This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

SAND2022-13473C

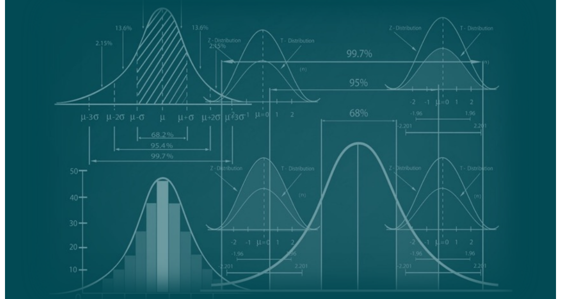
Enhancing Polynomial Chaos Expansion Surrogates Through Probabilistic Transfer Learning

SIAM Conference on Uncertainty Quantification

4/15/2022

PRESENTED BY

Moe Khalil
Sandia National Laboratories, Livermore
mkhalil@sandia.gov



Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

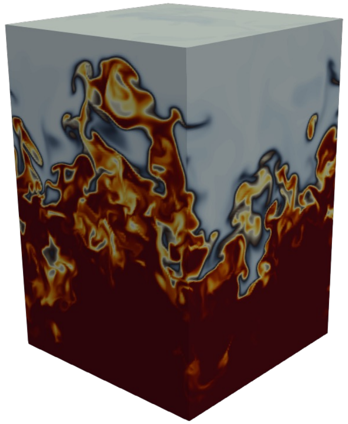
TL team

- **Sandia LDRD:**
  - Moe Khalil (Sandia): principal investigator, algorithm development
  - Reese Jones (Sandia): algorithm/software development – tempering transformations and objective functions
  - Uma Balakrishnan (Sandia): software development, numerical experiments
  - Wyatt Bridgman (Sandia): algorithm/software development – multi-fidelity inverse UQ
  - Jackie Chen, Bruno Soriano, Kisung Jung (Sandia): TL application in turbulent combustion modeling
  - Tarek Echekki (North Carolina State University): TL application in turbulent combustion modeling
  - Prof. Chris Pettit (US Naval Academy): TL for DoD relevant applications in atmospheric acoustics
  - Erin Swansen and Prof. Naira Hovakimyan (University of Illinois at Urbana–Champaign): TL application for adaptive control of urban air mobility

- **SBIR subcontract (Cyentech Consulting LLC, Houston):**
  - Moe Khalil (Sandia): principal investigator, algorithm development
  - Cosmin Safta (Sandia): algorithm/software development
  - Yueqin Huang (Cyentech Consulting LLC): customer - TL application for subsurface characterization
  - Xuqing Wu (University of Houston): academic collaborator, consultant
  - Jiefu Chen (University of Houston): academic collaborator, consultant
  - Han Lu (University of Houston): graduate student, algorithm/software development

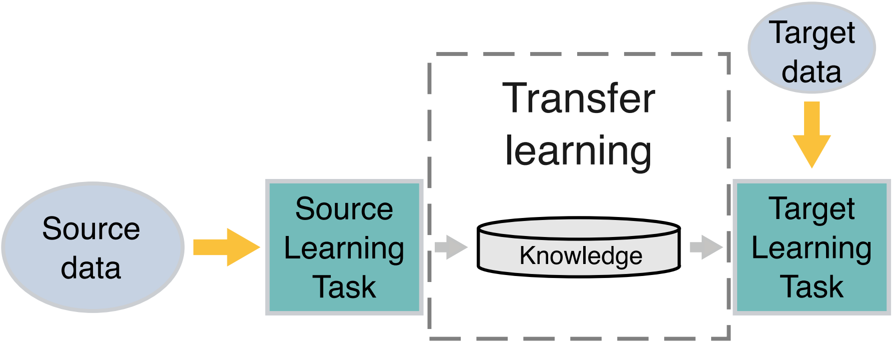
Alleviating Ill-Posedness in Training of ML Models with Sparse Data

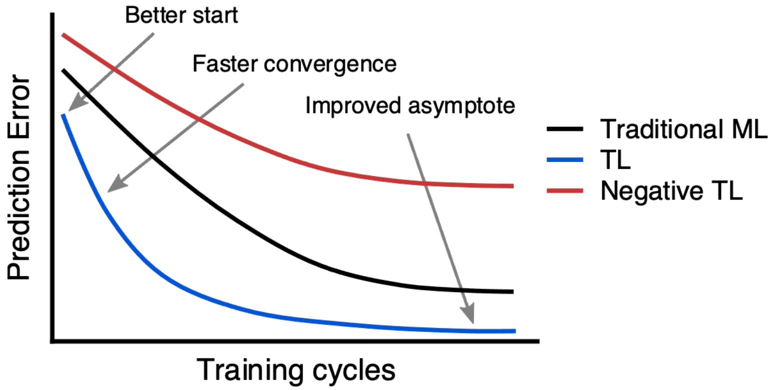
- **R&D Drivers:** Increasing attempts to deploy AI/ML for various scientific and engineering modeling tasks
- **Challenges:** Many application domains are defined by ***data sparsity***, effectively precluding the use of many predictive machine learning models:
  - ✗ Expensive computer simulations (e.g. Exemplar in Turbulent Combustion)
  - ✗ Sparse experimental data
  - ✗ Limited access to classified/sensitive data



Data acquisition cost: ~10 million CPU-hours per turbulent combustion simulation!

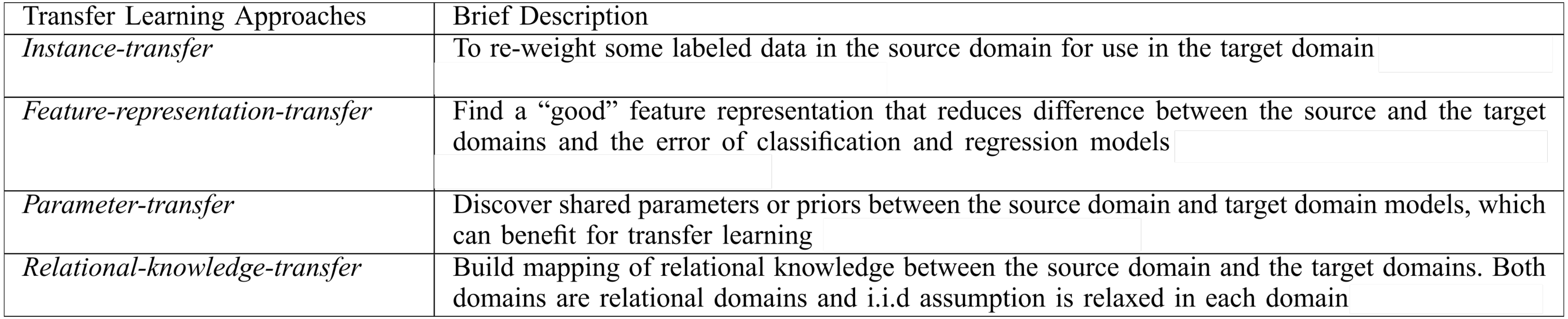
R&D Outcome: Novel ***probabilistic transfer learning*** (TL) framework that alleviates data sparsity while addressing key gaps in state-of-the-art TL techniques





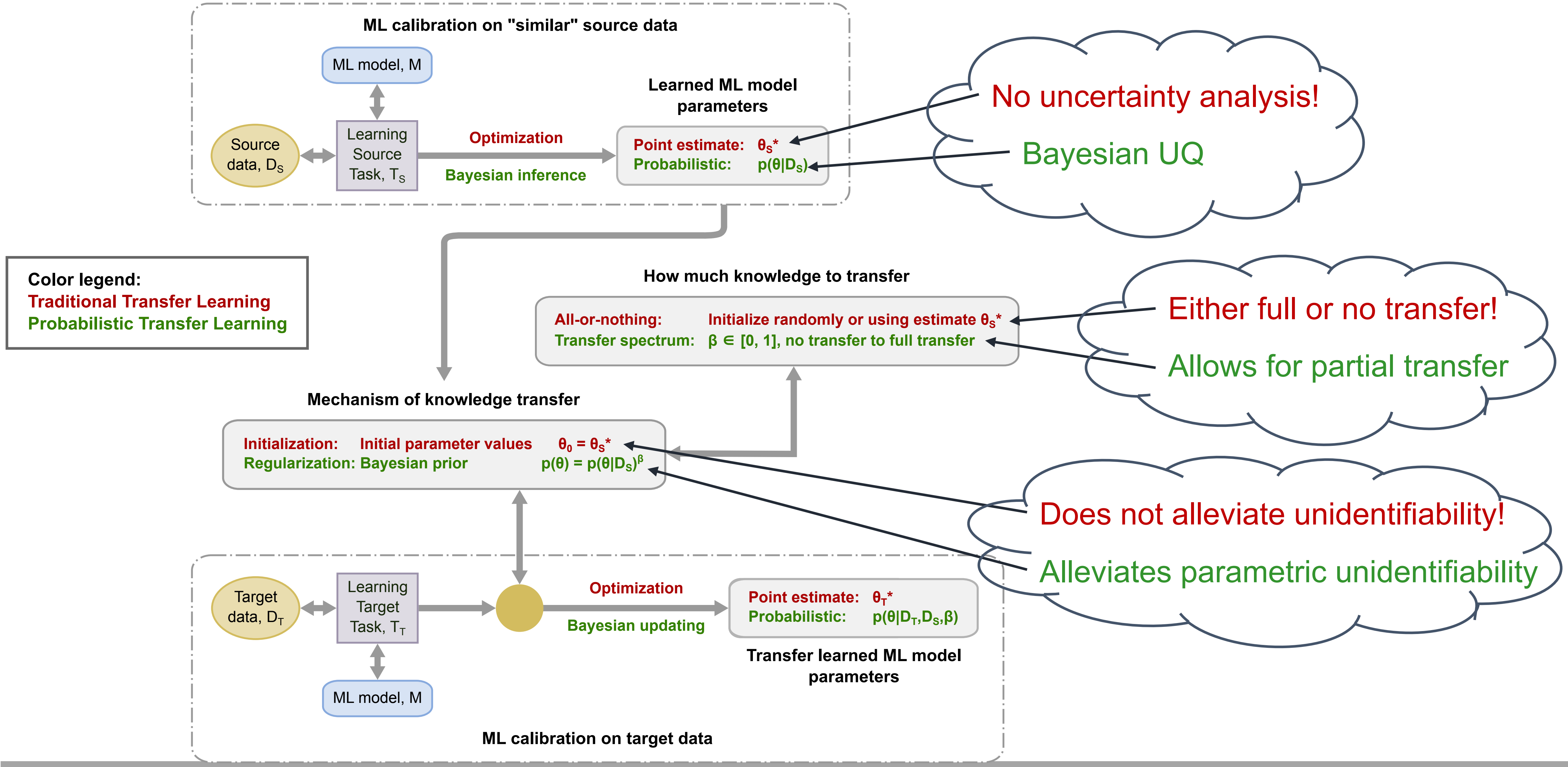
Shortcomings of existing TL methodologies

- State-of-the-art algorithms in TL
  - ✗ tend to be ML model-specific [George et al., 2018]
  - ✗ do not consider all (if any) types of uncertainties (data, parametric, model-form/fidelity) [Colbaugh et al., 2017, Raina et al., 2006]
  - ✗ use simplified (i.e. Gaussian) probability representations of data [Karbalayghareh et al., 2018].
- Most importantly, existing methods do not address key questions relating to
  - ✗ when it is worth applying TL (as opposed to traditional ML)
  - ✗ which ML model to use in TL (out of a set of plausible ones)
  - ✗ how much knowledge is to be transferred in order to safeguard against negative learning



Transfer Learning Approaches	Brief Description
Instance-transfer	To re-weight some labeled data in the source domain for use in the target domain
Feature-representation-transfer	Find a "good" feature representation that reduces difference between the source and the target domains and the error of classification and regression models
Parameter-transfer	Discover shared parameters or priors between the source domain and target domain models, which can benefit for transfer learning
Relational-knowledge-transfer	Build mapping of relational knowledge between the source domain and the target domains. Both domains are relational domains and i.i.d assumption is relaxed in each domain

Traditional vs Proposed Probabilistic Transfer Learning



ML calibration on "similar" source data

ML model, M

Source data, D_S

Learning Source Task, T_S

Optimization

Bayesian inference

Learned ML model parameters

Point estimate: θ_S^*

Probabilistic: $p(\theta|D_S)$

No uncertainty analysis!

Bayesian UQ

Color legend:

Traditional Transfer Learning

Probabilistic Transfer Learning

How much knowledge to transfer

All-or-nothing: Initialize randomly or using estimate θ_S^*

Transfer spectrum: $\beta \in [0, 1]$, no transfer to full transfer

Either full or no transfer!

Allows for partial transfer

Mechanism of knowledge transfer

Initialization: Initial parameter values $\theta_0 = \theta_S^*$

Regularization: Bayesian prior $p(\theta) = p(\theta|D_S)^\beta$

Does not alleviate unidentifiability!

Alleviates parametric unidentifiability

Target data, D_T

Learning Target Task, T_T

ML model, M

Optimization

Bayesian updating

Point estimate: θ_T^*

Probabilistic: $p(\theta|D_T,D_S,\beta)$

Transfer learned ML model parameters

ML calibration on target data

Technical Approach

- The proposed framework comprises of four inter-related tasks:
    - Capturing the knowledge to be transferred in training on source data
        - Provide flexibility in capturing PDFs
            - Low-fidelity Gaussian approximations (obtained using, for example, variational inference)
            - High-fidelity Gaussian mixture-models (GMM), able to characterize general non-Gaussian PDFs while enabling analytical scrutiny of the Bayesian framework.
        - Result in a spectrum of performance gains in TL
    - Propagating the knowledge to be transferred to target training tasks
        - Achieved via extensions of sequential (Bayesian) data assimilation
        - Rely on prior PDF tempering transformations (more on this later)
    - Determining how much knowledge to transfer given a choice of tempering transformation
        - Hierarchical or empirical Bayesian approaches for (joint) inference of tempering hyper-parameters
        - Information-theoretic measures; similarity and distance metrics
    - Selecting optimal ML model to use in TL
        - Probabilistic TL framework facilitates the use of Bayesian techniques for optimal model selection
        - Investigate feasibility of enhancing model complexity by leveraging Relevance Vector Machine learning techniques

Training of ML Models – Bayesian Approach

$$\mathcal{M}(x,\theta) = y \approx d + \epsilon$$

ML model

features

parameters

target

observation

noise

- Forward Problem: Given ML model, $\mathcal{M}$, model parameters, θ, and feature vector, x, predict “clean” targets, y
- Inverse Problem: Given a set of “noisy” observations, $D = \{d_1, \dots, d_N\}$, and feature vectors, $X = \{x_1, \dots, x_N\}$, infer parameters
    - Observations are
        - inherently noisy with unknown (or weakly known) noise model
        - sparse in space and time (insufficient resolution)
    - Problem typically ill-posed, i.e. no guarantee of solution existence nor uniqueness
- Solution: Probability density function (PDF) over the parameter space obtained using Bayes’ rule:

likelihood

prior

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

posterior

evidence

- $p(\theta)$ is the prior PDF of θ: describes prior knowledge, inducing regularization
- $p(d|\theta)$ is the likelihood PDF of θ: describes data fit
- $p(\theta|d)$ is the posterior PDF of θ: full Bayesian solution
    - Not a single point estimate
    - Completely characterizes the uncertainty in θ
    - Subsequently used in making predictions under uncertainty

Transferal of "Learning" in a Probabilistic Setting

- In a transfer learning context, we have a target task of interest (regression/classification) with associated target data $\{D_T, X_T\}$. We also have access to "supplementary" source data $\{D_S, X_S\}$.
- Extending on mechanisms of propagating knowledge in sequential data assimilation (e.g. Kalman-based filters), we can take the captured knowledge from the source data in the form of the likelihood function and use it as prior knowledge in the target task:

likelihood of target data

$$\underbrace{p(\theta|D_T, D_S)}_{\text{posterior}} \propto p(D_T|\theta) p_S(\theta)$$

posterior

prior from source data

- Sequential data assimilation would dictate that the prior PDF is in fact the likelihood PDF obtained using the source data, i.e. $p_S(\theta) = p(D_S|\theta)$

✗ This approach does not provide flexibility in allowing the modeler to dictate how much knowledge, if any, is transferred:
  - In a traditional setting of data assimilation, all data, whether source or target, can be captured by the same model with the same parameter values (or PDFs). This assumption is not longer guaranteed to be valid in a transfer learning setting
- Need a mechanism to control how much knowledge, if any, is transferred from source task to target task

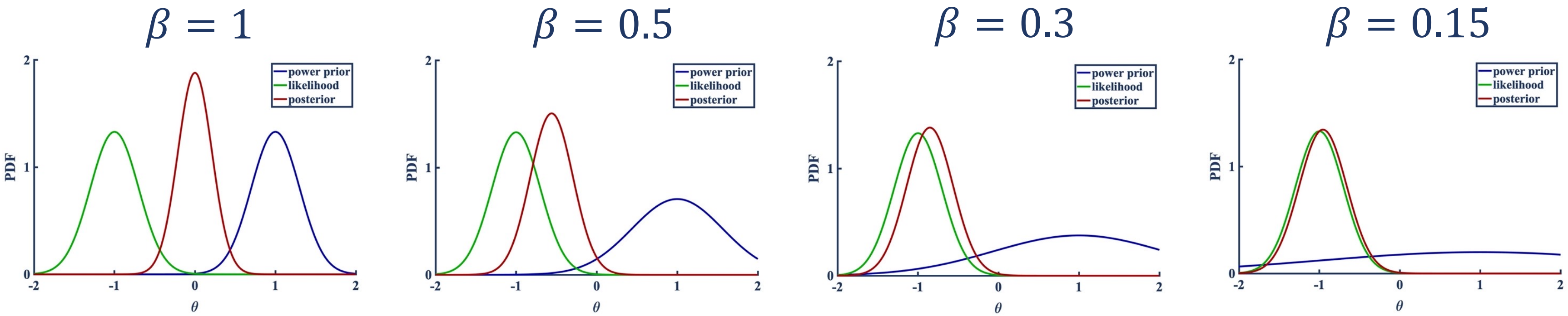
Tempering-based Diffusion of Knowledge

- How much knowledge to transfer: Tempering-based methodologies
- Tempering transformations allow us to “diffuse” or “concentrate” probabilistic knowledge (PDFs) gained through source domain learning tasks, effectively dictating how much knowledge is transferred to the target learning task
- Many PDF-tempering transformations that dictate how knowledge is transferred are envisaged
- Two proposed strategies consist of extensions/modifications of existing Bayesian priors:
  - $p_S(\theta \mid \beta) \propto p(D_S|\theta)^\beta$ Based on “power” priors
  - $p_S(\theta \mid \beta) = \beta\, p(D_S|\theta) + (1-\beta)\mathcal{N}(\theta; 0, \sigma^2 I)$ Based on “mixture” priors
- For the two types of transformations above
  - Full transfer: $\beta \to 1$ reverts back to the full likelihood from the source training task (i.e. traditional Bayes)
  - No transfer: $\beta \to 0$ results in a flat prior
  - Partial transfer: $0 < \beta < 1$
- Optimal choice of β depends on many factors, including:
  - ML model used (can capture local vs global trends)
  - Disparity between source and target domains
  - Degree of relative data sparsity (between source and target domains)
  - Relative intensity of noise in source and target domain data

Tempering-based Diffusion of Knowledge

- The following is an example of the extension of power-based and mixture-based prior tempering transformation to “diffuse” knowledge in the prior PDF
- The prior and likelihood PDFs are chosen to be Gaussian
- Note: one can show that for a Gaussian PDF, raising it to a power β is equivalent to scaling the associated covariance matrix by the same β (mean vector unaffected)
- Power prior: Gaussian posterior

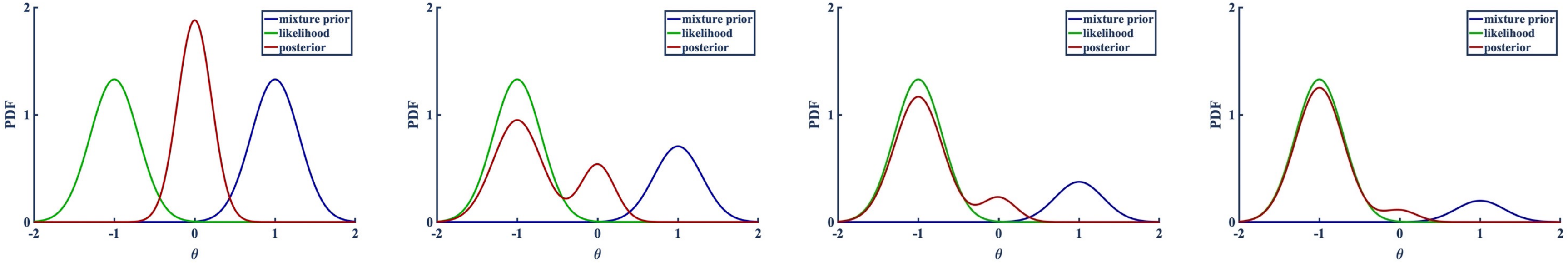
$$p(\theta|D,\beta) \propto p(D|\theta)p(\theta)^{\beta}$$



$\beta = 1$ $\beta = 0.5$ $\beta = 0.3$ $\beta = 0.15$

- Mixture prior: Gaussian-mixture posterior

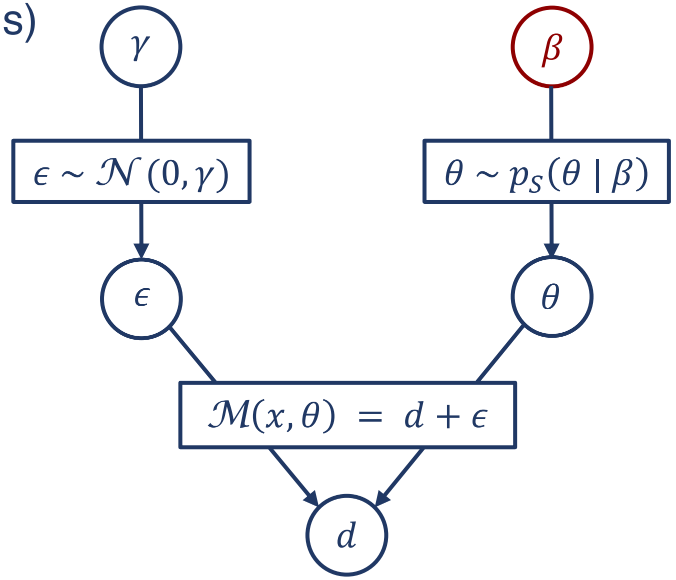
$$p(\theta|D,\beta) \propto p(D|\theta)[\beta\, p(\theta) + (1-\beta)\, \mathcal{N}(\theta; 0{,}1\times 10^{4})]$$



Degree of Knowledge/Learning to Transfer - Hierarchical Bayes

- The “tempering” hyper-parameter(s) β allow us to control the degree to which learning is transferred from the source task, characterized by the prior PDF, to the target task
- There are two approaches within a Bayesian context to determining β:
  1. Hierarchical Bayes
     - A fully Bayesian treatment of model parameters, θ, and noise and prior hyper-parameters, e.g. γ and β
     - Proceed with joint inference of all unknowns according to joint posterior:

$$
\begin{aligned}
p(\theta, \beta \mid D) &= p(\theta \mid \beta, D)\, p(\beta \mid D) && \text{(probability chain rule)} \\
&\propto p(D \mid \theta, \beta)\, p(\theta \mid \beta)\, p(\beta \mid D) && \text{(Bayes' rule)} \\
&= p(D \mid \theta)\, p(\theta \mid \beta)\, p(\beta) && \text{(independence assumptions)}
\end{aligned}
$$



- Posterior distribution over the ML model parameters, θ, can be obtained by marginalizing over the hyper-parameters
- ✓ Propagates uncertainty in hyper-parameters through to parameter posterior
- ✗ Added complexity associated with inference of “less relevant” parameters and propagation of uncertainty associated with it

Degree of Knowledge/Learning to Transfer - Empirical Bayes

- There are two approaches within a Bayesian context to determining β:
  2. Empirical Bayes
     - A pseudo-Bayesian treatment of prior hyper-parameter(s), β
     - ✓ Instead of inferring and subsequentially propagating uncertainties in the hyper-parameters, point estimates are obtained by maximizing *some* objective function
     - ✗ **What objective function?** Although empirical Bayes has been applied in numerous contexts for various purposes, there is not precedent for its use in transfer learning in determining such hyper-parameters
     - ✓ **Our solution:** Follow an ***information-theoretic*** approach, focusing on the **relative entropy** that measures the information geometry in moving from the prior to posterior:

$$\underbrace{\mathrm{D}_{\mathrm{KL}}[\,p(\theta \mid D,\beta), p(\theta \mid \beta)]}_{\substack{\text{Relative entropy}\\ \text{(prior \& posterior)}}} = -\underbrace{\mathrm{H}[\,p(\theta \mid D,\beta), p(D \mid \theta)]}_{\substack{\text{Cross entropy}\\ \text{(likelihood \& posterior)}}} = \underbrace{\int \log p(D \mid \theta)\, p(\theta \mid D,\beta)}_{\substack{\text{Expected data-fit}\\ \text{(posterior-averaged log-likelihood)}}}$$

     - This objective function (usually employed for Bayesian experimental design) has multiple interpretations, including an “average” log-likelihood one, resulting in a maximum “expected” likelihood estimate for hyper-parameter(s) β

Expected Data-fit to Dictate Degree of Knowledge/Learning to Transfer

- Current strategy: Empirical Bayesian treatment of the tempering hyper-parameter(s) with information-theoretic objective function in the form of the expected data-fit
- This is a data-driven approach: Optimal tempering hyper-parameters depend on the extent to which the source and target data provide "similar" information on the unknown model parameters:

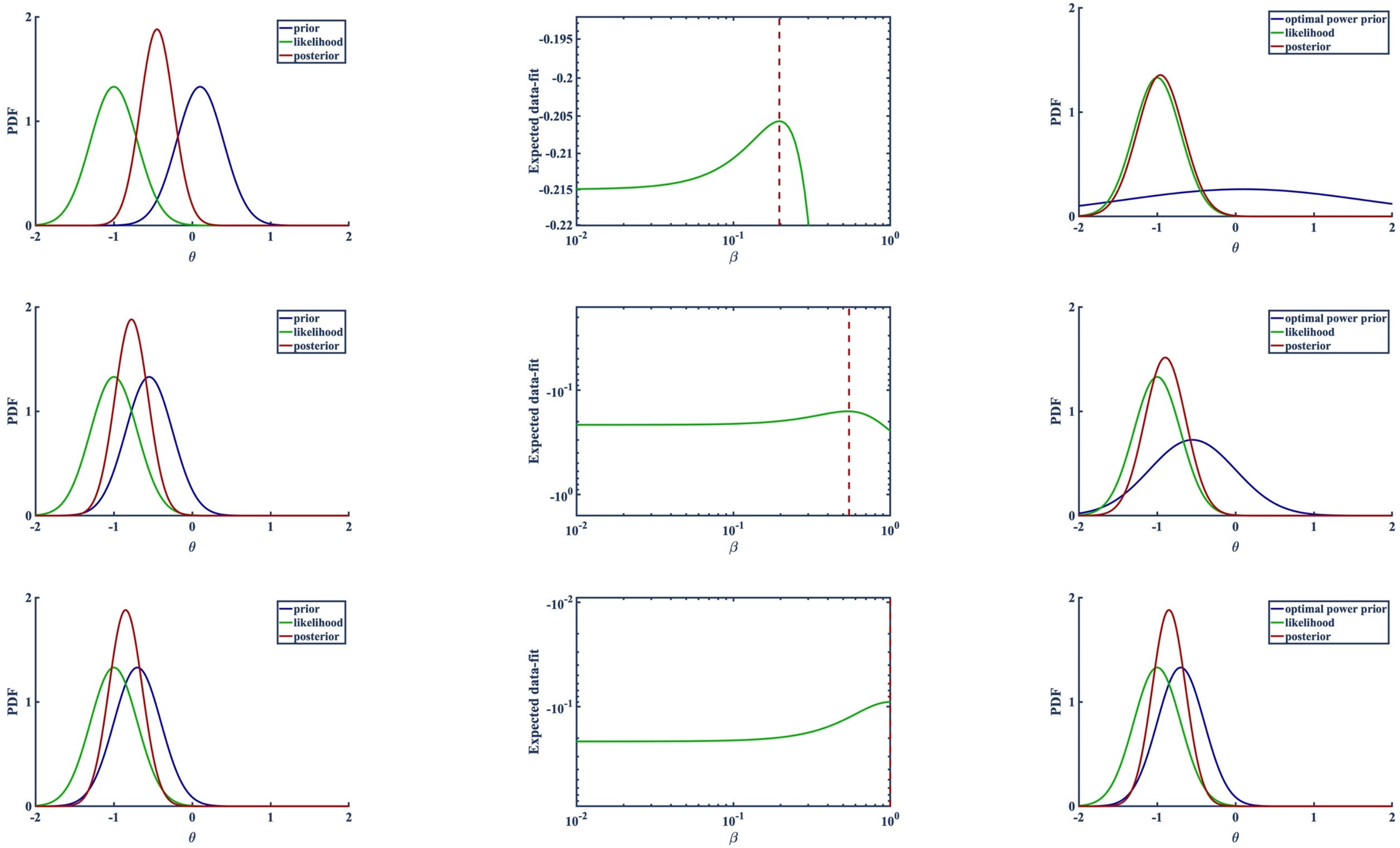
$$p(\theta|D,\beta) \propto p(D|\theta)p(\theta|\beta) \qquad \beta_{\text{opt}} = \max_{\beta} \mathrm{E}[\log p(D \mid \theta)] = \max_{\beta} \underbrace{\int \log p(D \mid \theta)\, p(\theta \mid D, \beta)\, \mathrm{d}\theta}_{\text{Integral over parameter-space!}}$$

- Desirable properties:
  - Utilizes whatever available sparse target data (no need to split target data into training and testing sets)
  - Full Bayesian treatment of parametric uncertainties
  - Generally applicable to any ML model and tempering transformation
  - For Gaussian distributions (linear-in-parameter ML models) and power prior tempering transformation, **closed form expression available for the objective function and gradient/Hessian**, thus facilitating optimization step

Expected Data-fit to Dictate Degree of Knowledge/Learning to Transfer

- Revisiting the 1-D example of dealing with Gaussian PDFs, let’s examine the result of utilizing the expected data-fit to drive the choice of tempering hyper-parameter(s)
- Note: for Gaussian PDFs and the use power-prior based tempering transformations, the expected data-fit is available analytically

$$p(\theta|D,\beta) \propto p(D|\theta)p(\theta)^{\beta} \qquad \beta_{\text{opt}} = \max_{\beta} E[\log p(D \mid \theta)] = \max_{\beta} \int \log p(D \mid \theta)\, p(\theta \mid \beta, D)\, \mathrm{d}\theta$$



Little transfer
$\beta_{\text{opt}} = 0.2$

Moderate transfer
$\beta_{\text{opt}} = 0.55$

Full transfer
$\beta_{\text{opt}} = 1$

Polynomial Surrogate Construction via Polynomial Chaos Expansions

- Polynomial chaos expansions (PCEs) are spectral expansions in the form

$$f(x) \approx \sum_{k=0}^{K-1} c_k \, \Psi(x)$$

with Ψ's being orthogonal polynomials with respect to some measure $p(x)$:

  - $p(x)$ is normal, $\Psi(x)$ are Hermite polynomials
  - $p(x)$ is uniform, $\Psi(x)$ are Legendre polynomials
- Truncation, i.e. choice of K, typically fixed based on chosen (total) order of expansion
- Some advantages over monomial basis:
  - Can use projection-based methods to identify c_k
  - Sensitivity indices (Sobol), output moments available analytically
- Coefficients c_k determined using:
  - Projection (integration): Monte Carlo, Quadrature, etc..
  - Regression: Least-squares, Bayesian

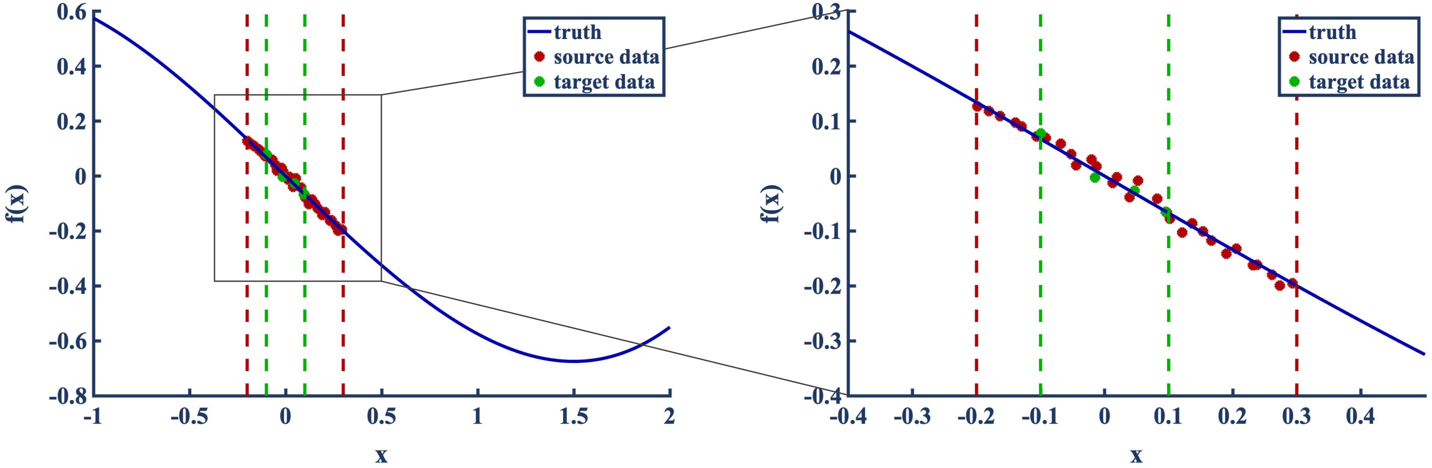
$$y_i = f(x_i) + \epsilon_i \qquad \epsilon_i \sim N(o, \gamma^2) \qquad p(y_1, \dots, y_N \,|\, \boldsymbol{c}) = \overbrace{\frac{1}{\left(\sqrt{2\,\pi}\,\gamma\right)^N} \prod_{i=1}^{N} e^{-\frac{1}{2\gamma^2}\left(y_i - \sum_{k=0}^{K-1} c_k \, \Psi(x_i)\right)^2}}^{\text{Gaussian (in parameter) Likelihood}}$$

Probabilistic Transfer Learning at Work – Polynomial Surrogates

- Assuming we're dealing with a "true" model given by

$$y_i = f(x) + \epsilon_i = 0.1x^3 - 0.75x + \epsilon_i \qquad\qquad \epsilon_i \sim \mathcal{N}(0, 1\times10^4)$$

- We have 30 data points from the source domain and 4 from the target domain

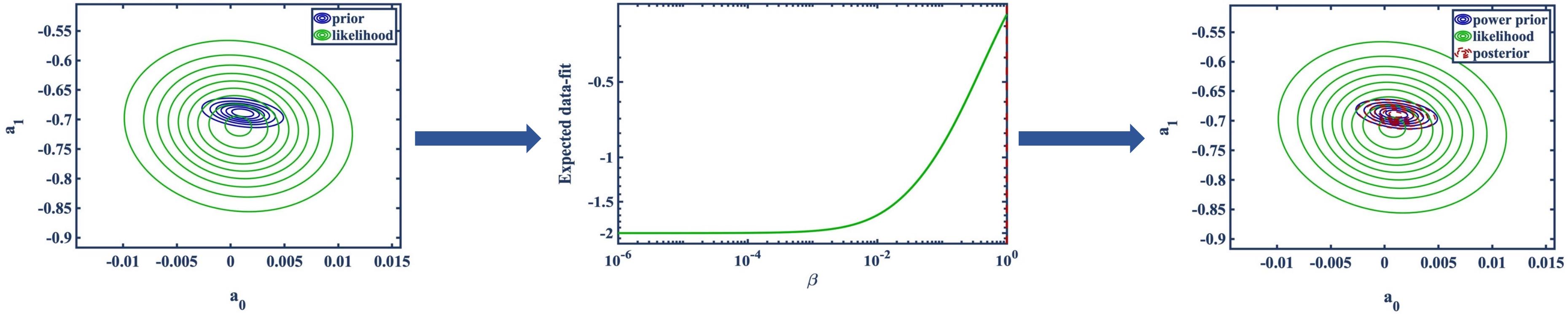


- Transfer learning task: Leverage the available source data to enhance accuracy of predictive model for target task, trained using the scarce target data
- We start with an approximate ML model to train. Let's assume a linear model:

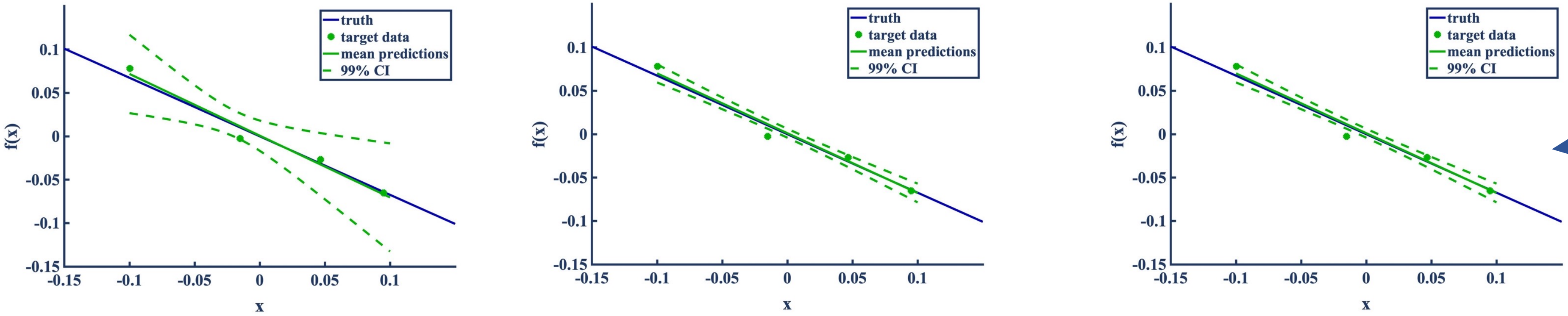
$$y_i \approx a_0 + a_1 x + \epsilon_i$$

Probabilistic Transfer Learning at Work – Polynomial Surrogates

- The source data, once assimilated, provide the prior PDF for subsequent use in the target training task
- Similarly, the target data provide the likelihood PDF
- We maximize the expected data-fit to arrive at an optimal power prior



- The following are posterior predictive mean estimates and confidence intervals



no transfer

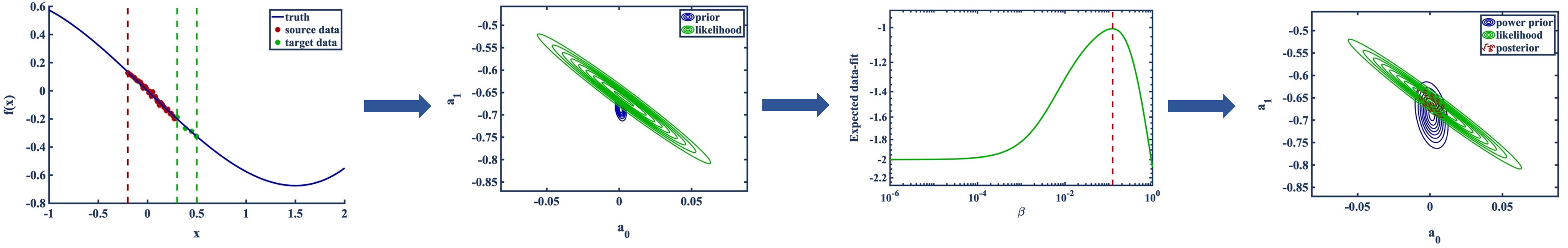
full transfer

optimal transfer, $\beta_{\text{opt}} = 1$

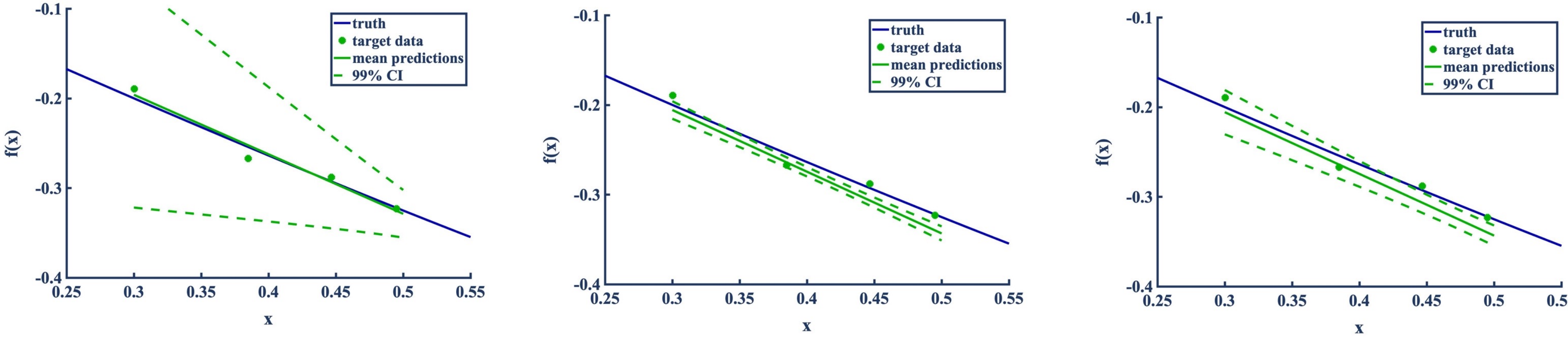
- Observation: overlapping source and target domains result in full transfer of learning

Probabilistic Transfer Learning at Work – Polynomial Surrogates

- Let’s repeat the procedure with a target domain that’s adjacent to the source domain (extrapolation)
- Again, we maximize the expected data-fit to arrive at an optimal power prior



- The following are posterior predictive mean estimates and confidence intervals



no transfer

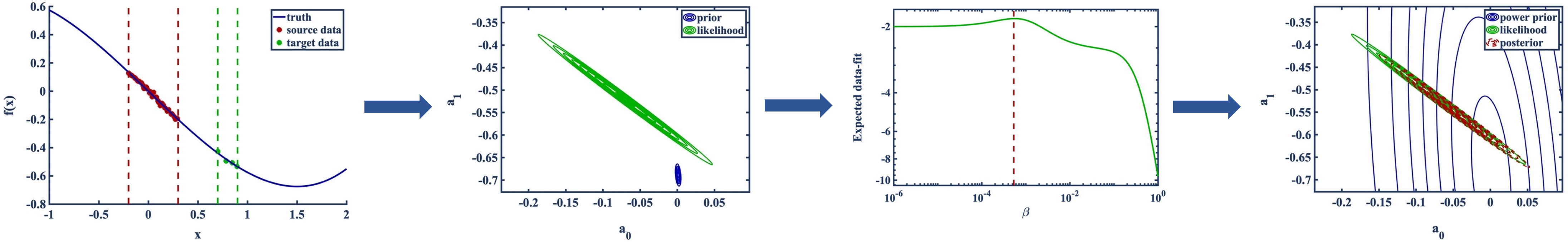
full transfer

optimal transfer, $\beta_{\text{opt}} \approx 0.1$

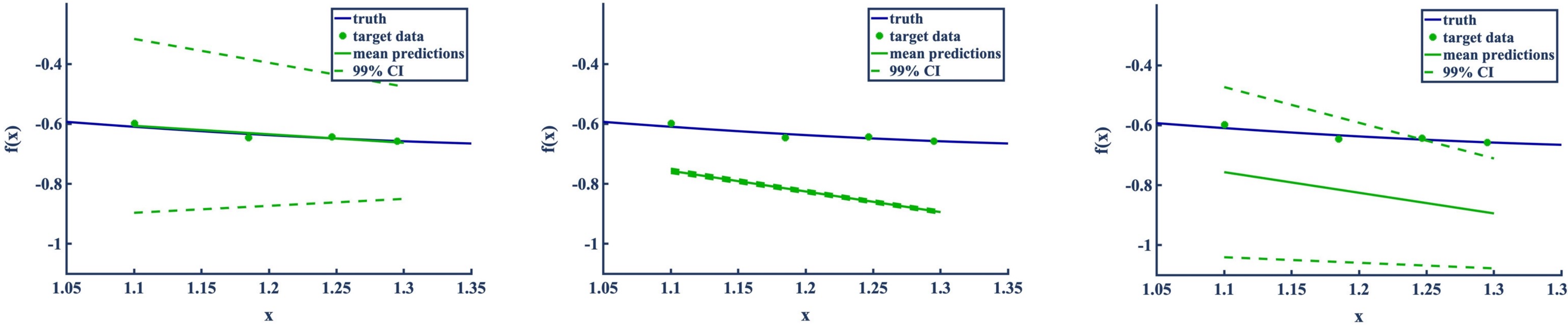
- Observation: Optimal transfer results in more accurate/precise predictions

Probabilistic Transfer Learning at Work – Polynomial Surrogates

- Lastly, we examine dissimilar target and source domains
- Again, we maximize the expected data-fit to arrive at an optimal power prior



- The following are posterior predictive mean estimates and confidence intervals



no transfer

full transfer

optimal transfer, $\beta_{\text{opt}} \approx 0.0005$

- Observation: Negligible transfer takes place with dissimilar tasks

Summary

- Developing a novel probabilistic framework for transfer learning:
    1. Reducing the validation/testing errors of such models by leveraging data from similar domains
    2. Propagating parametric, model-form, and data uncertainties towards predictions
    3. Allowing for optimal ML model selection within the TL paradigm
    4. Exhibiting moderate/strong computational scalability with increasing data volume and model complexity
    5. Safeguarding against negative learning (decreased accuracy due to task disparity, w.r.t. baseline)
    6. Consisting of strictly non-intrusive methods, applicable to most ML models without needing to modify the model (architecture) or implementation.
- Key technical steps:
    1. Capture the knowledge to be transferred in training on source data
        - Probability density functions on the calibrated ML model parameters/hyperparameters using Bayesian inversion
        - Captures data and modeling errors/uncertainties
    2. Propagate the knowledge to be transferred to target training tasks
        - Novel mechanisms for knowledge transfer that extends the traditional Bayesian approach via the application of prior PDF tempering transformations
    3. Determine how much knowledge to transfer given a choice of tempering transformation
        - Explore hierarchical or empirical Bayes approaches, based on information-theoretic measures and distance metrics
    4. Determine optimal ML model to use in TL
        - Determine the optimal ML model for TL using Bayesian model selection