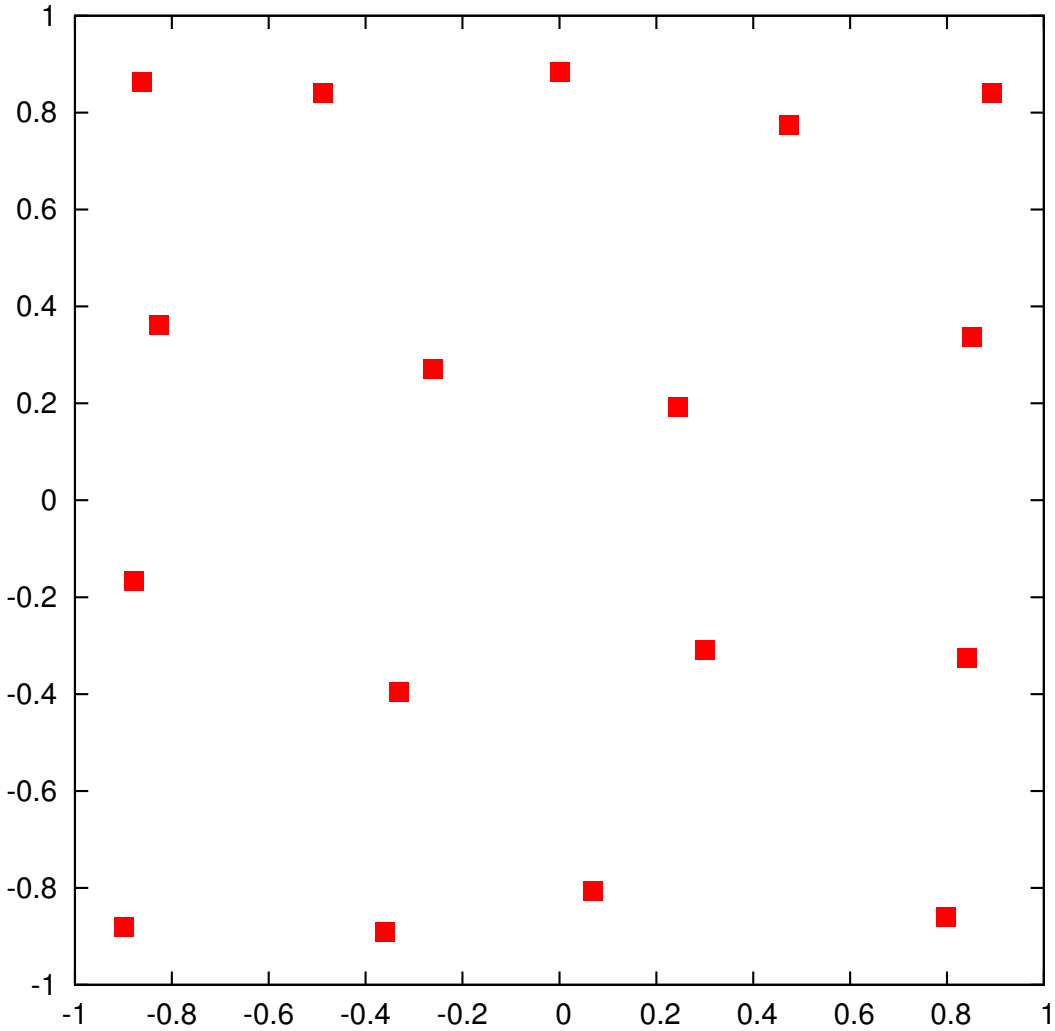
$$N_{sample}^{max} \sim O(10^4) \quad \rightarrow \quad D_{max} \sim O(10^2)$$

The above could not come for free, admissible penalty is:

- Enlargement of required evaluation budget provided that we still have

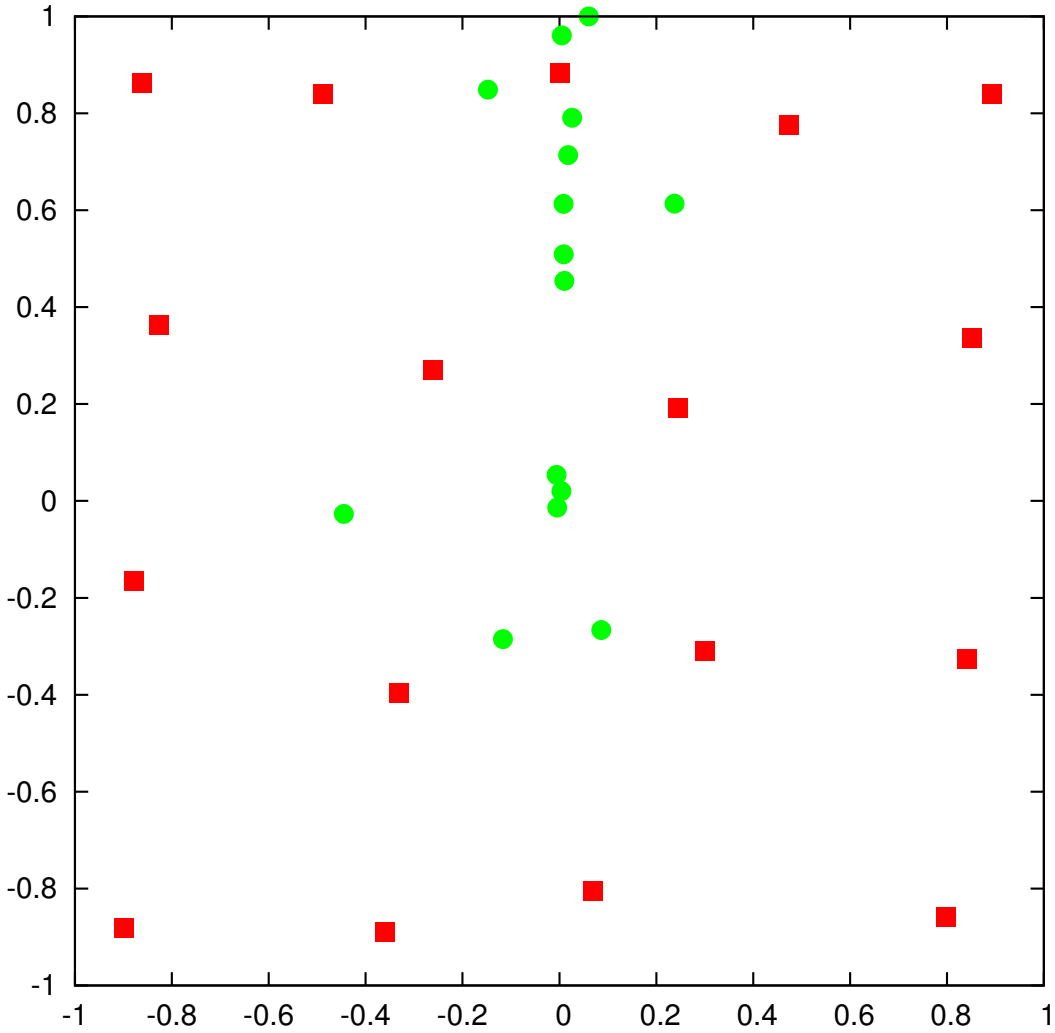
$$T_{external} \lesssim T_{internal}$$

Qualitative picture of SBO-inspired optimization process:



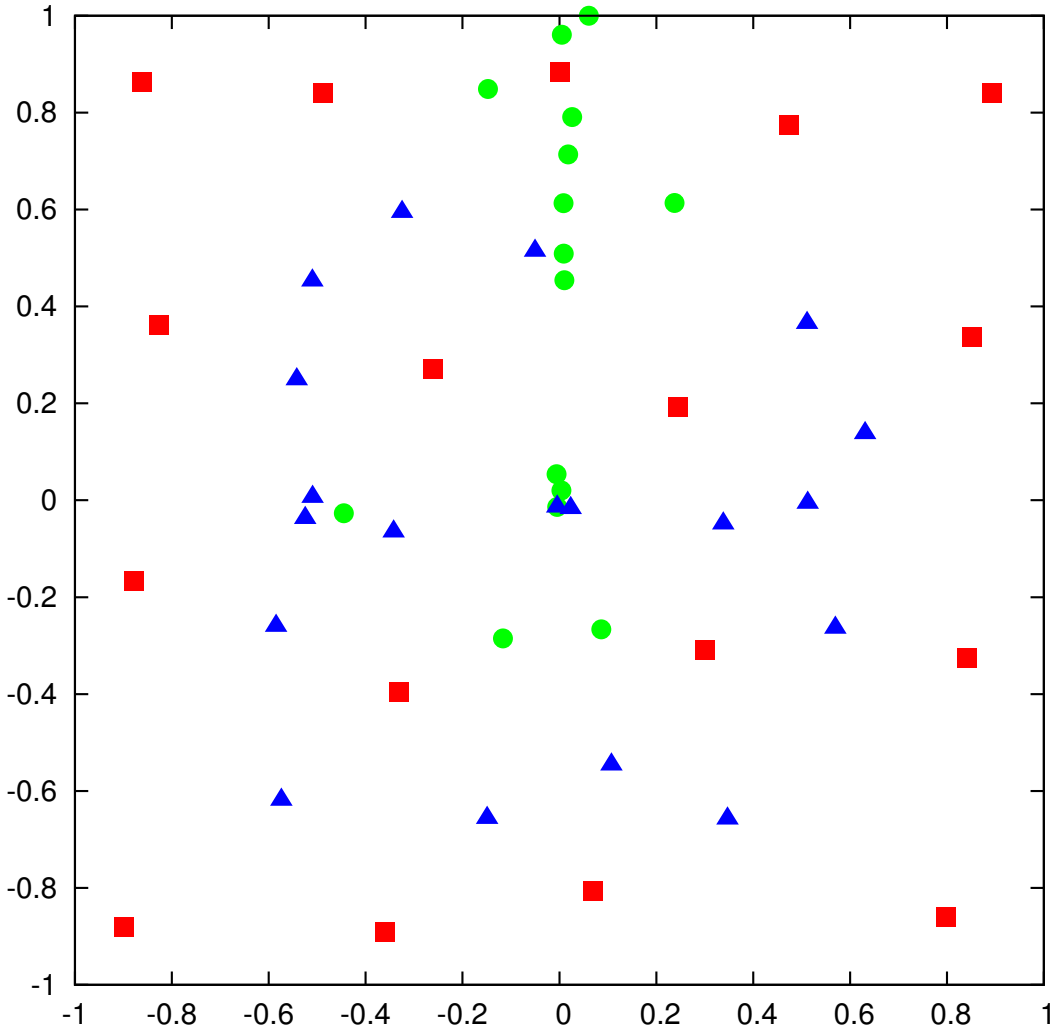
DoE sample

Qualitative picture of SBO-inspired optimization process:



After a few iterations

Qualitative picture of SBO-inspired optimization process:



Almost "converged" process

Prime observations:

1. Evaluated designs **cluster** in promising regions
2. **Hierarchy of length scales** could be observed:

   Let $\langle L \rangle_x$ denotes characteristic distance between nearest sampled designs around x. Then

   - DoE stage: $\langle L \rangle_x = L_0 \quad \forall x$
   - After a few iterations (Ω is some promising region):

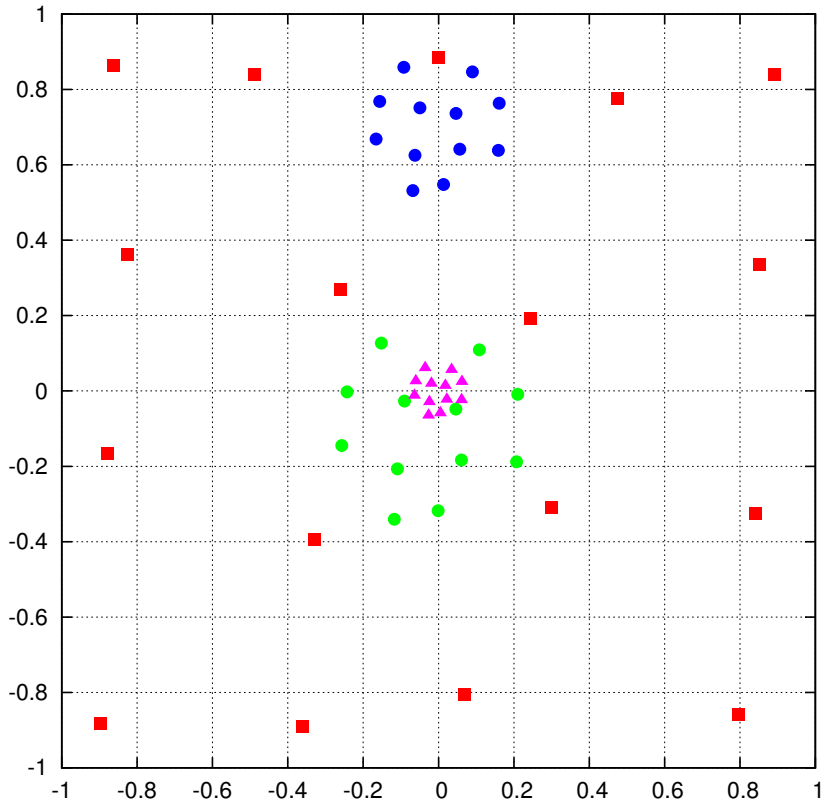
   $$\langle L \rangle_x = L_0 \quad x \notin \Omega \qquad \langle L \rangle_x = L_1 \quad x \in \Omega \qquad L_1 \lesssim L_0$$

   - At later stages (Ω_i are the nested promising regions):

   $$\begin{aligned} \langle L \rangle_x &= L_0 \quad x \notin \Omega \\ \langle L \rangle_x &= L_1 \quad x \in \Omega_1 \\ &\cdots \\ \langle L \rangle_x &= L_k \quad x \in \Omega_k \\ L_k \lesssim \cdots &\lesssim L_1 \lesssim L_0 \end{aligned}$$

At the expense of additional evaluations we could **enforce length scales hierarchy** at every iteration:

- Instead of single candidate evaluation we perform DoE sampling in candidate's vicinity, determined by upper region length scale.



Consequences:

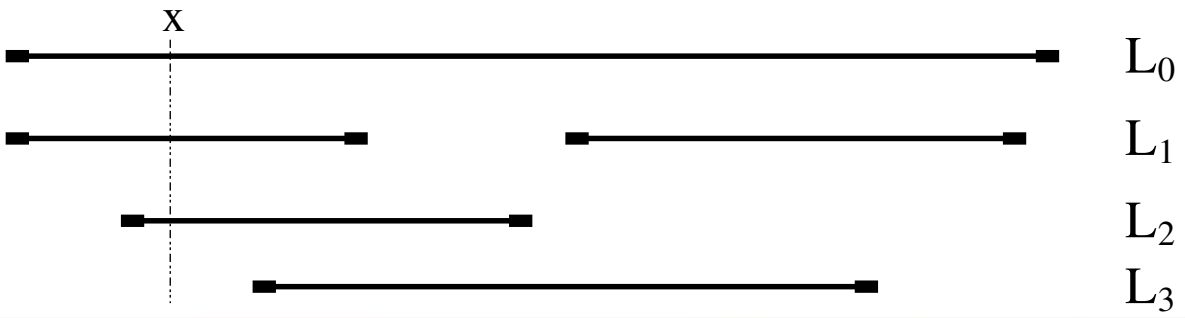
1. Underlying model is **not only probed** at candidate location x_c, but **is explored** in candidate's vicinity $\Omega(x_c)$

$$F(x_c) \rightarrow \{F(x_i)\}, i \in \Omega(x_c)$$

2. Every iteration induces well-defined smaller length scale L_k

$$L_k \lesssim \cdots \lesssim L_1 \lesssim L_0 \,,$$

each L_k being associated with particular nested regions.



Anzats for **multi-resolution** GP correlation function, which reflects the above hierarchy of length scales:

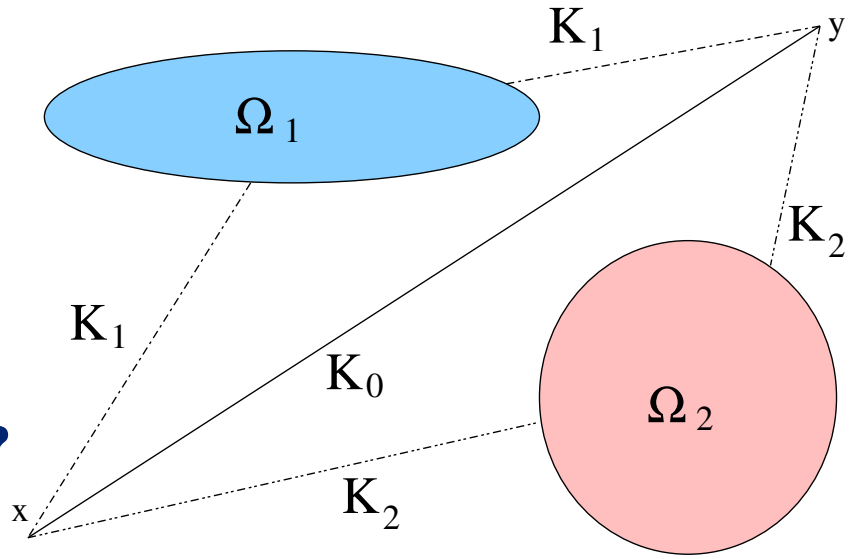
$$K(x,y) = K^{(0)}(x,y) + \sum_{\mu} \alpha_{\mu} \sum_{i,j} K^{(\mu)}(x, x_{\mu}^{i}) \, [K^{(\mu)}]^{-1}_{ij} \, K^{(\mu)}(x_{\mu}^{j}, y) \,,$$

where $K^{(\mu)}$ are Ω_{μ}-specific correlation vector/matrix.

Parameters to be determined:

- $K_{(\mu)}$-specific parameters
- Amplitudes $\alpha_{\mu} \geq 0$

Is that what we really wanted?



Clue is provided by basic observations:

- Prime parameter of every GP correlation function is its correlation length L, e.g.

$$K^{(\mu)}(x,y) \sim e^{-|x-y|/L}, \qquad K^{(\mu)}(x,y) \sim e^{-|x-y|^2/L^2}, \qquad \ldots$$

- For dimensional reasons L is ought be of order L_μ

$$L \sim L_\mu$$

Now we qualitatively argue that within considered context:

- Missing factor of order unity in $L \sim L_\mu$
- Other details of "exact" correlator $K^{(\mu)}$

are **not much important** and represent next order effects.

Regularization parameter ("nugget" term):

- Number of sampled designs is relatively small, hence fair estimation of data noise is not possible
- Underlying model evaluation is sufficiently expensive, we want to account for all measured data

Thus, GP models are ought to be almost interpolating. Nugget term **is to be taken as small** as permitted by numerical stability.

Length scale asymmetry in design space:

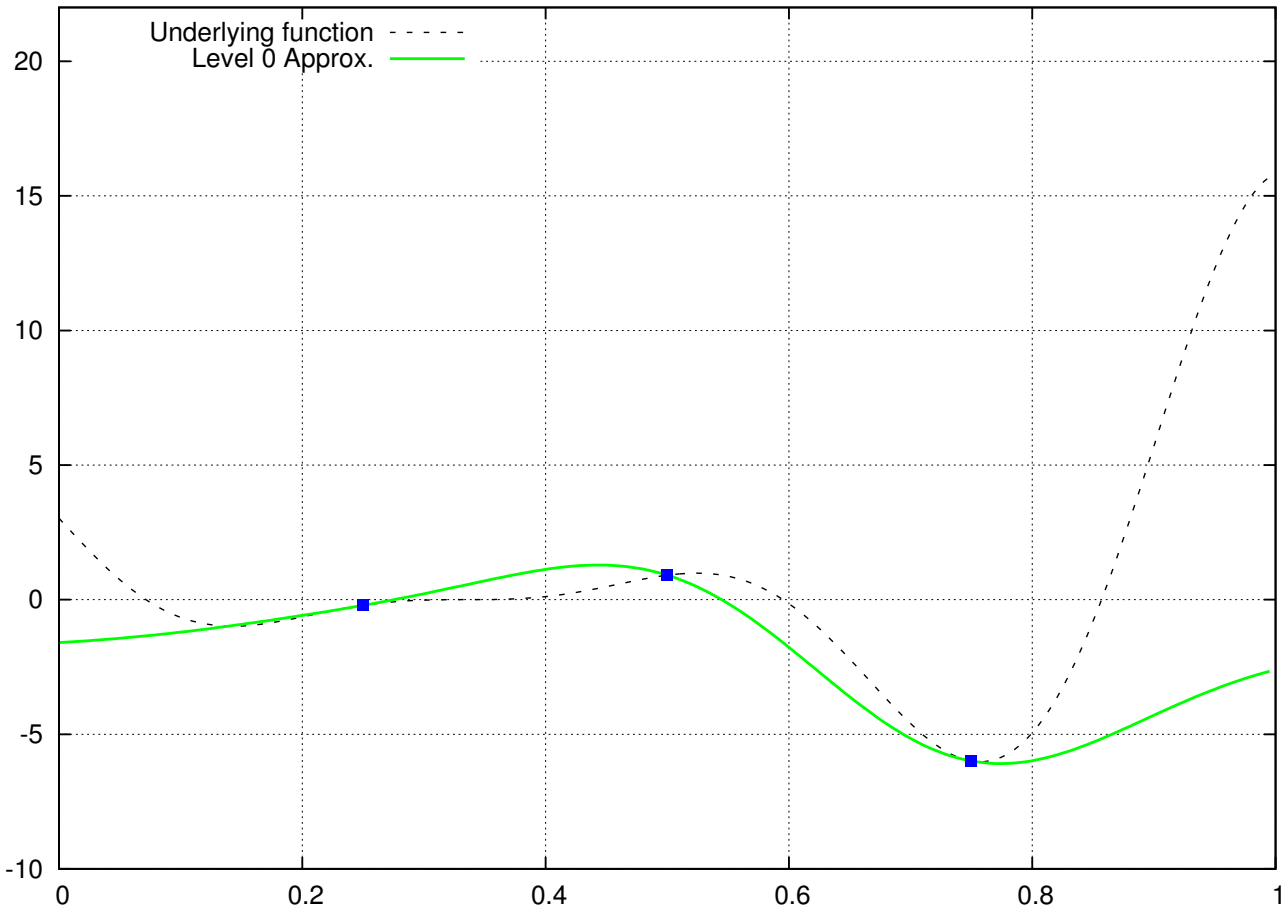
- Samples are taken regular in each nested domain (DoE-like), hence no large asymmetry factor could arise
- GP-based models are quite robust wrt sufficiently large variations of length scale around correct value (not in extrapolation regime, sure)

Thus, length scale asymmetry in design space might be **neglected**.

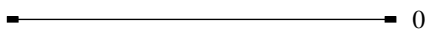
Summary:

- Prime gross features of $K^{(\mu)}$ are known in advance once length scale hierarchy is respected
- Seems that could avoid $K^{(\mu)}$-parameters tuning ("training") altogether
- Only amplitude α_μ is to be determined for every new region (every iteration)
- α_μ determination is cheap (no inversions of large matrices is involved)
- Knowledge of length scale hierarchy allows to predict the domains where model is changing upon sample augmentation.

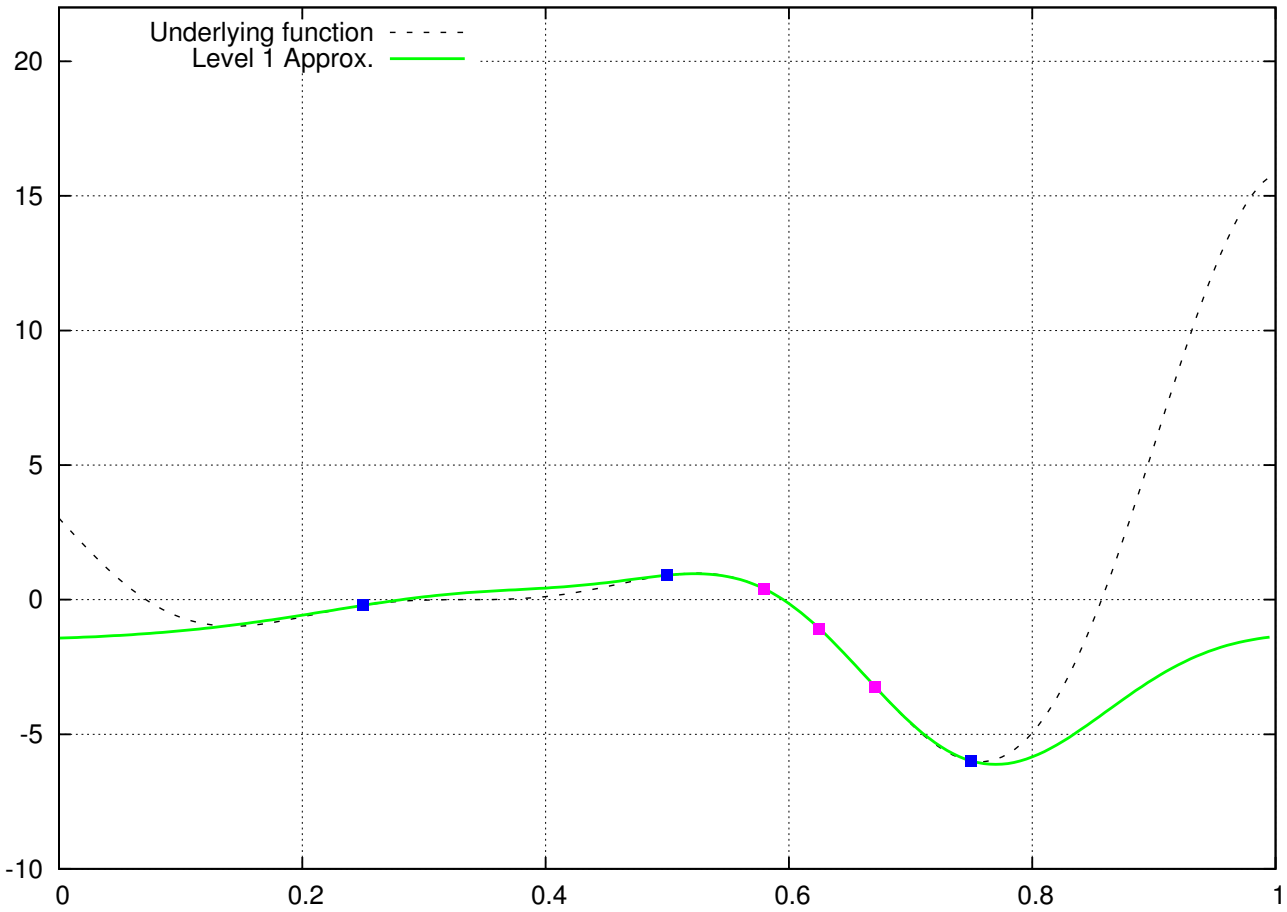
Hierarchical modeling of $(6x-2)^2 \sin[12x-4]$



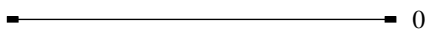
Model Scheme



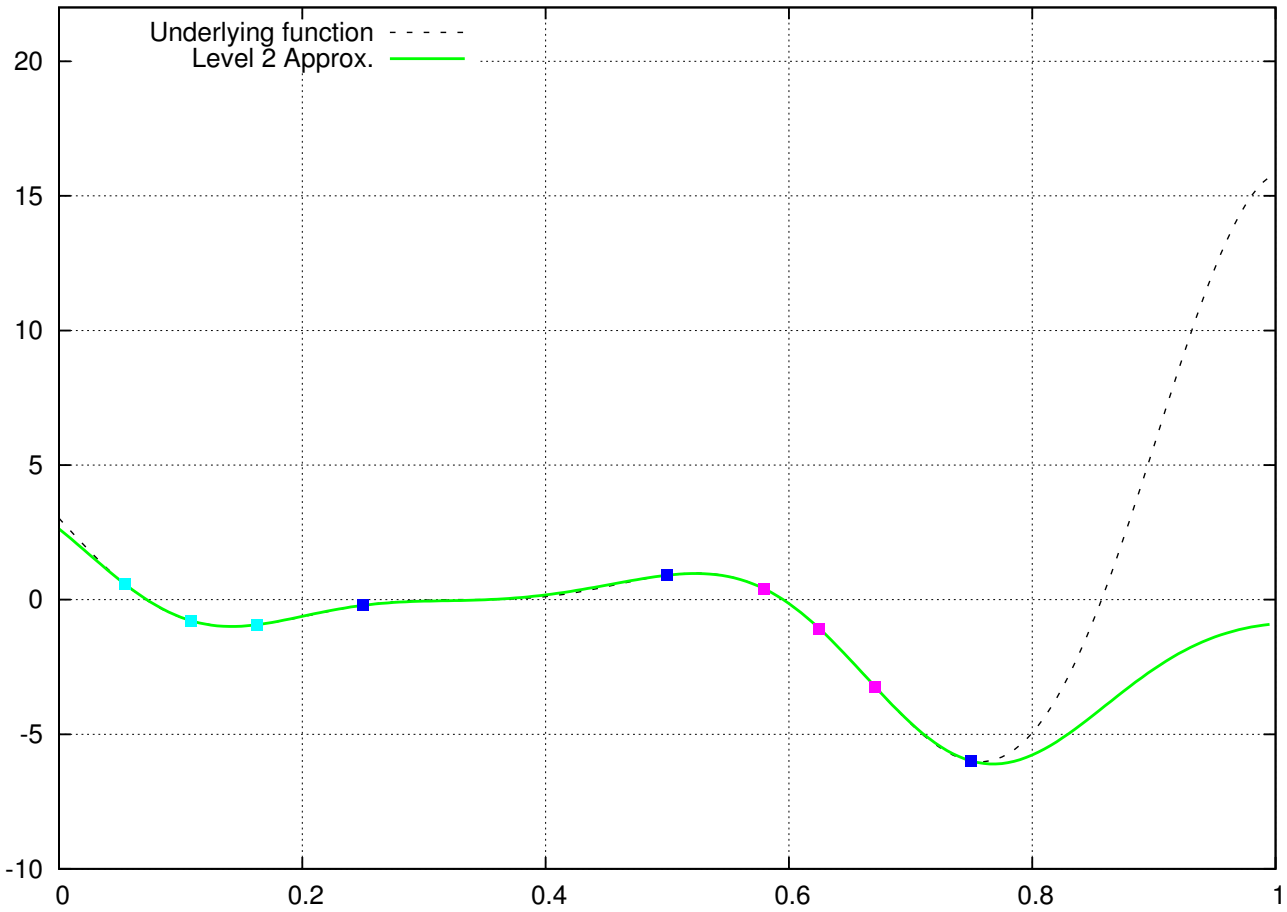
Hierarchical modeling of $(6x-2)^2 \sin[12x-4]$



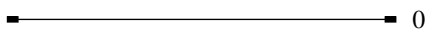
Model Scheme



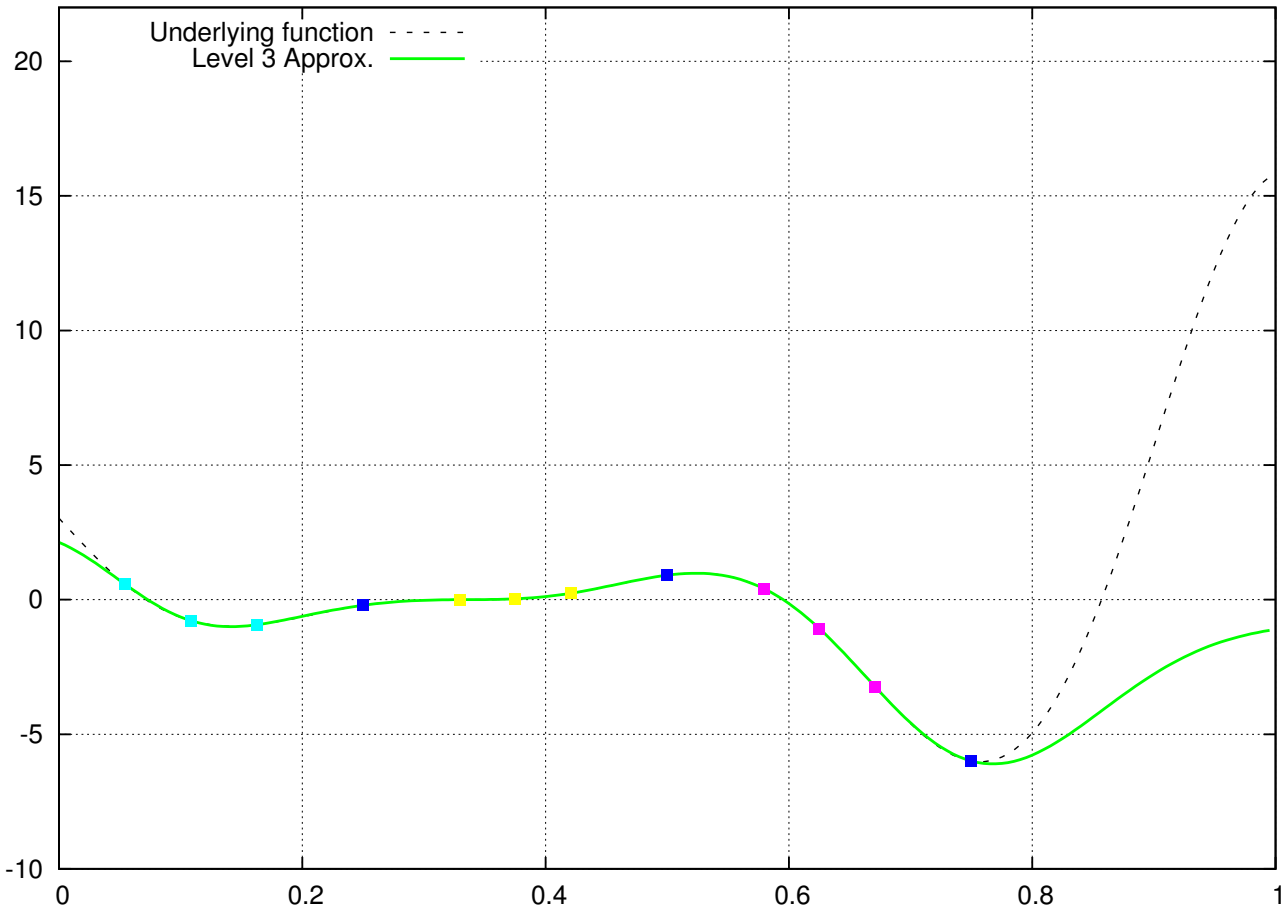
Hierarchical modeling of $(6x-2)^2 \sin[12x-4]$



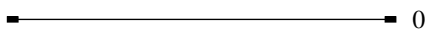
Model Scheme



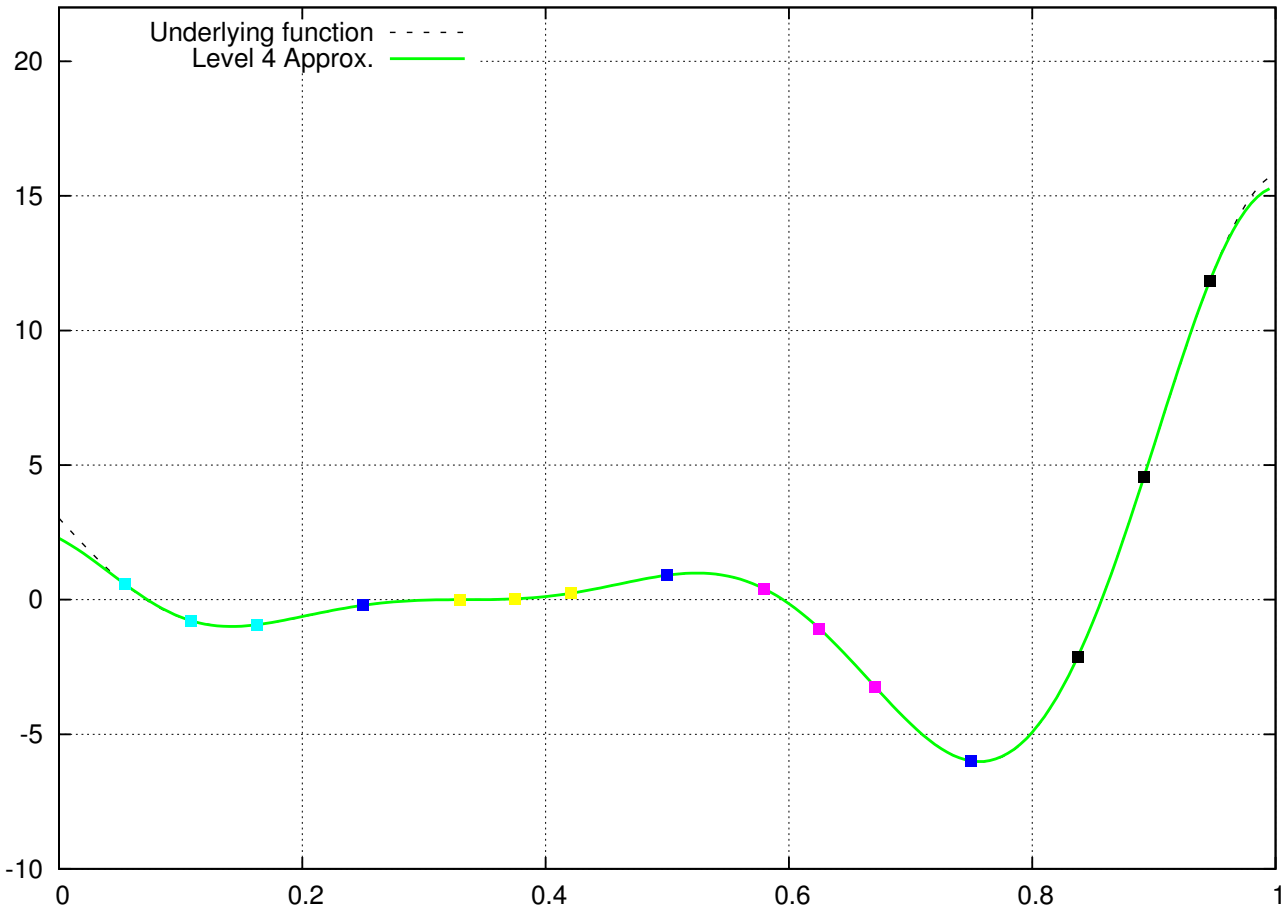
Hierarchical modeling of $(6x-2)^2 \sin[12x-4]$



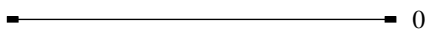
Model Scheme



Hierarchical modeling of $(6x-2)^2 \sin[12x-4]$



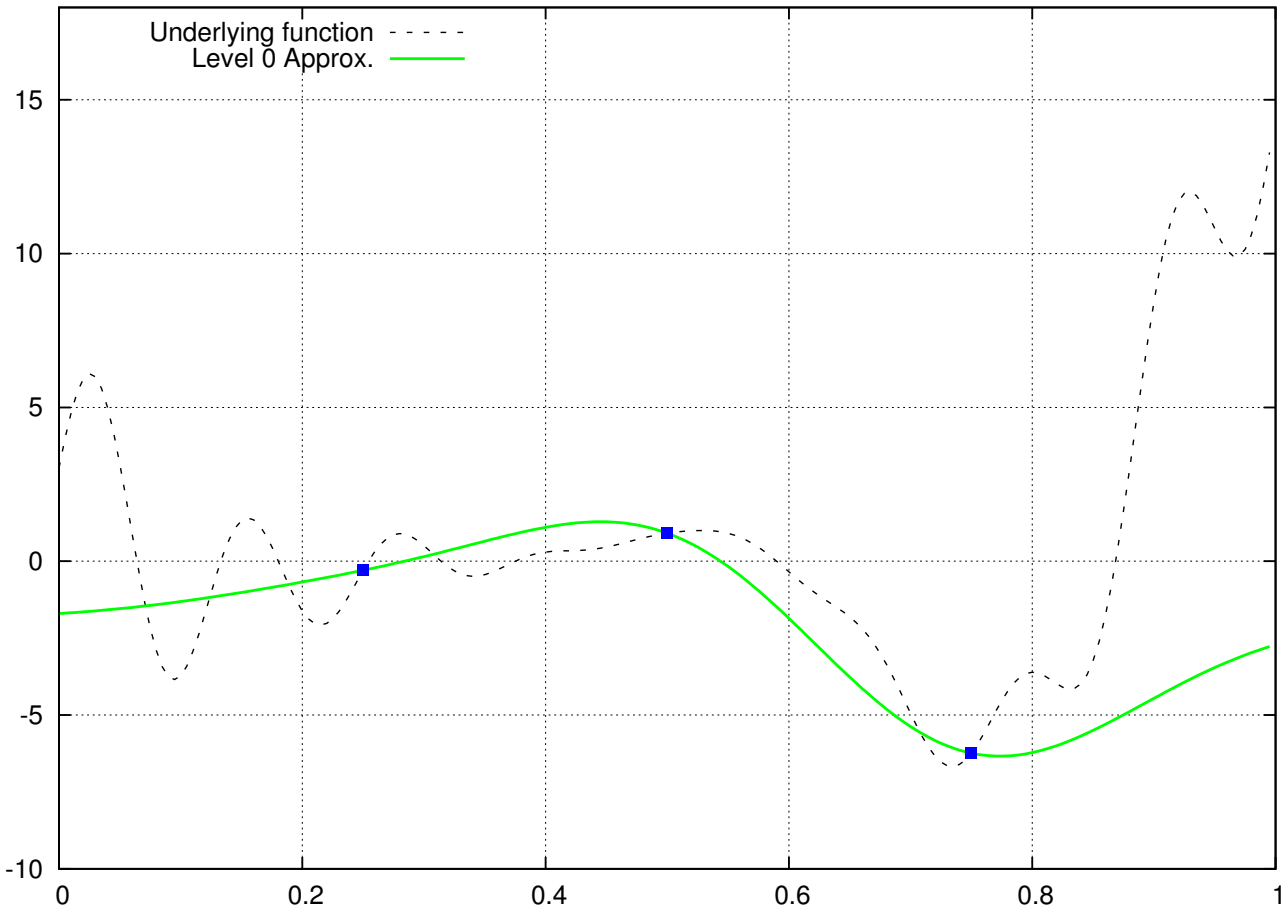
Model Scheme



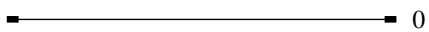
Notes:

- Conventional training process is **not** involved whatsoever
- Approximation quality is adequate at all steps
- Model scheme reflects “single” correlation length of input data
- Approximation changes **locally** upon insertion of new data points

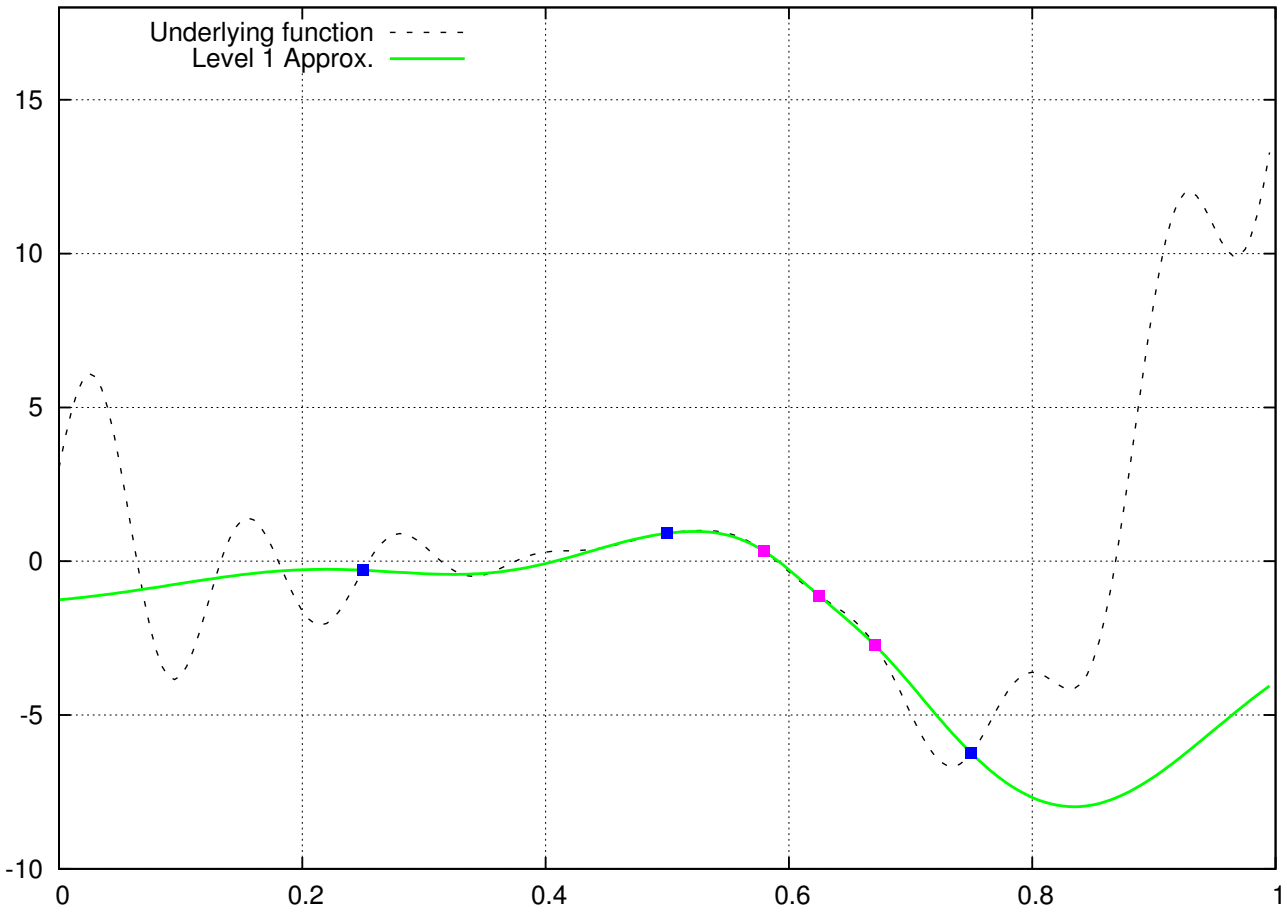
Treating $(6x-2)^2 \sin[12x-4] + 20(x-1/2)^2 \sin[50x]$



Model Scheme

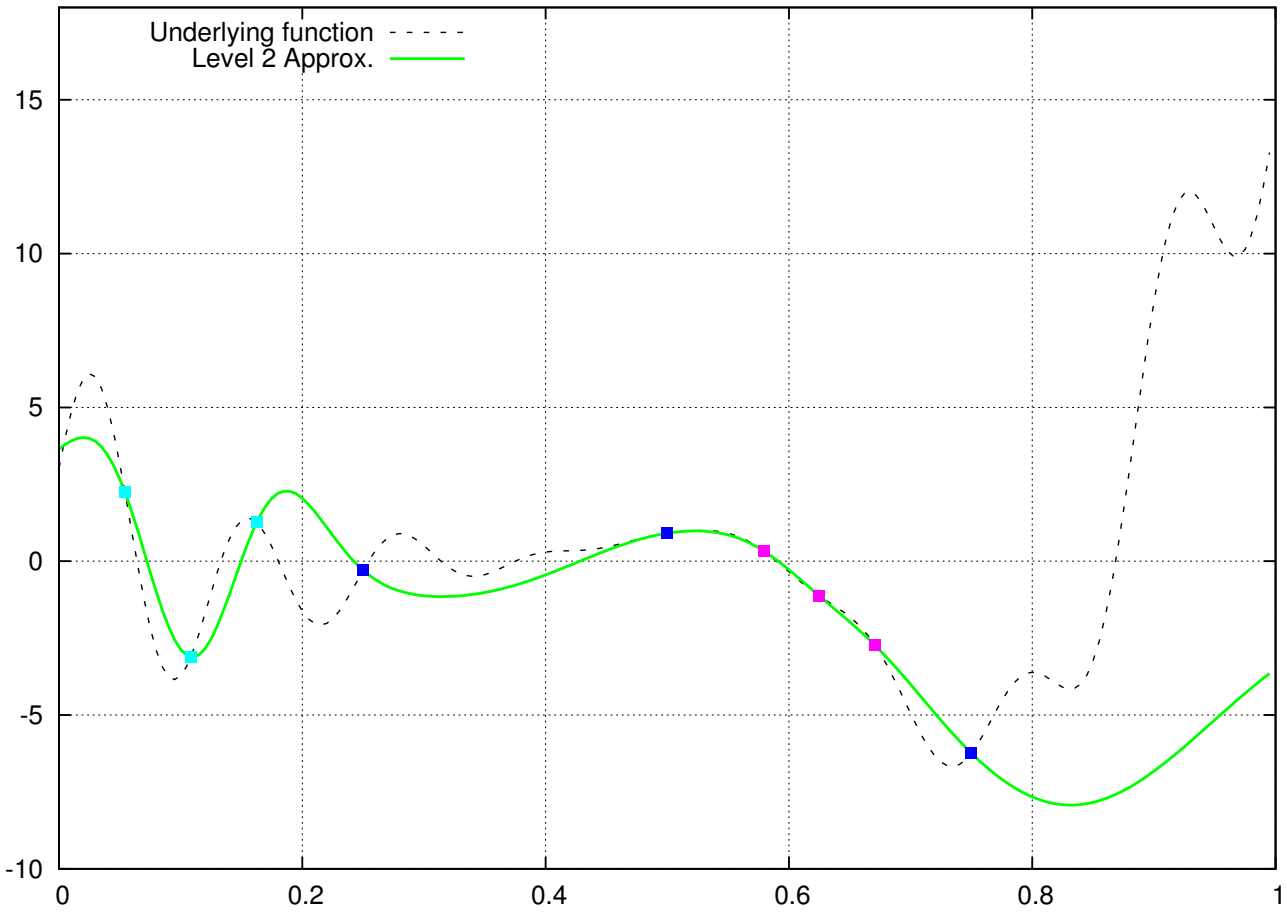


Treating $(6x-2)^2 \sin[12x-4] + 20(x-1/2)^2 \sin[50x]$



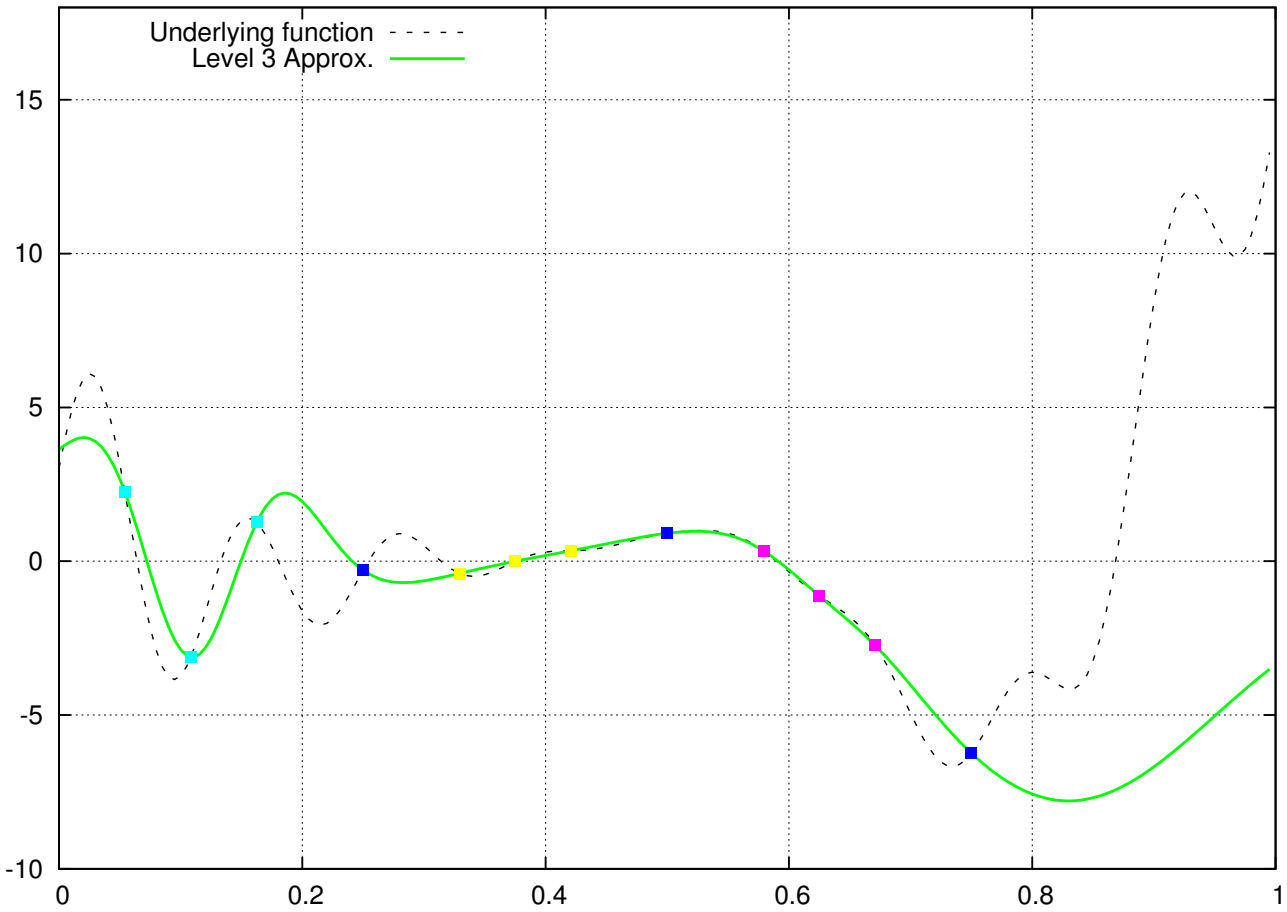
Model Scheme

Treating $(6x-2)^2 \sin[12x-4] + 20(x-1/2)^2 \sin[50x]$



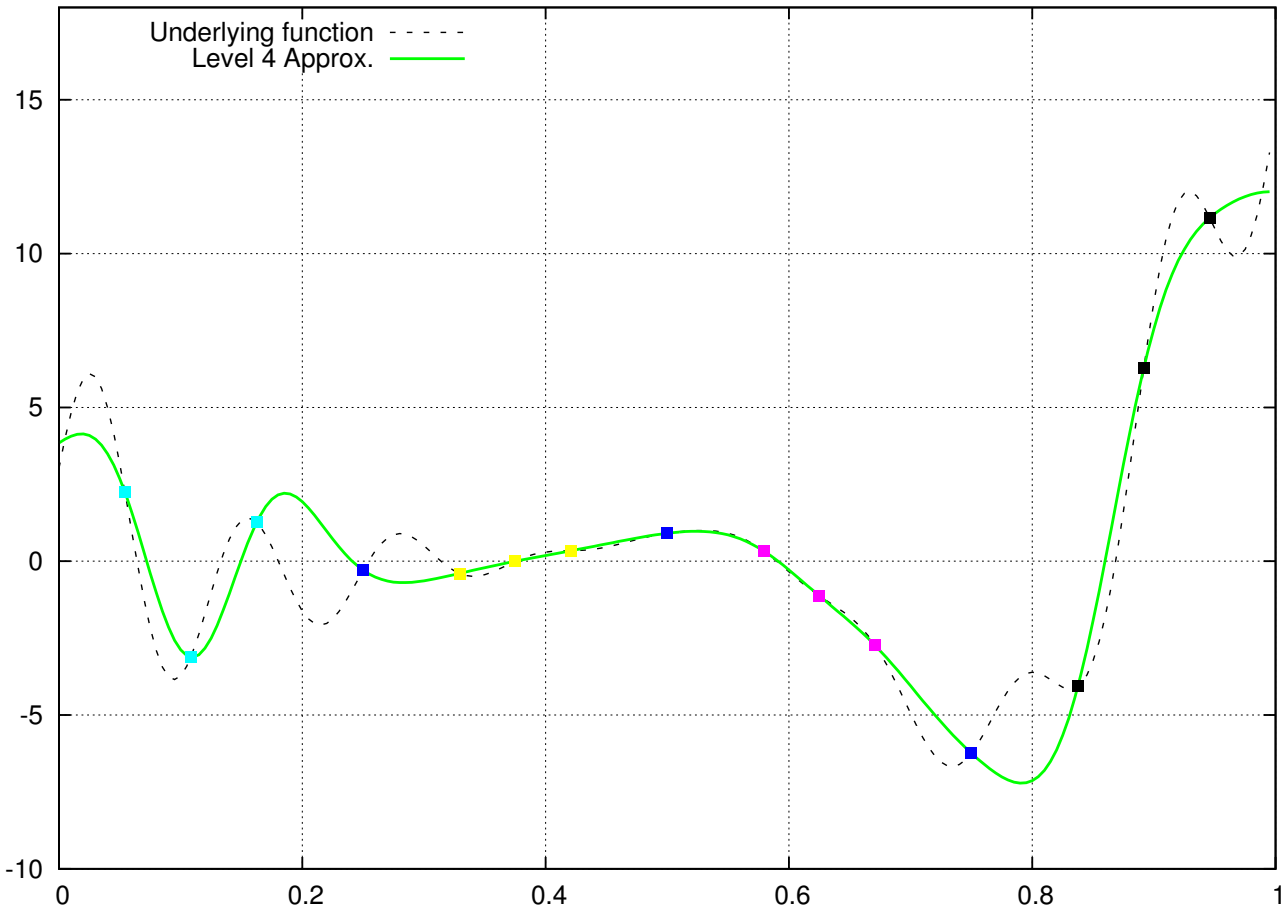
Model Scheme

Treating $(6x-2)^2 \sin[12x-4] + 20(x-1/2)^2 \sin[50x]$



Model Scheme

Treating $(6x-2)^2 \sin[12x-4] + 20(x-1/2)^2 \sin[50x]$



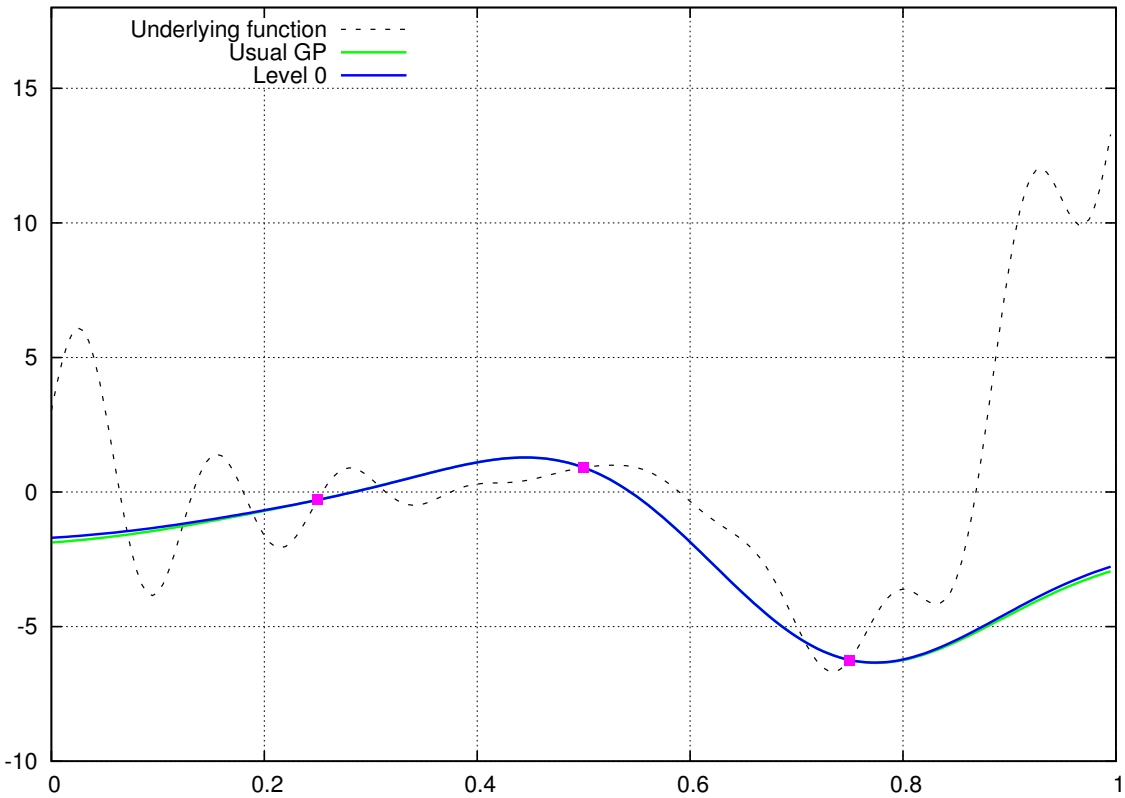
Model Scheme

Notes:

- Model scheme correctly reflects different correlation lengths of input data
- Approximation is adequate, moreover, it changes only locally upon the sample augmentation
- No usual training is involved

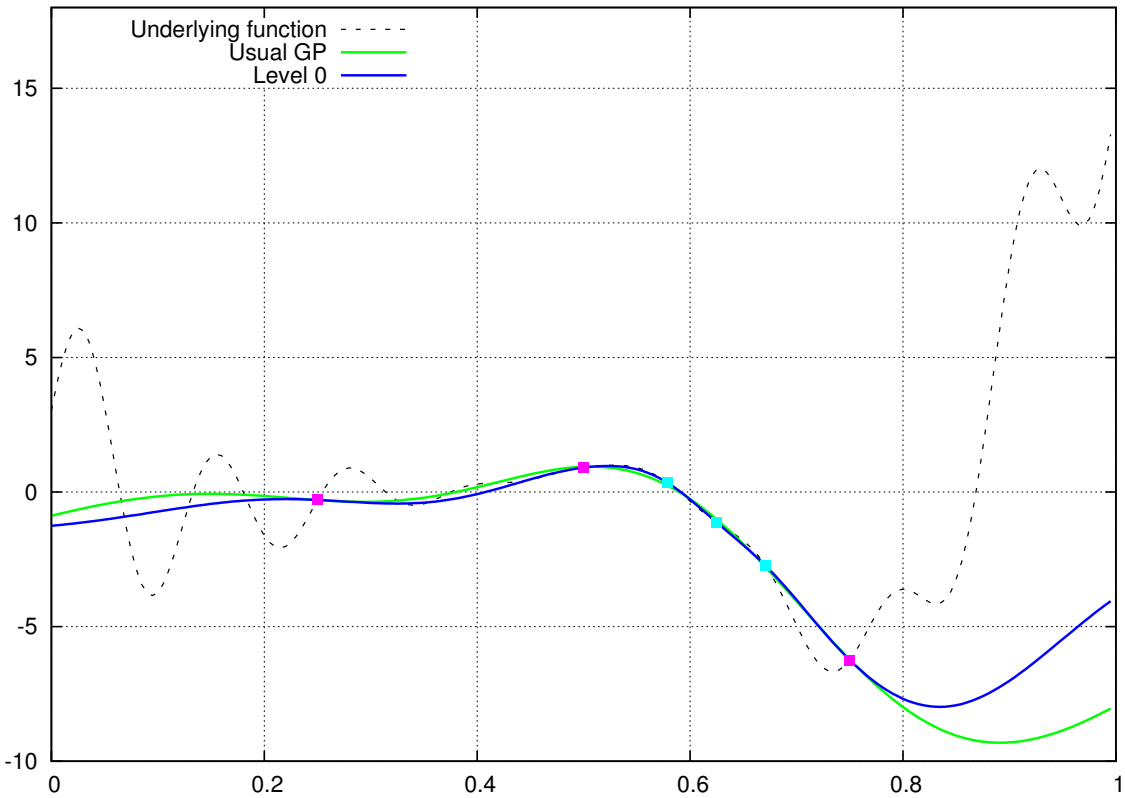
$$(6x-2)^2 \sin[12x-4] + 20(x-1/2)^2 \sin[50x]$$

Hierarchical treatment vs. usual GP



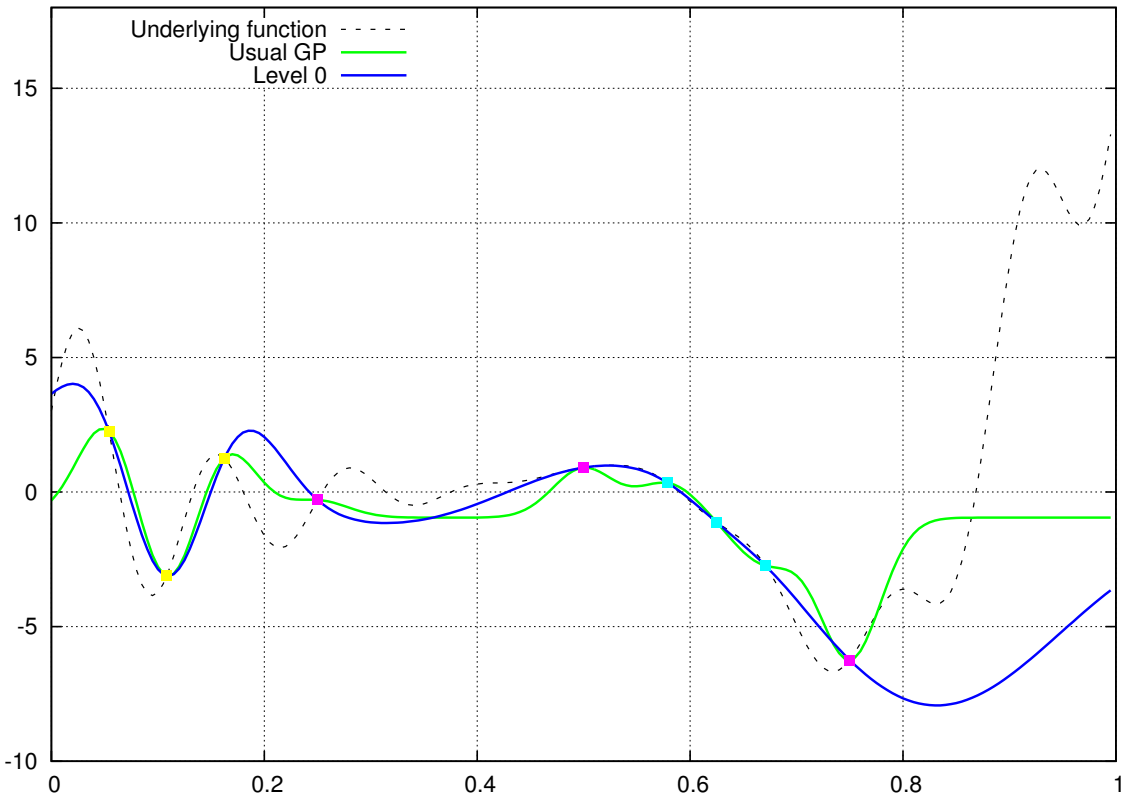
$$(6x-2)^2 \sin[12x-4] + 20(x-1/2)^2 \sin[50x]$$

Hierarchical treatment vs. usual GP



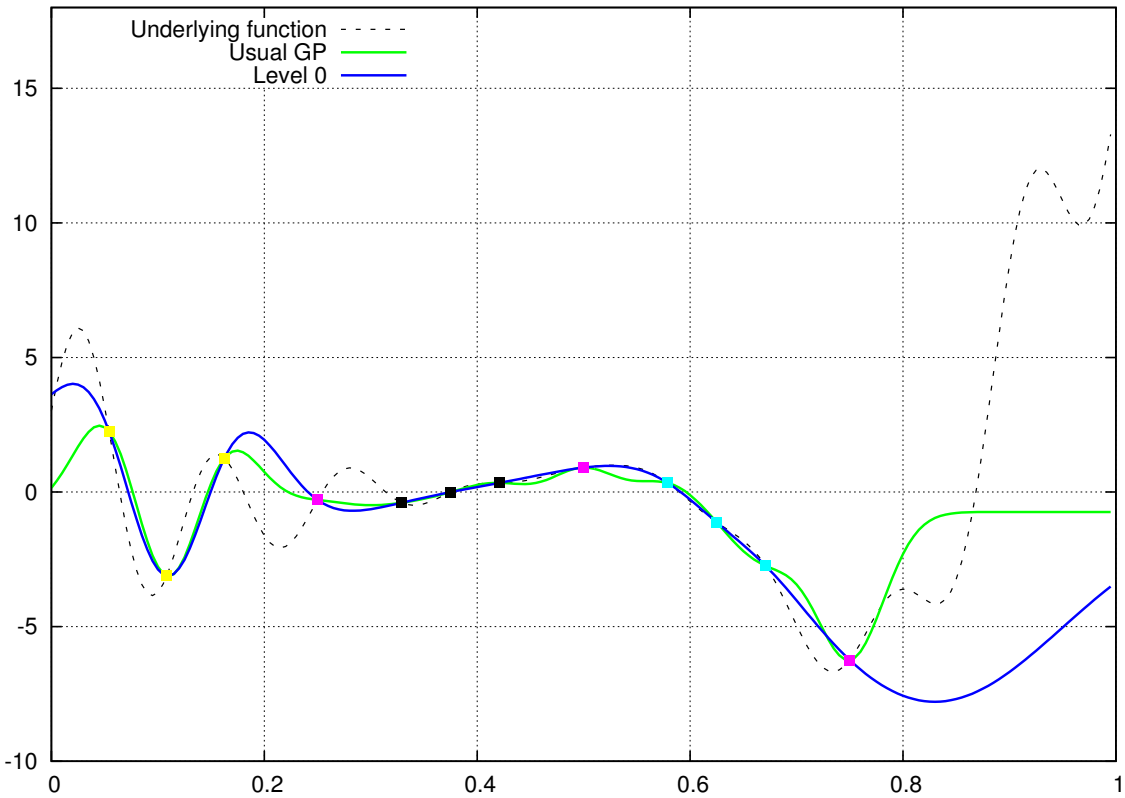
$$(6x-2)^2 \sin[12x-4] + 20(x-1/2)^2 \sin[50x]$$

Hierarchical treatment vs. usual GP



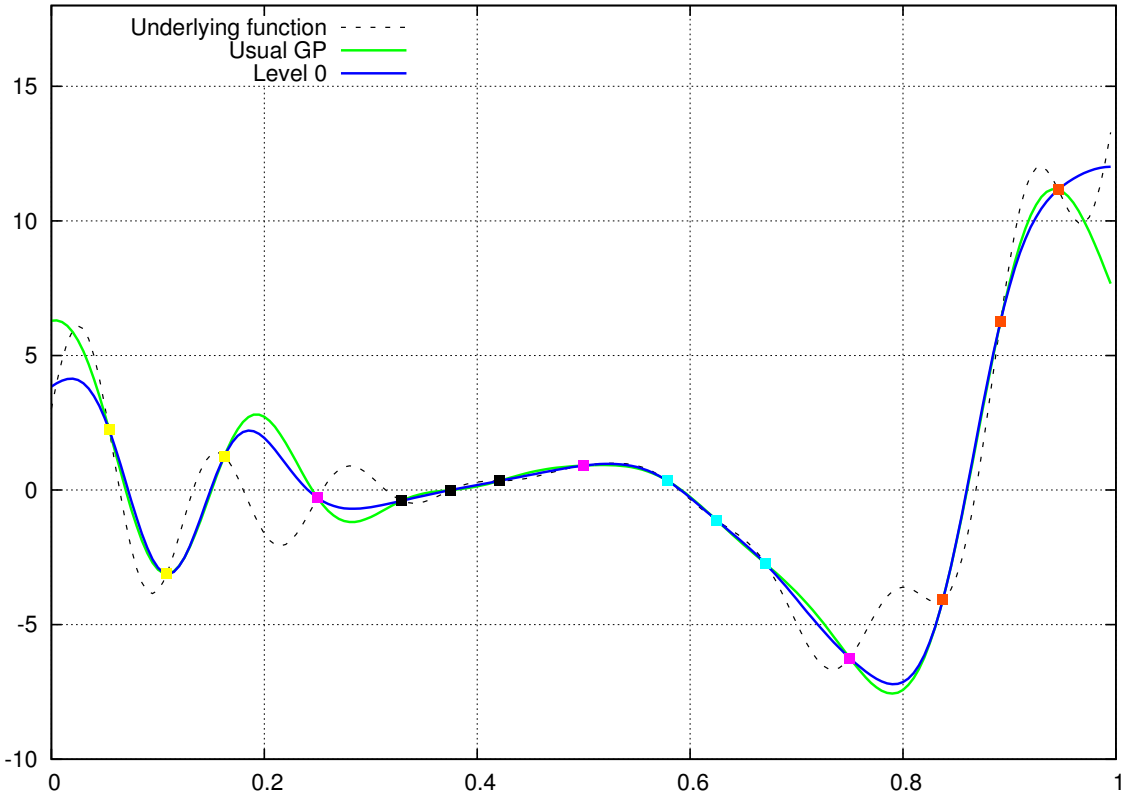
$$(6x-2)^2 \sin[12x-4] + 20(x-1/2)^2 \sin[50x]$$

Hierarchical treatment vs. usual GP



$$(6x-2)^2 \sin[12x-4] + 20(x-1/2)^2 \sin[50x]$$

Hierarchical treatment vs. usual GP



Notes:

- Conventional GP is **unstable** wrt sample augmentation: predictions might change globally upon only a few points insertion.