Tighter Bounds on the Log Marginal Likelihood for Gaussian Process Regression

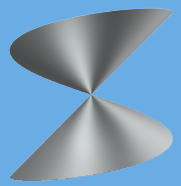
Artem Artemev*[1,2], David R. Burt*[3] , Mark van der Wilk[1]

[1]Imperial College London, [2]Secondmind, [3]University of Cambridge

ICML, 2021

Imperial College London

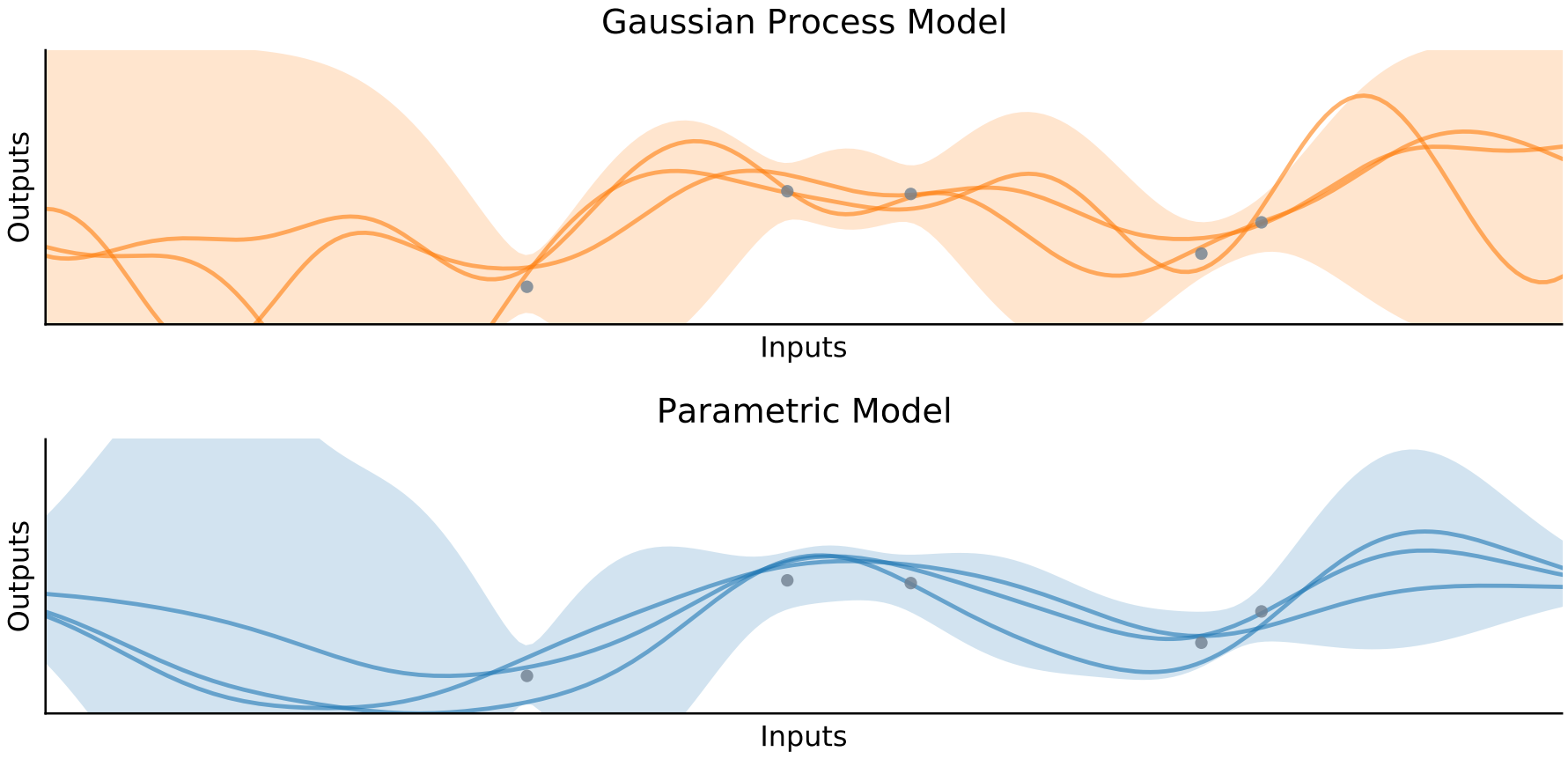
UNIVERSITY OF CAMBRIDGE



Secondmind

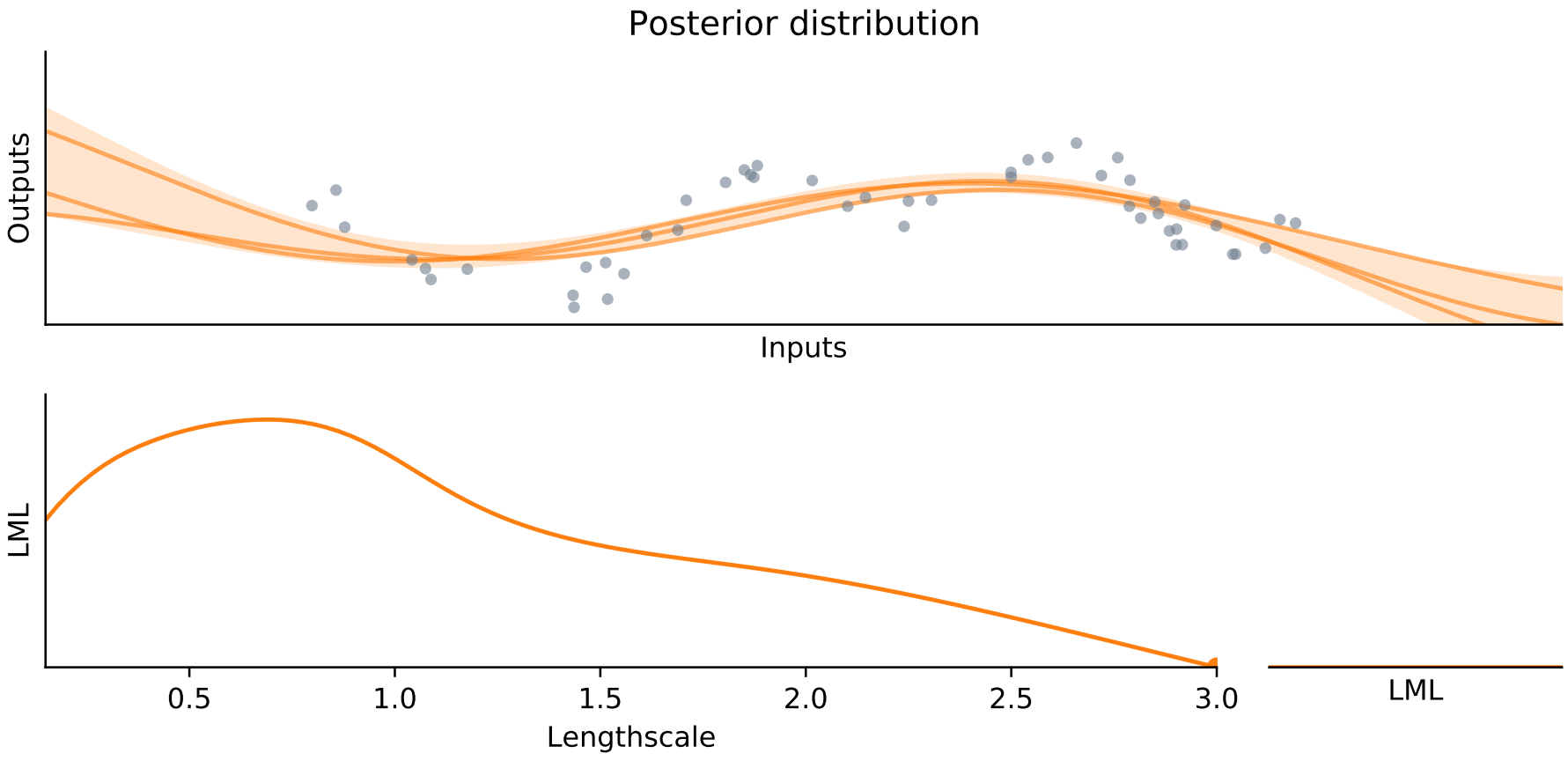
Advantages of Gaussian Process Regression

Non-parametric flexibility



- Automatically add capacity as new data is observed.
- Retain uncertainty in regions with little data.

Model selection with GPR



- LML can balance data-fit and model complexity. Maximize log marginal likelihood (LML) $\Rightarrow$ automatically select hyperparameters.

Approximating Gaussian Process Regression

Scalability concerns with GP regression

Do we really need to approximate? The posterior of a GP can be computed with linear algebra.

- Direct implementations involve computing and factoring kernel matrix (e.g. Cholesky) $\Rightarrow$ costly.
- Many approximations developed:
  - Variational (sparse) approximations (Titsias, 2009).
  - Iterative approximations using e.g. conjugate gradient (Gibbs and Mackay, 1997; Gardener, Pleiss, Bindel, Weinberger, Wilson, 2018).

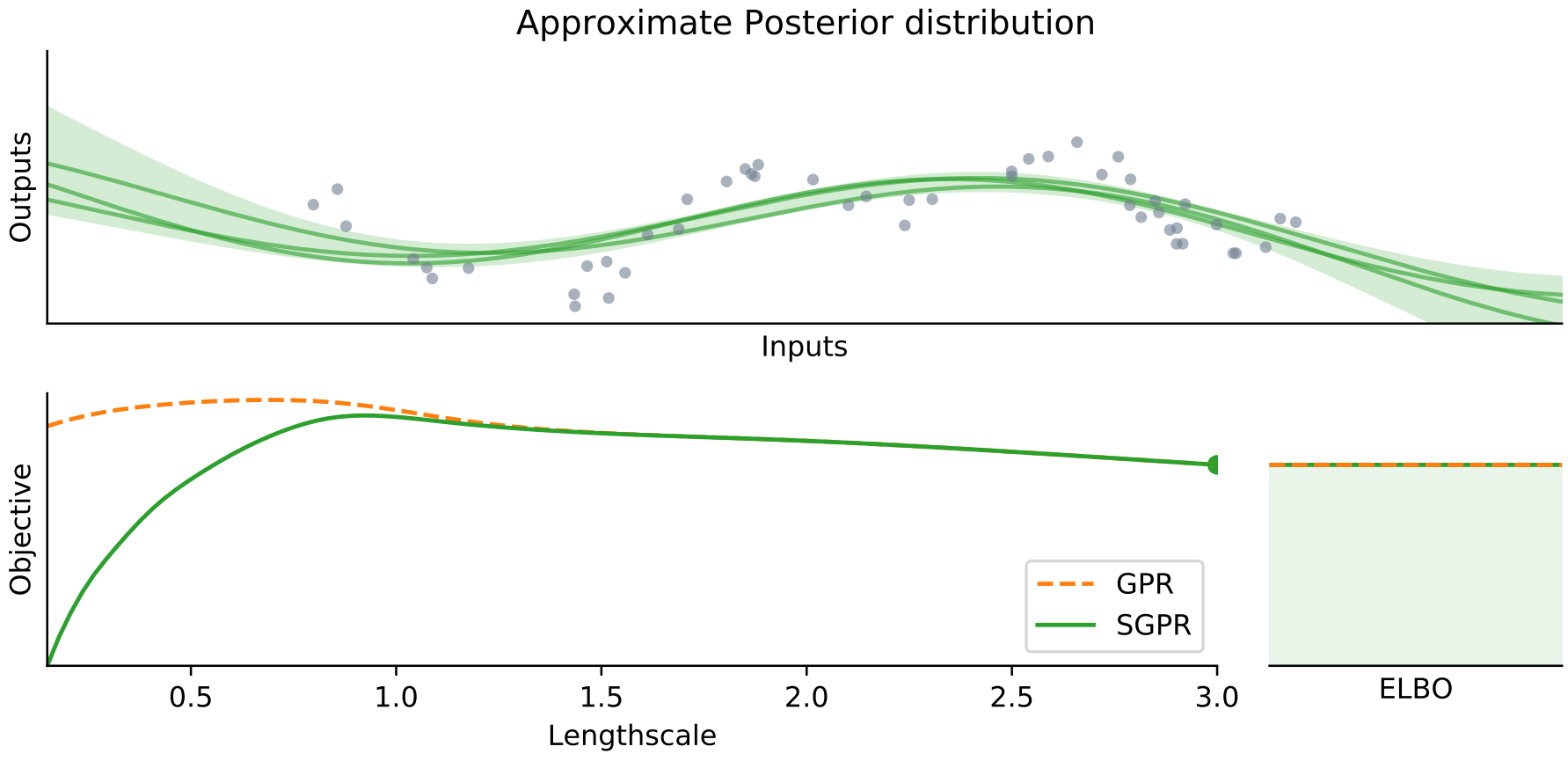
Desiderata for Approximate GP Regression

What do we want from a scalable GP approximation?

- Accurately approximate non-parametric posterior.
- Fast, easy, effective hyperparameter selection.
- Don't introduce lots of extra parameters that are hard to tune.

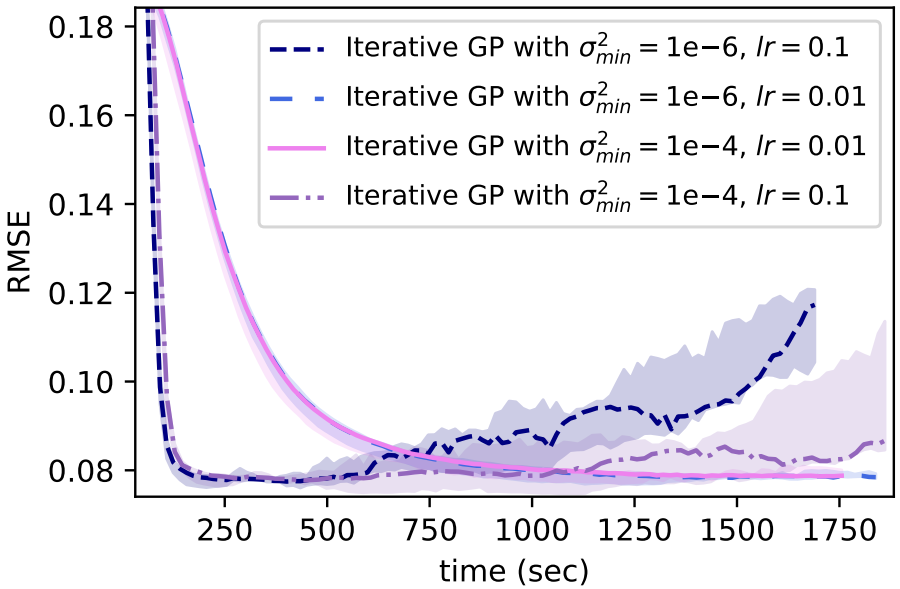
Sparse variational methods

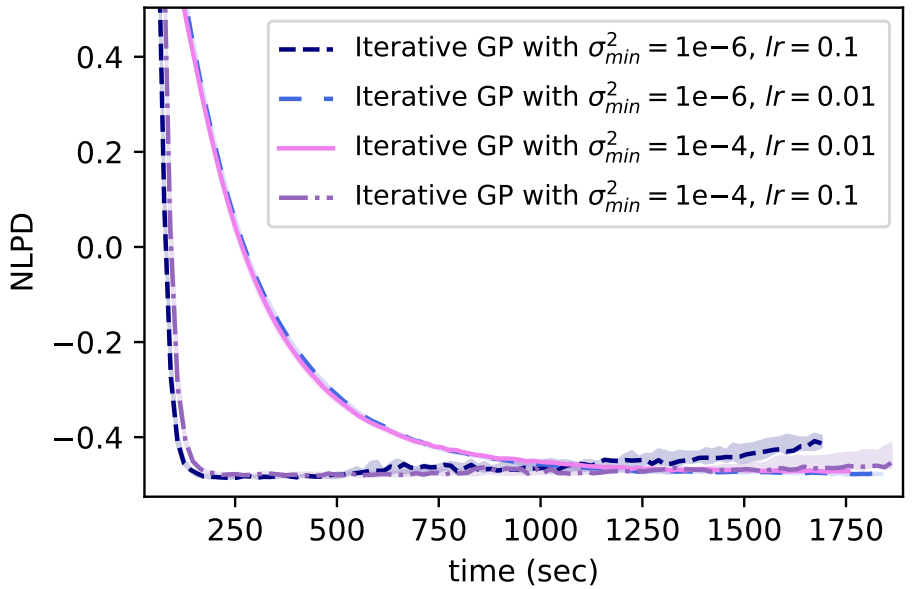
- Non-parametric posterior, but limited capacity mean function!
- Easy to select hyperparameters, but biased!
- Easy to choose approximation parameters $\Rightarrow$ Maximize ELBO.



Iterative methods

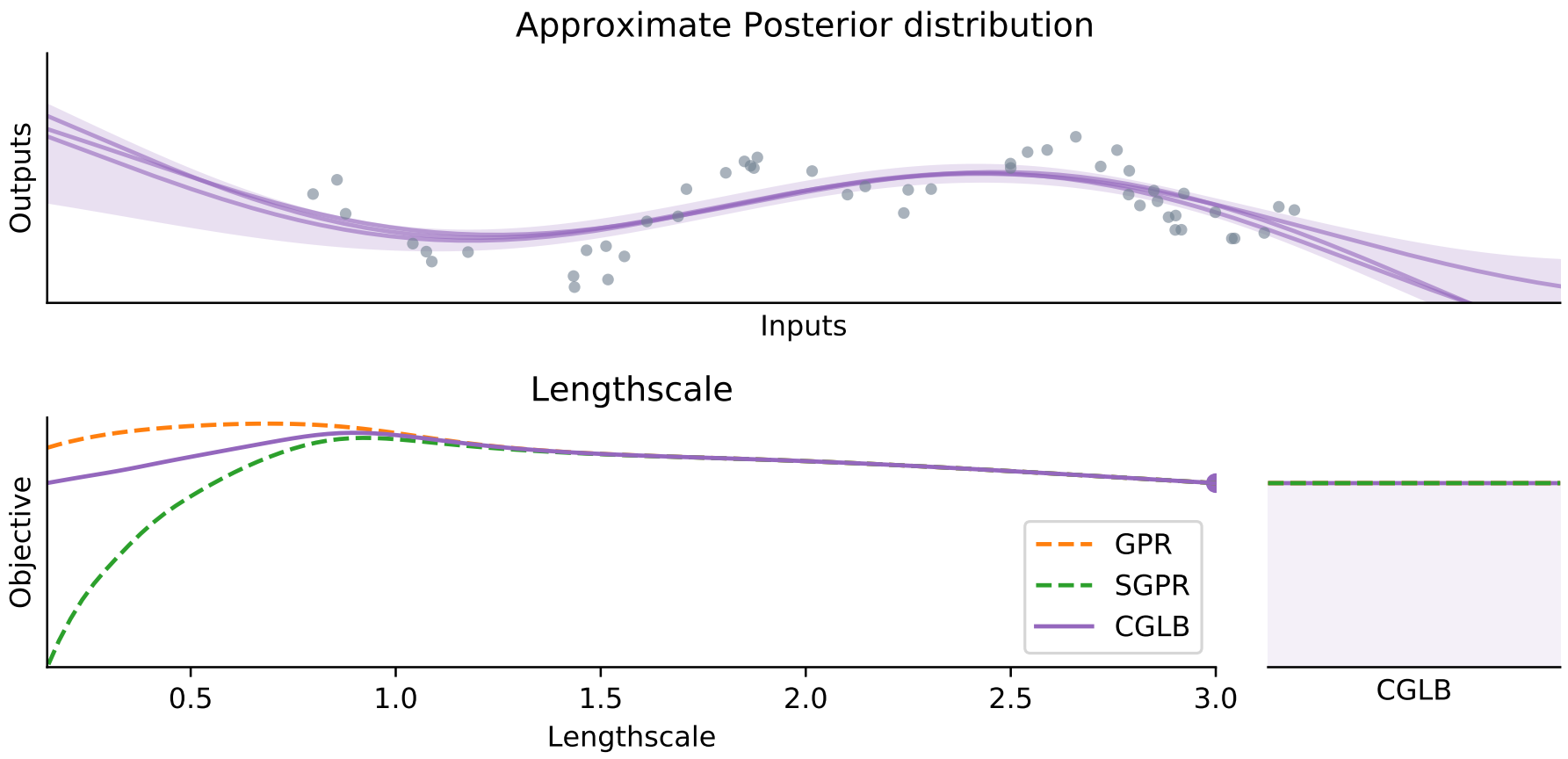
- Can give *very* accurate approximations to predictive mean.
- Bias can be small but
  - stochastic objective ⇒ slower convergence, more parameters to pick.
  - bias can be hard to assess.
- Setting approximation parameters is less automatic.





Conjugate Gradient Lower Bound

Combine training speed and reliability of SGPR with low bias and good predictive mean of iterative methods



The Conjugate Gradient Lower Bound

The log marginal likelihood

Gaussian Process regression LML:

$$\log p_Y(y;\theta) = c - \frac{1}{2}\underbrace{\log|\mathsf{K}|}_{\text{log det.}} - \frac{1}{2}\underbrace{y^\top \mathsf{K}^{-1} y.}_{\text{quad. term}} \tag{1}$$

- We will upper bound both terms individually.
- Sparse GPR ELBO:

$$\log|\mathsf{K}| \leq \log|\mathsf{Q}| + \frac{1}{\sigma^2}\mathrm{tr}(\mathsf{K}-\mathsf{Q}) \tag{2}$$

and

$$y^\top \mathsf{K}^{-1} y \leq y^\top \mathsf{Q}^{-1} y, \tag{3}$$

with Q a (specific) low-rank plus diagonal approximation to K.

Bounding the log determinant term

Technical Contribution 1 (Apply tighter bound to log-det.).

We use the arithmetic-geometric inequality ⇒ always at least as tight as SGPR bound (similar bound in Vakili, Khezeli, Picheny 2021).

Bound on $\log|\mathrm{K}| - \log|\mathrm{Q}|$

SGPR	$\frac{\mathrm{tr}(\mathrm{K}-\mathrm{Q})}{\sigma^2}$	$\mathcal{O}(nm^2)$	loose
O-SGPR[1]	$\mathrm{tr}\left(\mathrm{Q}^{-1}(\mathrm{K}-\mathrm{Q})\right)$	$\mathcal{O}(n^2m)$	tighter
CGLB	$n\log\left(1+\frac{\mathrm{tr}(\mathrm{K}-\mathrm{Q})}{n\sigma^2}\right)$	$\mathcal{O}(nm^2)$	tighter
CGLB-expensive	$n\log\left(\frac{\mathrm{tr}\left(\mathrm{Q}^{-1}\mathrm{K}\right)}{n}\right)$	$\mathcal{O}(n^2m)$	tightest

[1] Shi, Titsias, Mnih, 2020.

Bounding the quadratic term

Technical Contribution 2 (New bound on quadratic term).

Derive a bound on quadratic term that is tight if

- *Works if SGPR would work OR*
- *If iterative method would work AND*
- *Allows us to determine when we should stop CG.*

Derivation:

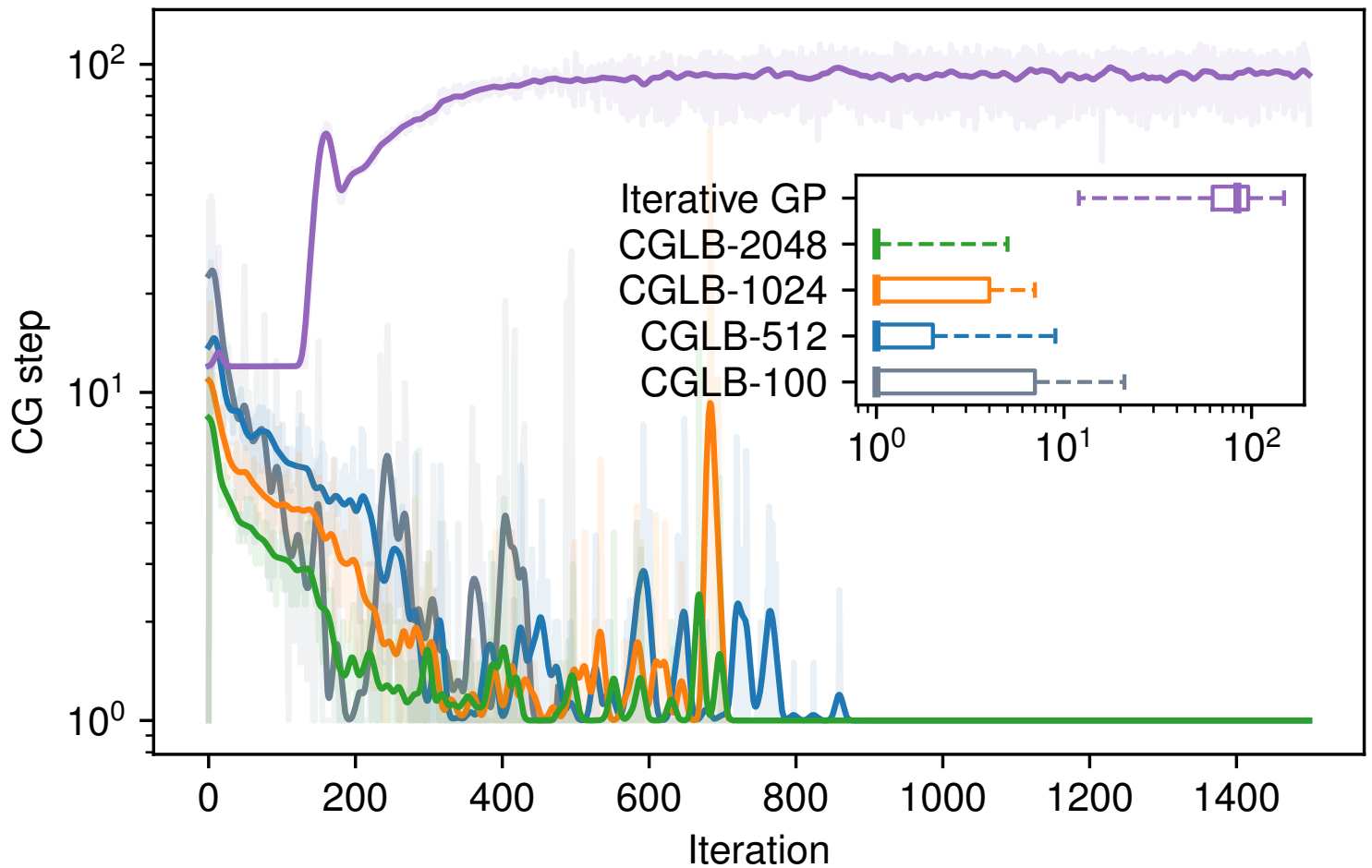
- Introduce auxiliary parameter $v \in \mathbb{R}^n$ as an approximation to $\mathrm{K}^{-1}y$.

Let $r := y - \mathrm{K}v$ be the residual, then

$$\begin{aligned} y^\top \mathrm{K}^{-1} y &= (r + \mathrm{K}v)^\top \mathrm{K}^{-1}(r + \mathrm{K}v) \\ &= r^\top \mathrm{K}^{-1} r + 2r^\top v + v^\top \mathrm{K} v \\ &\leq r^\top \mathrm{Q}^{-1} r + 2r^\top v + v^\top \mathrm{K} v. \end{aligned}$$

Improving the inner loop of conjugate gradient

- Terminate CG based on a criteria that tells us running it more steps could only improve the bound by ϵ.
- Restart CG solution at solution found in last iteration.



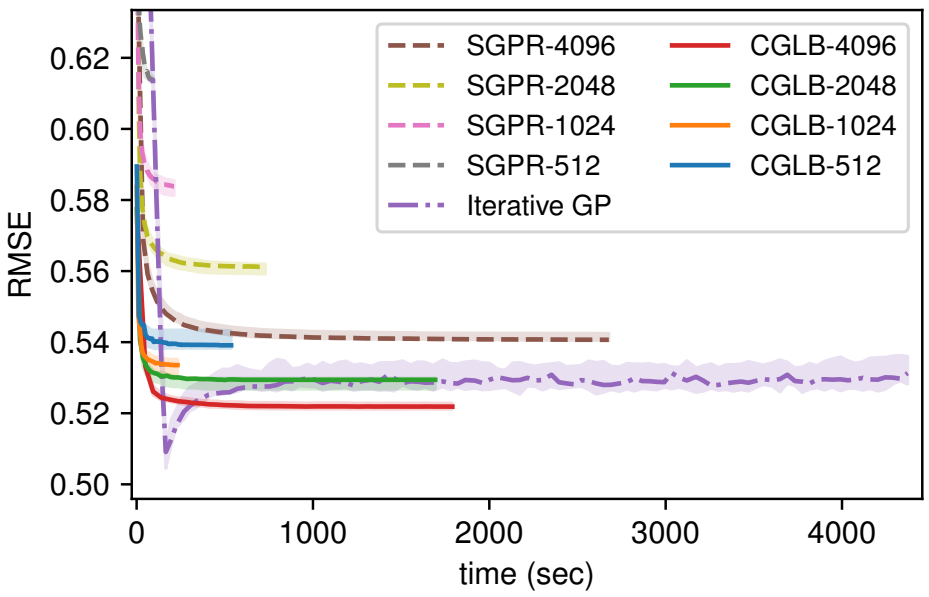
Choosing approximation parameters

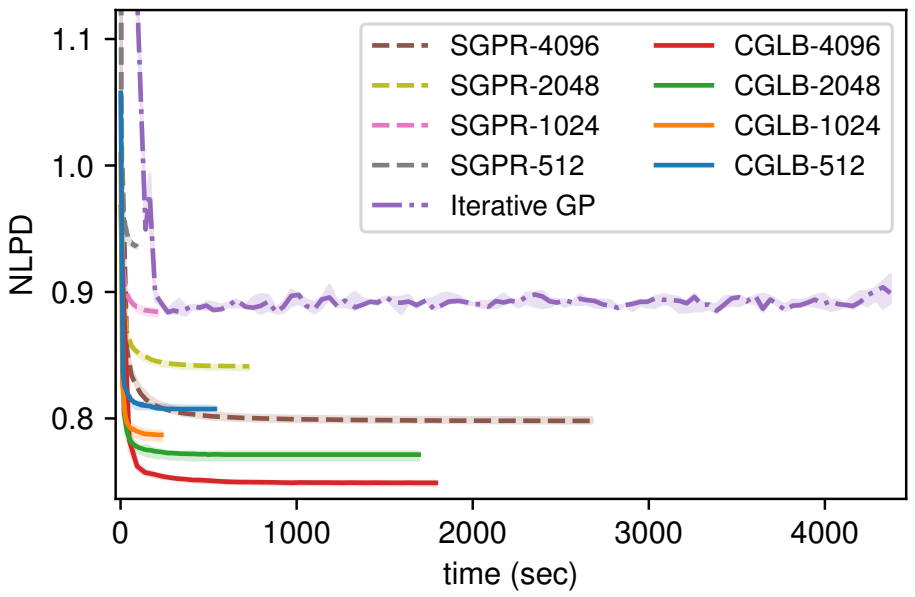
- Rank of Q: Like SGPR, higher rank is better. But low-rank doesn't result in as much bias as SGPR.
- Criterion for stopping CG: Directly relates to objective.
- Optimization procedure: L-BFGS converges quickly and has established default settings. No need to tune learning rate.

Comparing CGLB to Sparse and Iterative Methods

Performance on real data:

- CGLB enjoys fast convergence, good predictive performance and is easy to tune.





Conclusions and an open problem

CGLB combines many of the benefits of SGPR and iterative methods.

- Reduce bias relative to SGPR using better bounds and CG.
- Deterministic lower bound objective $\Rightarrow$ fast and stable training.
- Accurate mean approximation.

Still several limitations to overcome:

- Better posterior covariance estimates?
- Can an iterative approach refine the log determinant bound farther?

Open Problem.

We derived a lower bound on the LML. Does there exist a family of posterior distributions such that this lower bound is the corresponding ELBO?