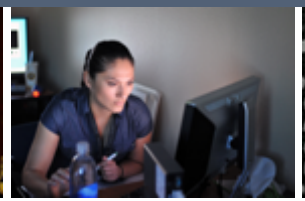
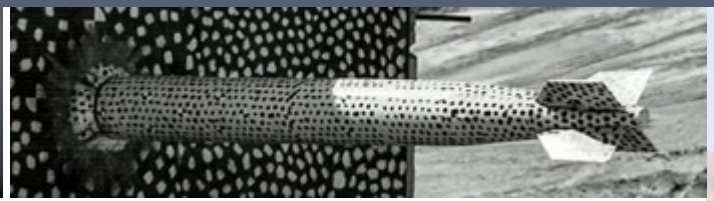
SAND2022-9310C

Sandia National Laboratories

Probabilistic Approaches to Transfer Learning





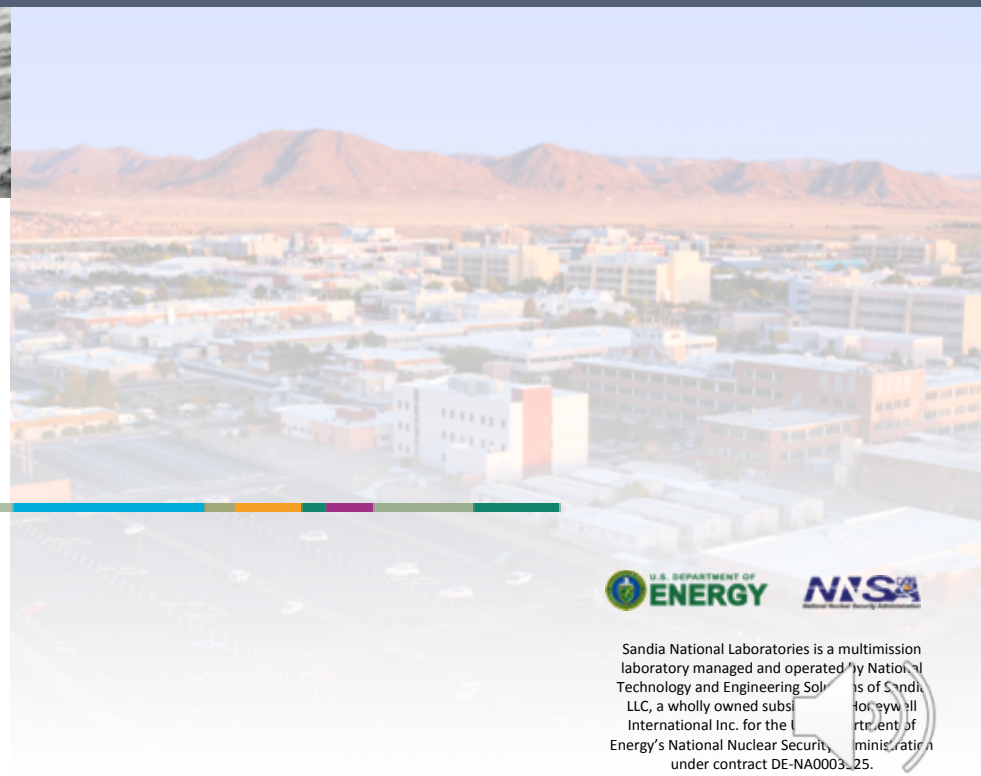
6th annual Sandia Machine Learning and Deep Learning Workshop

7/22/2022

PRESENTED BY

Wyatt Bridgman (SNL)

Collaborators:
Moe Khalil (PI), Fulton Wang, Justin Jacobs, Ahmad Rushdi, Reese Jones, Jackie Chen, Martin Rieth, Yuki Shimizu & Bruno Soriano (SNL), Tarek Echekki (NCSU), Chris Pettit (USNA), Minh Do & Molly Dasso (UIUC)



U.S. DEPARTMENT OF ENERGY NNSA

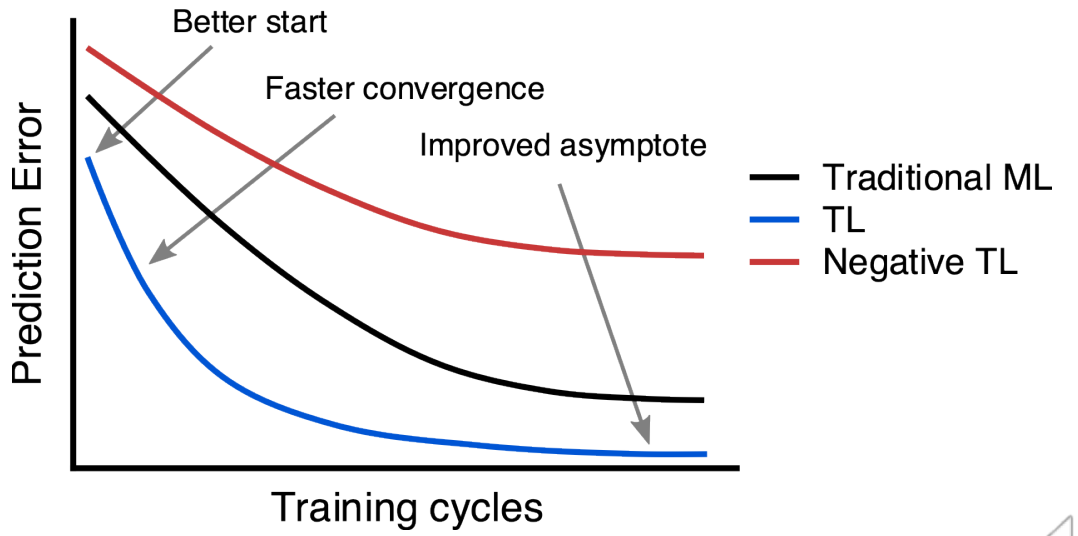
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

SAND NO: TBD

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.

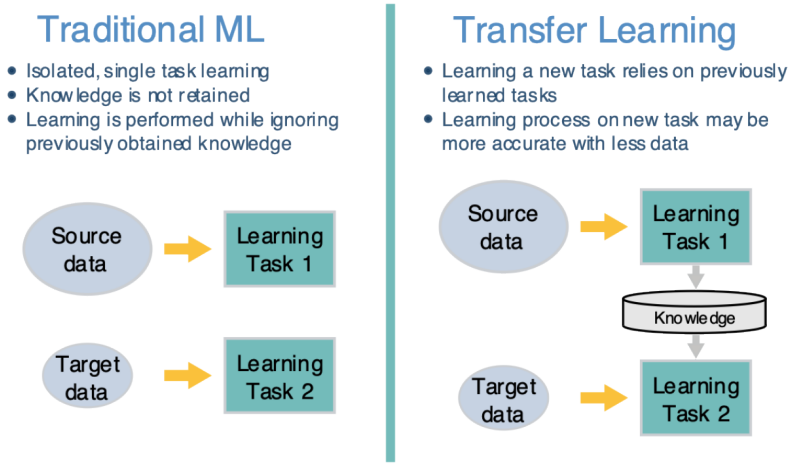
Alleviating Ill-Posedness in Training of ML Models with Sparse Data

- **Challenge:** Many Sandia mission domains are defined by a lack of reliable data, effectively precluding the use of many modern deep learning/machine learning techniques for predictive modeling:
  - ✘ Excessive expense of computer simulations
  - ✘ Prohibitive experimental data acquisition cost
  - ✘ Limited access to classified ad/or sensitive data
- **Goal:** *Enhance the trust in machine learning* (ML) model predictions within *noisy and sparse data settings*
- **Proposed Solution:** Novel probabilistic transfer learning framework.
- Transfer learning (TL): knowledge gained through similar training tasks is used to possibly improve the training process on a target domain having limited/noisy data:
  - ✓ *Improved initialization*
  - ✓ *Increased rate of convergence*
  - ✓ *Greater achievable performance*
- Proposed framework will aim to alleviate potential negative transfer: TL resulting in decreased Performance



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



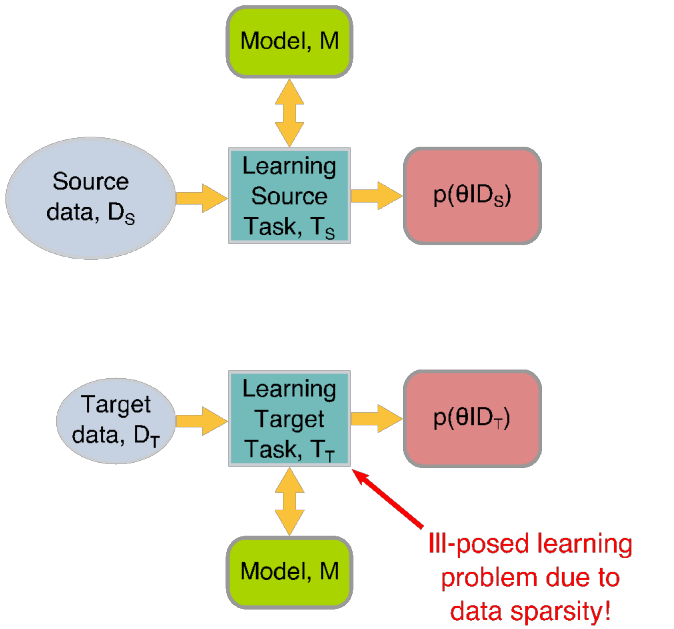
Transfer Learning Approaches	Brief Description
Instance-transfer	To re-weight some labeled data in the source domain for use in the target domain
Feature-representation-transfer	Find a "good" feature representation that reduces difference between the source and the target domains and the error of classification 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

Technical Approach

- Proposed TL framework aims to address the shortcomings in existing methodologies.
- It determines when to apply TL, which model to use, and how much knowledge to transfer.
- It relies on probability/measure theories to characterize and propagate uncertainties, thereby enhancing the trustworthiness of ML models in making predictions based on noisy and sparse training data.

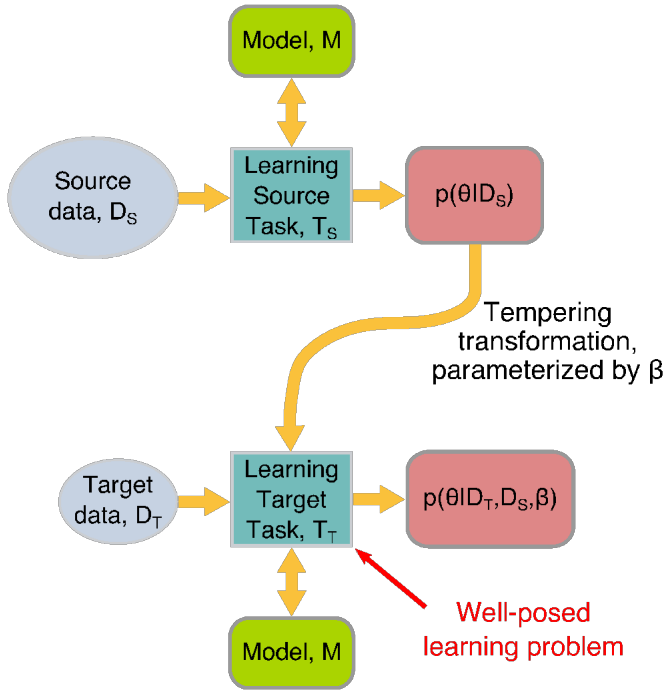
Traditional ML

- Isolated, single task learning
- Knowledge is not retained
- Learning is performed while ignoring previously obtained knowledge



Proposed TL

- Learnings a new task relies on previously learned tasks
- Learning process on new task may be more accurate with sparse 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:

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

posterior, likelihood, prior, 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

posterior $p(\theta|D_T, D_S) \propto p(D_T|\theta)p_S(\theta)$ 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

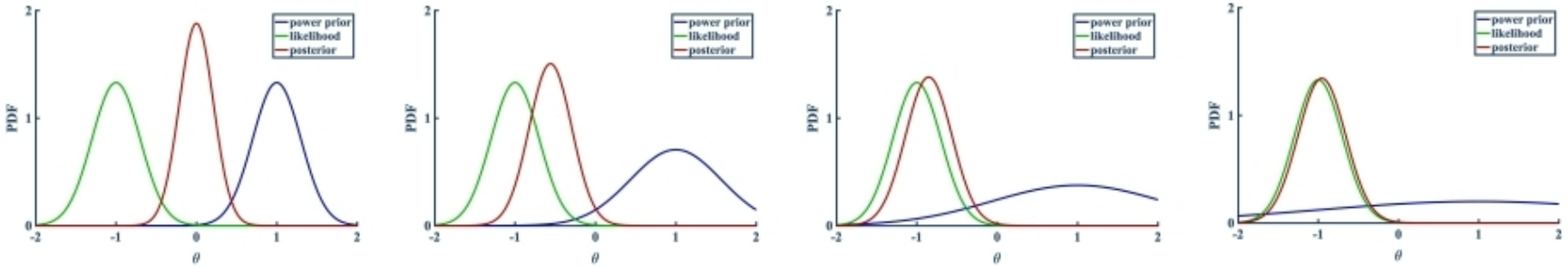
- How much knowledge to transfer: Tempering-based methodologies
- Tempering transformations allow us to "diffuse" or "concentrate" probabilistic knowledge (PDFs) gained through source domain learning tasks, effectively dictating how much knowledge is transferred to the target learning task
- Many PDF-tempering 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

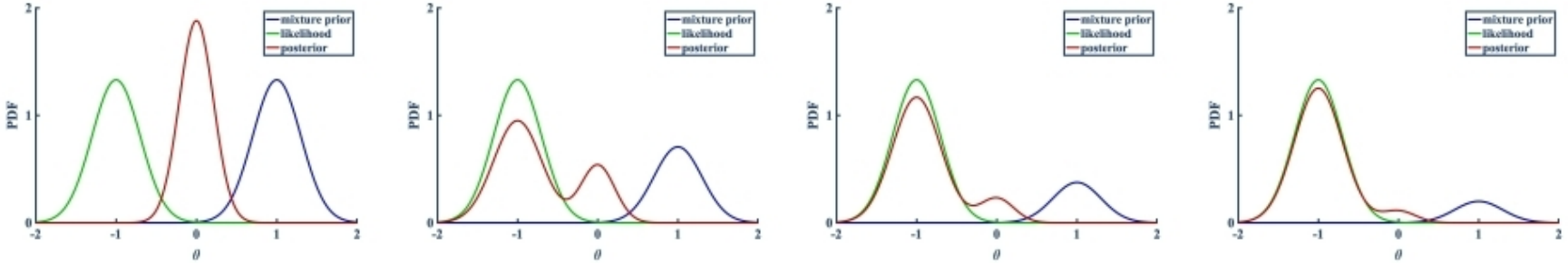
$$p(\theta|D,\beta) \propto p(D|\theta)p(\theta)^{\beta}$$

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



- Mixture prior: Gaussian-mixture posterior

$$p(\theta|D,\beta) \propto p(D|\theta)[\beta\, p(\theta) + (1-\beta)\, \mathcal{N}(\theta; 0{,}1\times10^{4})]$$



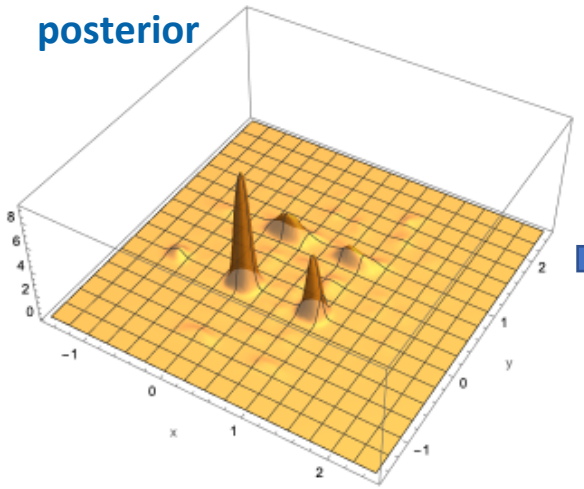
Tempering intractable posteriors - high fidelity PDF approximations

- Bayesian inverse problems for nonlinear models can yield intractable posteriors with multimodal behavior.
- **Problem:** multimodal, intractable source posteriors require high-fidelity, closed-form approximations in order to apply tempering transformations
- **Strategy:** combine global optimization with Laplace approximations to efficiently obtain a Gaussian mixture model (GMM) approximation source posterior/target prior:

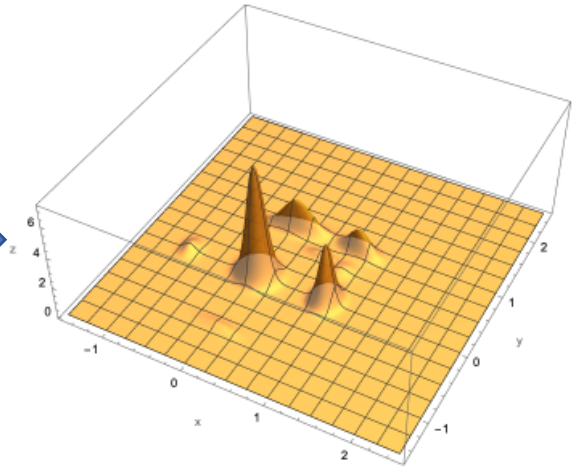
$$p_S(\theta|D_S) \approx q(\theta) = \sum_k \pi_k \mathcal{N}(\theta|\mu_k, \Sigma_k), \qquad \Sigma_k = \mathrm{H}_{-\log p_S}(\mu_k)^{-1}$$

- Can be used as an initialization strategy for obtaining efficient Variational Inference (VI) approximations.

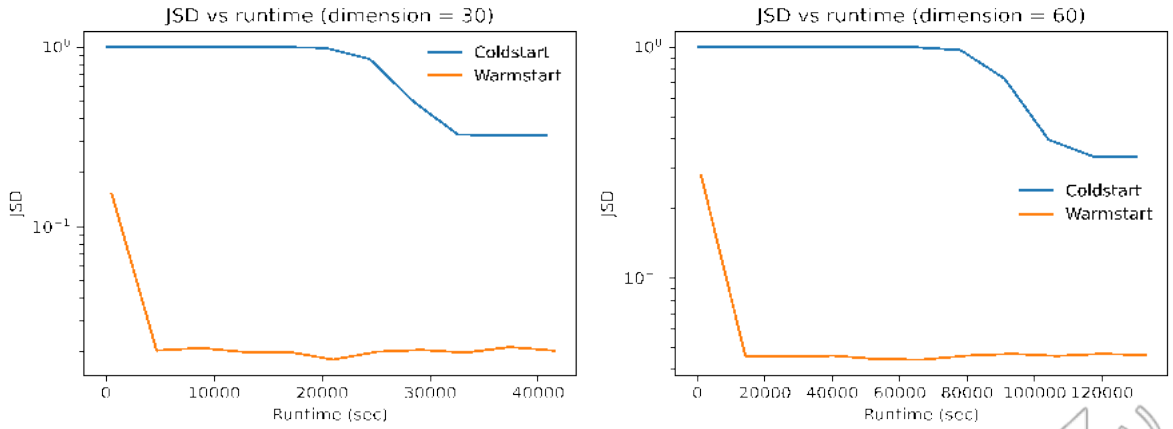
Multimodal, intractable posterior



High-fidelity GMM approximation



Scalability of VI initialized by GMM approx.

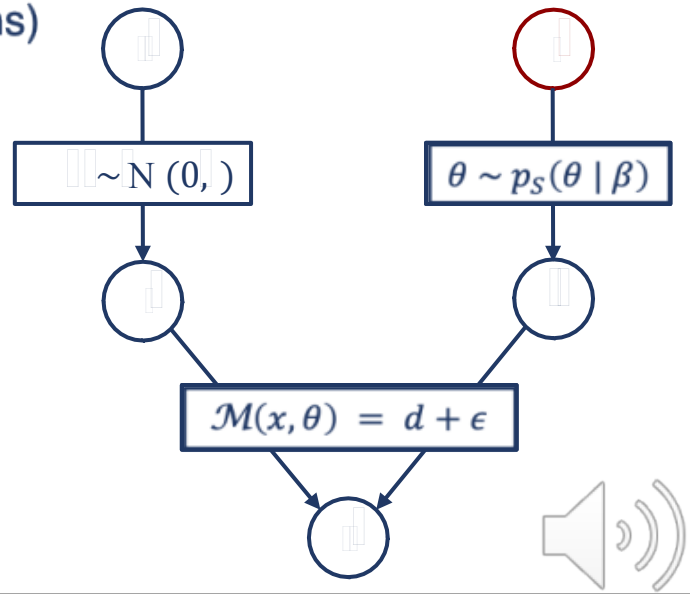


Degree of Knowledge/Learning to Transfer - Hierarchical Bayes

- 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
     - ✗ 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
     - ✓ Idea: for the objective function, we will follow an *information-theoretic* approach that relates to the Bayesian model evidence, oftentimes used in data-informed model selection:

$$\underbrace{\log p(D \mid \beta)}_{\text{log-evidence}} = \underbrace{E[\log p(D \mid \theta)]}_{\text{expected data-fit}} - \underbrace{E\left[\log \frac{p(\theta \mid D, \beta)}{p(\theta \mid \beta)}\right]}_{\text{expected information gain}}$$

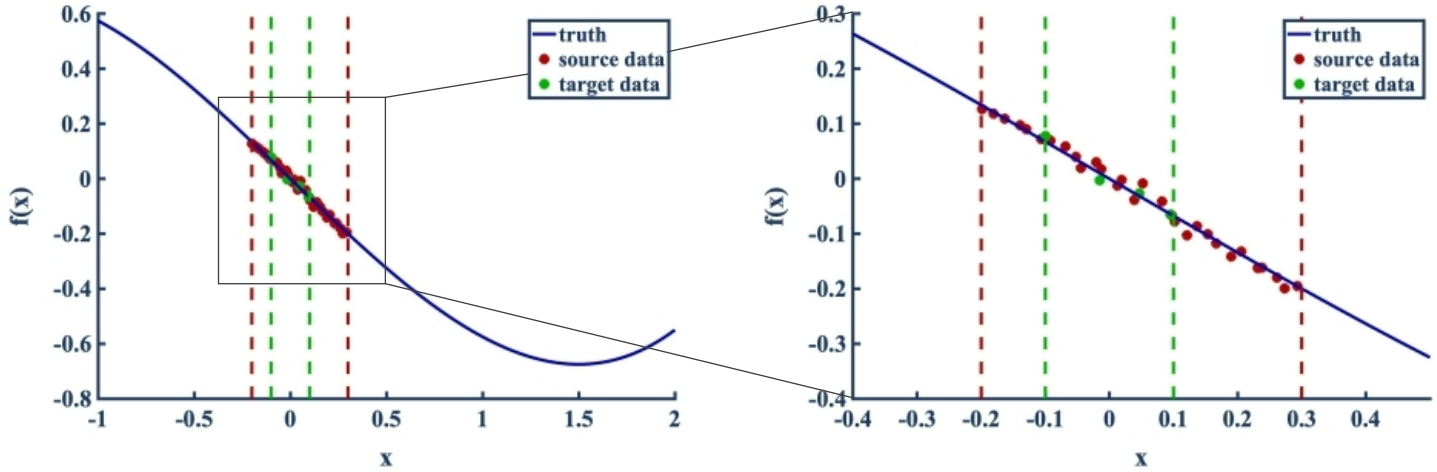
- In the above, the expectations are with respect to the parameter posterior $p(\theta \mid D, \beta)$
- The (log) evidence is comprised of a data-fit term and a term which provides a penalty against more "complex" models, the expected information gain: This has the tendency to drive β to zero in many settings (akin to behavior seen in automatic relevance determination in relevance vector machines)
- Result: use the expected data-fit as the objective function to maximize for the optimal β

Probabilistic Transfer Learning at Work – Polynomial Surrogates

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

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

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

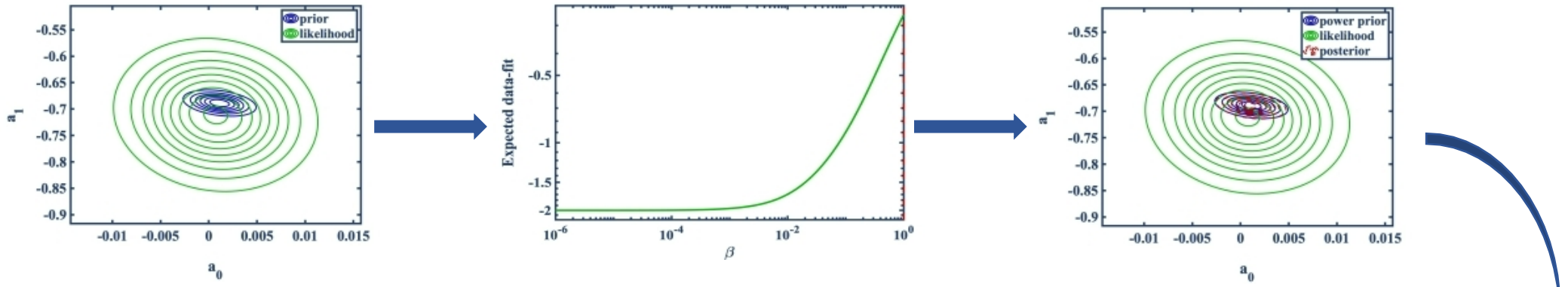


- Transfer learning task: Leverage the available source data to enhance accuracy of predictive model for target task, trained using the scarce target data
- We start with an approximate ML model to train. Let's assume a linear model:

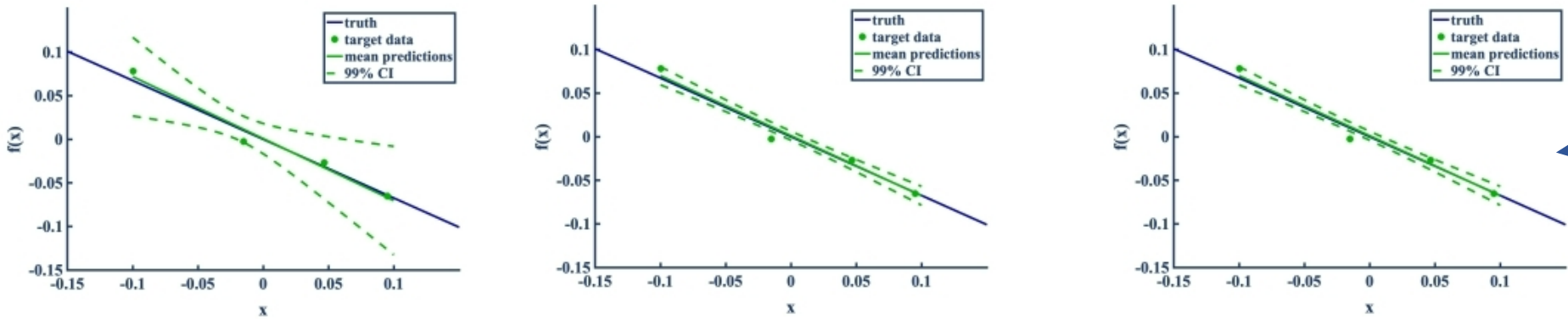
$$y_i \approx a_0 + a_1 x + \epsilon_i$$

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

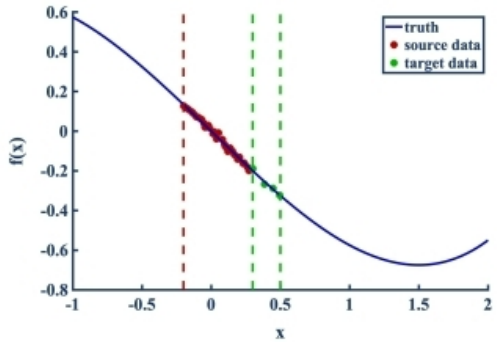


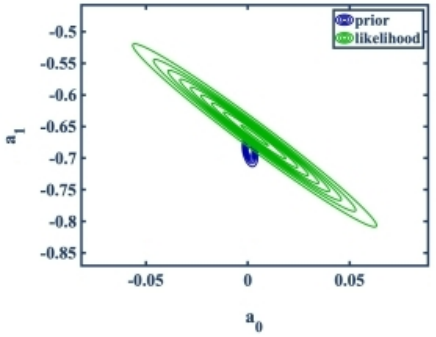
no transfer | full transfer | optimal transfer, $\beta_{\text{opt}} = 1$

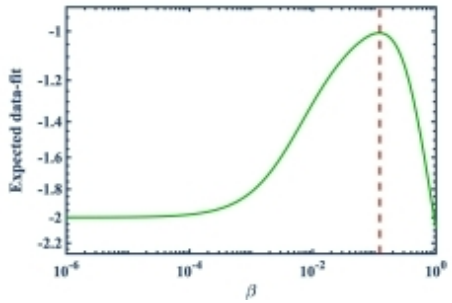
- Observation: overlapping source and target domains result in full transfer of learning

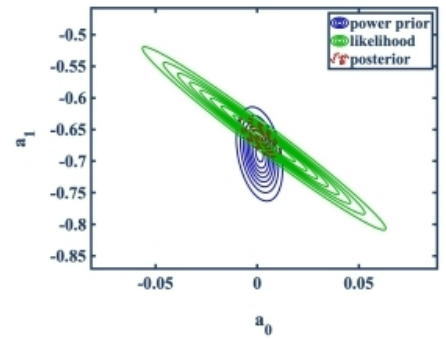
Probabilistic Transfer Learning at Work – Polynomial Surrogates

- 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

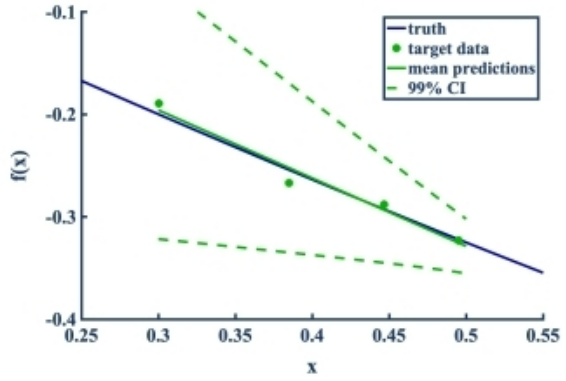




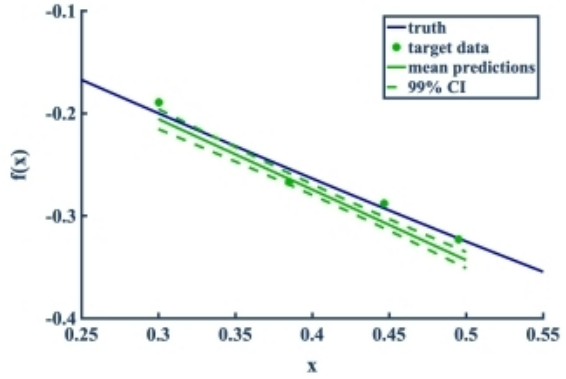




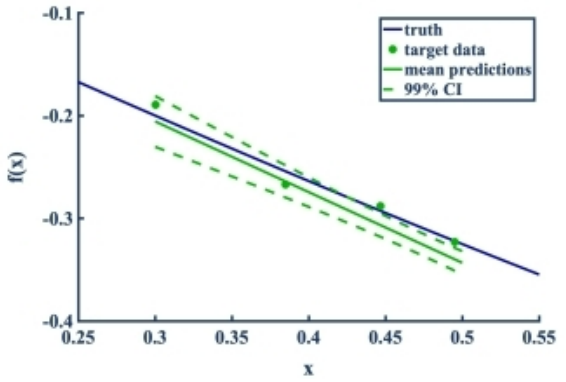
- The following are posterior predictive mean estimates and confidence intervals



no transfer



full transfer

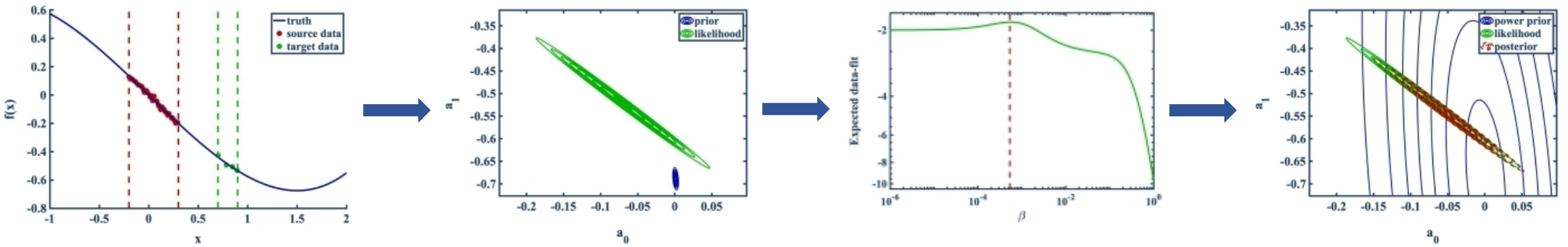


optimal transfer, $\beta_{\text{opt}} \approx 0.1$

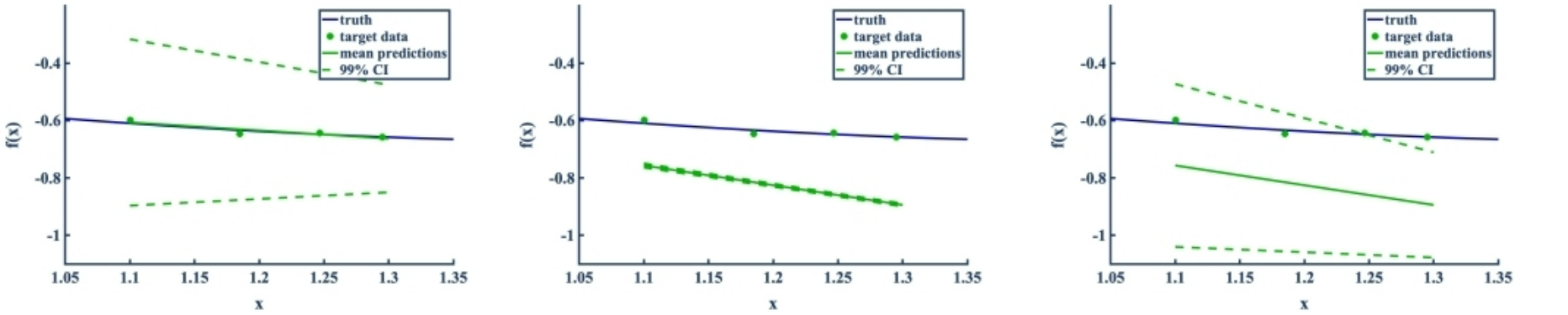
- Observation: Optimal transfer results in more accurate/precise predictions

Probabilistic Transfer Learning at Work – Polynomial Surrogates DELETE

- 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

- **Goals for probabilistic transfer learning**
  1. **Model performance:** improve model performance by leveraging data from similar domains
  2. **UQ:** propagating parametric, model-form, and data uncertainties towards predictions
  3. **Model selection:** allowing for optimal ML model selection within the TL paradigm
  4. **Scalability:** exhibiting moderate/strong computational scalability with increasing data volume and model complexity
  5. **Optimal transfer:** safeguarding against negative learning.
- **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
     - Efficient, high-fidelity approximations of parameter PDFs using Gaussian Mixture Models
  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

Thank you!