Bayesian Probabilistic Numerical Integration with Tree-Based Models

Xing Liu

Imperial College London

MCM 2021

November 8, 2021

This talk is based on

- H. Zhu, **X. Liu**, R. Kang, Z. Shen, S. Flaxman, F.-X. Briol (2020). *Bayesian probabilistic numerical integration with tree-based models*. NeurIPS 2020.

Overview of Today's Talk

1. Bayesian Probabilistic Numerical Integration (BPNI)
2. Bayesian Quadrature
3. Bayesian Additive Regression Trees (BART) and BART Integration
4. Experiments
5. Summary

Numerical Integration

Numerical integration concerns the estimation of an *intractable integral*

$$\Pi[f] := \int_{\mathcal{X}} f(x) d\Pi(x) = \int_{\mathcal{X}} f(x)\pi(x)dx,$$

where $f : \mathcal{X} \to \mathbb{R}$ ($\mathcal{X} \subset \mathbb{R}^d$) is assumed to be square-integrable w.r.t. a distribution Π on $\mathcal{X}$ that attains a density π.

- **Examples**: Posterior expectations, EM algorithm, differential equations.
- **Methods**: Monte Carlo integration (MI), MCMC, SMC, QMC...
- They are all **quadrature rules**:

$$\hat{\Pi}[f] = \sum_{i=1}^{n} w_i f(x_i),$$

  for some design points $\{x_i\}_{i=1}^n \subset \mathcal{X}$ and weights $\{w_i\}_{i=1}^n$.
- **Problem**: Not straightforward to quantify uncertainties about $\Pi[f]$ given only a *small* number of function evaluations!

Numerical Integration

Numerical integration concerns the estimation of an *intractable integral*

$$\Pi[f] := \int_{\mathcal{X}} f(x) d\Pi(x) = \int_{\mathcal{X}} f(x)\pi(x)dx,$$

where $f : \mathcal{X} \to \mathbb{R}$ ($\mathcal{X} \subset \mathbb{R}^d$) is assumed to be square-integrable w.r.t. a distribution Π on $\mathcal{X}$ that attains a density π.

- **Examples**: Posterior expectations, EM algorithm, differential equations.
- **Methods**: Monte Carlo integration (MI), MCMC, SMC, QMC...
- They are all **quadrature rules**:

$$\hat{\Pi}[f] = \sum_{i=1}^{n} w_i f(x_i),$$

  for some design points $\{x_i\}_{i=1}^n \subset \mathcal{X}$ and weights $\{w_i\}_{i=1}^n$.
- **Problem**: Not straightforward to quantify uncertainties about $\Pi[f]$ given only a *small* number of function evaluations!

Numerical Integration

Numerical integration concerns the estimation of an *intractable integral*

$$\Pi[f] := \int_{\mathcal{X}} f(x) d\Pi(x) = \int_{\mathcal{X}} f(x)\pi(x) dx,$$

where $f : \mathcal{X} \to \mathbb{R}$ ($\mathcal{X} \subset \mathbb{R}^d$) is assumed to be square-integrable w.r.t. a distribution Π on $\mathcal{X}$ that attains a density π.

- **Examples**: Posterior expectations, EM algorithm, differential equations.
- **Methods**: Monte Carlo integration (MI), MCMC, SMC, QMC...
- They are all **quadrature rules**:

$$\hat{\Pi}[f] = \sum_{i=1}^{n} w_i f(x_i),$$

  for some design points $\{x_i\}_{i=1}^n \subset \mathcal{X}$ and weights $\{w_i\}_{i=1}^n$.
- **Problem**: Not straightforward to quantify uncertainties about $\Pi[f]$ given only a *small* number of function evaluations!

Numerical Integration

Numerical integration concerns the estimation of an *intractable integral*

$$\Pi[f] := \int_{\mathcal{X}} f(x) d\Pi(x) = \int_{\mathcal{X}} f(x)\pi(x) dx,$$

where $f : \mathcal{X} \to \mathbb{R}$ ($\mathcal{X} \subset \mathbb{R}^d$) is assumed to be square-integrable w.r.t. a distribution Π on $\mathcal{X}$ that attains a density π.

- **Examples**: Posterior expectations, EM algorithm, differential equations.
- **Methods**: Monte Carlo integration (MI), MCMC, SMC, QMC...
- They are all **quadrature rules**:

$$\hat{\Pi}[f] = \sum_{i=1}^{n} w_i f(x_i),$$

  for some design points $\{x_i\}_{i=1}^n \subset \mathcal{X}$ and weights $\{w_i\}_{i=1}^n$.
- **Problem**: Not straightforward to quantify uncertainties about $\Pi[f]$ given only a *small* number of function evaluations!

Numerical Integration

Numerical integration concerns the estimation of an *intractable integral*

$$\Pi[f] := \int_{\mathcal{X}} f(x)d\Pi(x) = \int_{\mathcal{X}} f(x)\pi(x)dx,$$

where $f : \mathcal{X} \to \mathbb{R}$ ($\mathcal{X} \subset \mathbb{R}^d$) is assumed to be square-integrable w.r.t. a distribution Π on $\mathcal{X}$ that attains a density π.

- **Examples**: Posterior expectations, EM algorithm, differential equations.
- **Methods**: Monte Carlo integration (MI), MCMC, SMC, QMC...
- They are all **quadrature rules**:

  $$\hat{\Pi}[f] = \sum_{i=1}^{n} w_i f(x_i),$$

  for some design points $\{x_i\}_{i=1}^n \subset \mathcal{X}$ and weights $\{w_i\}_{i=1}^n$.
- **Problem**: Not straightforward to quantify uncertainties about $\Pi[f]$ given only a *small* number of function evaluations!

Bayesian Quadrature and BPNI

Bayesian quadrature (BQ): frame the problem as a statistical estimation task, so that probabilistic statements can be used to quantify uncertainty about $\Pi[f]$ for finite n.

- carries a Bayesian interpretation
- takes the form of a quadrature rule

Bayesian Probabilistic Numerical Integration (BPNI): any Bayesian estimators that can be used to estimate an intractable integral (not necessarily a quadrature rule).

Recipe of BQ:

1. Posit a GP prior distribution for the integrand f.
2. Compute the posterior distribution given values of f at some design points.
3. Push the distribution forward through $\Pi[\cdot]$ to get the implied distribution on $\Pi[f]$.

Bayesian Quadrature and BPNI

Bayesian quadrature (BQ): frame the problem as a statistical estimation task, so that probabilistic statements can be used to quantify uncertainty about $\Pi[f]$ for finite n.

- carries a Bayesian interpretation
- takes the form of a quadrature rule

Bayesian Probabilistic Numerical Integration (BPNI): any Bayesian estimators that can be used to estimate an intractable integral (not necessarily a quadrature rule).

Recipe of BQ:

1. Posit a GP prior distribution for the integrand f.
2. Compute the posterior distribution given values of f at some design points.
3. Push the distribution forward through $\Pi[\cdot]$ to get the implied distribution on $\Pi[f]$.

Bayesian Quadrature and BPNI

Bayesian quadrature (BQ): frame the problem as a statistical estimation task, so that probabilistic statements can be used to quantify uncertainty about $\Pi[f]$ for finite n.

- carries a Bayesian interpretation
- takes the form of a quadrature rule

Bayesian Probabilistic Numerical Integration (BPNI): any Bayesian estimators that can be used to estimate an intractable integral (not necessarily a quadrature rule).

Recipe of BQ:

1. Posit a GP prior distribution for the integrand f.
2. Compute the posterior distribution given values of f at some design points.
3. Push the distribution forward through $\Pi[\cdot]$ to get the implied distribution on $\Pi[f]$.

Bayesian Quadrature and BPNI

Bayesian quadrature (BQ): frame the problem as a statistical estimation task, so that probabilistic statements can be used to quantify uncertainty about $\Pi[f]$ for finite n.

- carries a Bayesian interpretation
- takes the form of a quadrature rule

Bayesian Probabilistic Numerical Integration (BPNI): any Bayesian estimators that can be used to estimate an intractable integral (not necessarily a quadrature rule).

Recipe of BQ:

1. Posit a GP prior distribution for the integrand f.
2. Compute the posterior distribution given values of f at some design points.
3. Push the distribution forward through $\Pi[\cdot]$ to get the implied distribution on $\Pi[f]$.

Bayesian Quadrature and BPNI

Bayesian quadrature (BQ): frame the problem as a statistical estimation task, so that probabilistic statements can be used to quantify uncertainty about $\Pi[f]$ for finite n.

- carries a Bayesian interpretation
- takes the form of a quadrature rule

Bayesian Probabilistic Numerical Integration (BPNI): any Bayesian estimators that can be used to estimate an intractable integral (not necessarily a quadrature rule).

Recipe of BQ:

1. Posit a GP prior distribution for the integrand f.
2. Compute the posterior distribution given values of f at some design points.
3. Push the distribution forward through $\Pi[\cdot]$ to get the implied distribution on $\Pi[f]$.

Bayesian Quadrature and BPNI

Bayesian quadrature (BQ): frame the problem as a statistical estimation task, so that probabilistic statements can be used to quantify uncertainty about $\Pi[f]$ for finite n.

- carries a Bayesian interpretation
- takes the form of a quadrature rule

Bayesian Probabilistic Numerical Integration (BPNI): any Bayesian estimators that can be used to estimate an intractable integral (not necessarily a quadrature rule).

Recipe of BQ:

1. Posit a GP prior distribution for the integrand f.
2. Compute the posterior distribution given values of f at some design points.
3. Push the distribution forward through $\Pi[\cdot]$ to get the implied distribution on $\Pi[f]$.

Bayesian Quadrature and BPNI

Bayesian quadrature (BQ): frame the problem as a statistical estimation task, so that probabilistic statements can be used to quantify uncertainty about $\Pi[f]$ for finite n.

- carries a Bayesian interpretation
- takes the form of a quadrature rule

Bayesian Probabilistic Numerical Integration (BPNI): any Bayesian estimators that can be used to estimate an intractable integral (not necessarily a quadrature rule).

Recipe of ~~BQ~~ BPNI:

1. Positing an ~~GP~~ arbitrary prior distribution for the integrand f
2. Computing the posterior distribution given values of f at some design points.
3. Pushing the distribution forward through $\Pi[\cdot]$ to get the implied distribution on $\Pi[f]$.

Bayesian Quadrature and BPNI

Bayesian quadrature (BQ): frame the problem as a statistical estimation task, so that probabilistic statements can be used to quantify uncertainty about $\Pi[f]$ for finite n.

- carries a Bayesian interpretation
- takes the form of a quadrature rule

Bayesian Probabilistic Numerical Integration (BPNI): any Bayesian estimators that can be used to estimate an intractable integral (not necessarily a quadrature rule).

Recipe of ~~BQ~~ ~~BPNI~~ BART Integration (BART-Int):

1. Positing an ~~GP~~ ~~arbitrary~~ BART prior distribution for the integrand f
2. Computing the posterior distribution given values of f at some design points.
3. Pushing the distribution forward through $\Pi[\cdot]$ to get the implied distribution on $\Pi[f]$.

Why not Stick to GPs?

Advantages:

- Posterior distribution for integrand f (and hence $\Pi[f]$) has a closed-form (assuming integrals of the form $\Pi[k(\cdot, x)]$ are available, where k is the covariance function of the GP).
- Different covariance functions k can be selected to accommodate integrands with different properties (smoothness, periodicity etc.).

Disadvantages

- **Discontinuities**: Hard to choose k when f is non-smooth or discontinuous.
- **Computational cost**: $\mathcal{O}(n^3)$. Prohibitive for large n.
- **High dimensions**: Applications of BQ are often limited to low-dimensional problems due to the curse of dimensionality, since the number of points needed will grow exponentially with d.

We have chosen a GP prior as the model for f, but this is not necessarily the only choice! We consider instead **tree-structured models**.

Why not Stick to GPs?

Advantages:

- Posterior distribution for integrand f (and hence $\Pi[f]$) has a closed-form (assuming integrals of the form $\Pi[k(\cdot, x)]$ are available, where k is the covariance function of the GP).
- Different covariance functions k can be selected to accommodate integrands with different properties (smoothness, periodicity etc.).

Disadvantages

- **Discontinuities**: Hard to choose k when f is non-smooth or discontinuous.
- **Computational cost**: $\mathcal{O}(n^3)$. Prohibitive for large n.
- **High dimensions**: Applications of BQ are often limited to low-dimensional problems due to the curse of dimensionality, since the number of points needed will grow exponentially with d.

We have chosen a GP prior as the model for f, but this is not necessarily the only choice! We consider instead **tree-structured models**.

Why not Stick to GPs?

Advantages:

- Posterior distribution for integrand f (and hence $\Pi[f]$) has a closed-form (assuming integrals of the form $\Pi[k(\cdot, x)]$ are available, where k is the covariance function of the GP).
- Different covariance functions k can be selected to accommodate integrands with different properties (smoothness, periodicity etc.).

Disadvantages

- **Discontinuities**: Hard to choose k when f is non-smooth or discontinuous.
- **Computational cost**: $\mathcal{O}(n^3)$. Prohibitive for large n.
- **High dimensions**: Applications of BQ are often limited to low-dimensional problems due to the curse of dimensionality, since the number of points needed will grow exponentially with d.

We have chosen a GP prior as the model for f, but this is not necessarily the only choice! We consider instead **tree-structured models**.

Why not Stick to GPs?

Advantages:

- Posterior distribution for integrand f (and hence $\Pi[f]$) has a closed-form (assuming integrals of the form $\Pi[k(\cdot, x)]$ are available, where k is the covariance function of the GP).
- Different covariance functions k can be selected to accommodate integrands with different properties (smoothness, periodicity etc.).

Disadvantages

- **Discontinuities**: Hard to choose k when f is non-smooth or discontinuous.
- **Computational cost**: $\mathcal{O}(n^3)$. Prohibitive for large n.
- **High dimensions**: Applications of BQ are often limited to low-dimensional problems due to the curse of dimensionality, since the number of points needed will grow exponentially with d.

We have chosen a GP prior as the model for f, but this is not necessarily the only choice! We consider instead **tree-structured models**.

Why not Stick to GPs?

Advantages:

- Posterior distribution for integrand f (and hence $\Pi[f]$) has a closed-form (assuming integrals of the form $\Pi[k(\cdot, x)]$ are available, where k is the covariance function of the GP).
- Different covariance functions k can be selected to accommodate integrands with different properties (smoothness, periodicity etc.).

Disadvantages

- **Discontinuities**: Hard to choose k when f is non-smooth or discontinuous.
- **Computational cost**: $\mathcal{O}(n^3)$. Prohibitive for large n.
- **High dimensions**: Applications of BQ are often limited to low-dimensional problems due to the curse of dimensionality, since the number of points needed will grow exponentially with d.

We have chosen a GP prior as the model for f, but this is not necessarily the only choice! We consider instead **tree-structured models**.

Why not Stick to GPs?

Advantages:

- Posterior distribution for integrand f (and hence $\Pi[f]$) has a closed-form (assuming integrals of the form $\Pi[k(\cdot, x)]$ are available, where k is the covariance function of the GP).
- Different covariance functions k can be selected to accommodate integrands with different properties (smoothness, periodicity etc.).

Disadvantages

- **Discontinuities**: Hard to choose k when f is non-smooth or discontinuous.
- **Computational cost**: $\mathcal{O}(n^3)$. Prohibitive for large n.
- **High dimensions**: Applications of BQ are often limited to low-dimensional problems due to the curse of dimensionality, since the number of points needed will grow exponentially with d.

We have chosen a GP prior as the model for f, but this is not necessarily the only choice! We consider instead **tree-structured models**.

Why not Stick to GPs?

Advantages:

- Posterior distribution for integrand f (and hence $\Pi[f]$) has a closed-form (assuming integrals of the form $\Pi[k(\cdot, x)]$ are available, where k is the covariance function of the GP).
- Different covariance functions k can be selected to accommodate integrands with different properties (smoothness, periodicity etc.).

Disadvantages

- **Discontinuities**: Hard to choose k when f is non-smooth or discontinuous.
- **Computational cost**: $\mathcal{O}(n^3)$. Prohibitive for large n.
- **High dimensions**: Applications of BQ are often limited to low-dimensional problems due to the curse of dimensionality, since the number of points needed will grow exponentially with d.

We have chosen a GP prior as the model for f, but this is not necessarily the only choice! We consider instead **tree-structured models**.

Why not Stick to GPs?

Advantages:

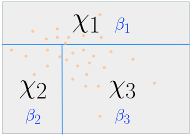
- Posterior distribution for integrand f (and hence $\Pi[f]$) has a closed-form (assuming integrals of the form $\Pi[k(\cdot, x)]$ are available, where k is the covariance function of the GP).
- Different covariance functions k can be selected to accommodate integrands with different properties (smoothness, periodicity etc.).

Disadvantages

- **Discontinuities**: Hard to choose k when f is non-smooth or discontinuous.
- **Computational cost**: $\mathcal{O}(n^3)$. Prohibitive for large n.
- **High dimensions**: Applications of BQ are often limited to low-dimensional problems due to the curse of dimensionality, since the number of points needed will grow exponentially with d.

We have chosen a GP prior as the model for f, but this is not necessarily the only choice! We consider instead **tree-structured models**.

Bayesian Additive Regression Trees (BART)



1. A **regression tree** is a step function:

$$g_{\mathcal{T},\beta}(x) = \sum_{k=1}^{K} \beta_k \mathbb{1}_{\chi_k}(x),$$

where $\beta := (\beta_1, \ldots, \beta_K)^\top \in \mathbb{R}^K$ are the **leaf values**, and $\chi_k \subset \mathcal{X}$ so that $\mathcal{T} := \{\chi_k\}_{k=1}^K$ forms a **partition** of $\mathcal{X}$.

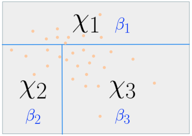
2. A T**-additive regression tree** is a sum of regression trees:

$$g_{\mathcal{E},\mathcal{B}}(x) := \sum_{t=1}^{T} g_{\mathcal{T}_t,\beta_t}(x),$$

where $\mathcal{B} := \{\beta_t\}_{t=1}^T$ and $\mathcal{E} := \{\mathcal{T}_t\}_{t=1}^T$.

3. A **Bayesian additive regression tree** (BART) is any distribution on the family of T-additive regression trees
   - This can be done by specifying a (prior) distribution on the leaf values $\mathcal{B}$ and partition $\mathcal{E}$ (Chipman et al. 1998, 2010).

Bayesian Additive Regression Trees (BART)



1. A **regression tree** is a step function:

$$g_{\mathcal{T},\beta}(x) = \sum_{k=1}^{K} \beta_k \mathbb{1}_{\chi_k}(x),$$

where $\beta := (\beta_1, \ldots, \beta_K)^\top \in \mathbb{R}^K$ are the **leaf values**, and $\chi_k \subset \mathcal{X}$ so that $\mathcal{T} := \{\chi_k\}_{k=1}^K$ forms a **partition** of $\mathcal{X}$.

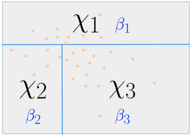
2. A T**-additive regression tree** is a sum of regression trees:

$$g_{\mathcal{E},\mathcal{B}}(x) := \sum_{t=1}^{T} g_{\mathcal{T}_t,\beta_t}(x),$$

where $\mathcal{B} := \{\beta_t\}_{t=1}^T$ and $\mathcal{E} := \{\mathcal{T}_t\}_{t=1}^T$.

3. A **Bayesian additive regression tree** (BART) is any distribution on the family of T-additive regression trees
   - This can be done by specifying a (prior) distribution on the leaf values $\mathcal{B}$ and partition $\mathcal{E}$ (Chipman et al. 1998, 2010).

Bayesian Additive Regression Trees (BART)



1. A **regression tree** is a step function:

$$g_{\mathcal{T},\beta}(x) = \sum_{k=1}^{K} \beta_k \mathbb{1}_{\chi_k}(x),$$

where $\beta := (\beta_1, \ldots, \beta_K)^\top \in \mathbb{R}^K$ are the **leaf values**, and $\chi_k \subset \mathcal{X}$ so that $\mathcal{T} := \{\chi_k\}_{k=1}^K$ forms a **partition** of $\mathcal{X}$.

2. A T**-additive regression tree** is a sum of regression trees:

$$g_{\mathcal{E},\mathcal{B}}(x) := \sum_{t=1}^{T} g_{\mathcal{T}_t,\beta_t}(x),$$

where $\mathcal{B} := \{\beta_t\}_{t=1}^T$ and $\mathcal{E} := \{\mathcal{T}_t\}_{t=1}^T$.

3. A **Bayesian additive regression tree** (BART) is any distribution on the family of T-additive regression trees
   - This can be done by specifying a (prior) distribution on the leaf values $\mathcal{B}$ and partition $\mathcal{E}$ (Chipman et al. 1998, 2010).

From BART to BART-Int

Modelling f with BART: Posit a BART prior on function f → condition on data $\{x_i, y_i\}_{i=1}^n$ → induce a posterior distribution $\mathbb{P}_n$ (with density p_n), whose mean is

$$f(x) \approx g^n(x) \coloneqq \mathbb{E}_{\mathbb{P}_n}[g_{\mathcal{E},\mathcal{B}}(x)] = \int_\Omega g_{\mathcal{E},\mathcal{B}}(x) p_n(\mathcal{E},\mathcal{B}) d\mathcal{E} d\mathcal{B}.$$

This posterior mean is intractable, but can be estimated by drawing m MCMC samples of trees $\{g_j^n\}_{j=1}^m$ from the BART posterior:

$$g^n(x) \approx \hat{g}^n(x) = \frac{1}{m}\sum_{j=1}^m g_j^n(x) = \frac{1}{m}\sum_{j=1}^m \sum_{t=1}^T \sum_{k=1}^{K_{t,j}} \beta_{t,k}^j \mathbb{1}_{\chi_{t,k}^j}(x).$$

BART-Int estimates $\Pi[f]$ by pushing the above forward through Π:

Definition (BART-Int Estimator (Our contribution))

$$\Pi[\hat{g}^n] = \frac{1}{m}\sum_{j=1}^m \Pi\left[g_j^n\right] = \frac{1}{m}\sum_{j=1}^m \sum_{t=1}^T \sum_{k=1}^{K_{t,j}} \beta_{t,k}^j \Pi\left[\mathbb{1}_{\chi_{t,k}^j}\right].$$

From BART to BART-Int

Modelling f with BART: Posit a BART prior on function f → condition on data $\{x_i, y_i\}_{i=1}^n$ → induce a posterior distribution $\mathbb{P}_n$ (with density p_n), whose mean is

$$f(x) \approx g^n(x) \coloneqq \mathbb{E}_{\mathbb{P}_n}[g_{\mathcal{E},\mathcal{B}}(x)] = \int_\Omega g_{\mathcal{E},\mathcal{B}}(x) p_n(\mathcal{E},\mathcal{B}) d\mathcal{E} d\mathcal{B}.$$

This posterior mean is intractable, but can be estimated by drawing m MCMC samples of trees $\{g_j^n\}_{j=1}^m$ from the BART posterior:

$$g^n(x) \approx \hat{g}^n(x) = \frac{1}{m}\sum_{j=1}^m g_j^n(x) = \frac{1}{m}\sum_{j=1}^m \sum_{t=1}^T \sum_{k=1}^{K_{t,j}} \beta_{t,k}^j \mathbb{1}_{\chi_{t,k}^j}(x).$$

BART-Int estimates $\Pi[f]$ by pushing the above forward through Π:

Definition (BART-Int Estimator (Our contribution))

$$\Pi[\hat{g}^n] = \frac{1}{m}\sum_{j=1}^m \Pi\left[g_j^n\right] = \frac{1}{m}\sum_{j=1}^m \sum_{t=1}^T \sum_{k=1}^{K_{t,j}} \beta_{t,k}^j \Pi\left[\mathbb{1}_{\chi_{t,k}^j}\right].$$

From BART to BART-Int

Modelling f with BART: Posit a BART prior on function f $\rightarrow$ condition on data $\{x_i, y_i\}_{i=1}^n$ $\rightarrow$ induce a posterior distribution $\mathbb{P}_n$ (with density p_n), whose mean is

$$f(x) \approx g^n(x) \coloneqq \mathbb{E}_{\mathbb{P}_n}[g_{\mathcal{E},\mathcal{B}}(x)] = \int_\Omega g_{\mathcal{E},\mathcal{B}}(x) p_n(\mathcal{E},\mathcal{B}) d\mathcal{E} d\mathcal{B}.$$

This posterior mean is intractable, but can be estimated by drawing m MCMC samples of trees $\{g_j^n\}_{j=1}^m$ from the BART posterior:

$$g^n(x) \approx \hat{g}^n(x) = \frac{1}{m}\sum_{j=1}^m g_j^n(x) = \frac{1}{m}\sum_{j=1}^m \sum_{t=1}^T \sum_{k=1}^{K_{t,j}} \beta_{t,k}^j \mathbb{1}_{\chi_{t,k}^j}(x).$$

BART-Int estimates $\Pi[f]$ by pushing the above forward through Π:

Definition (BART-Int Estimator (Our contribution))

$$\Pi[\hat{g}^n] = \frac{1}{m}\sum_{j=1}^m \Pi\left[g_j^n\right] = \frac{1}{m}\sum_{j=1}^m \sum_{t=1}^T \sum_{k=1}^{K_{t,j}} \beta_{t,k}^j \Pi\left[\mathbb{1}_{\chi_{t,k}^j}\right].$$

BART-Int (Our Contribution)

Advantages:

- A T-additive regression tree is a step function, so is discontinuous in nature.
- Computational cost: $\mathcal{O}(Tmn)$ (Pratola et al. 2014).

Disadvantage:

- Unlike GPs, BART posteriors are intractable, so needs to be approximated (e.g. using MCMC).
- BART-Int requires probabilities of the form $\Pi[\mathbb{1}_{\mathcal{X}^j_{t,k}}]$, which are also intractable.
  - Can be approximated, e.g., by using another sample from Π.
  - Corresponds to the issue of intractable kernel means $\Pi[k(\cdot, X)]$ for BQ.

BART-Int (Our Contribution)

Advantages:

- A T-additive regression tree is a step function, so is discontinuous in nature.
- Computational cost: $\mathcal{O}(Tmn)$ (Pratola et al. 2014).

Disadvantage:

- Unlike GPs, BART posteriors are intractable, so needs to be approximated (e.g. using MCMC).
- BART-Int requires probabilities of the form $\Pi[\mathbb{1}_{\chi^j_{t,k}}]$, which are also intractable.
  - Can be approximated, e.g., by using another sample from Π.
  - Corresponds to the issue of intractable kernel means $\Pi[k(\cdot, X)]$ for BQ.

BART-Int (Our Contribution)

Advantages:

- A T-additive regression tree is a step function, so is discontinuous in nature.
- Computational cost: $\mathcal{O}(Tmn)$ (Pratola et al. 2014).

Disadvantage:

- Unlike GPs, BART posteriors are intractable, so needs to be approximated (e.g. using MCMC).
- BART-Int requires probabilities of the form $\Pi[\mathbb{1}_{\chi^j_{t,k}}]$, which are also intractable.
  - Can be approximated, e.g., by using another sample from Π.
  - Corresponds to the issue of intractable kernel means $\Pi[k(\cdot, X)]$ for BQ.

BART-Int (Our Contribution)

Advantages:

- A T-additive regression tree is a step function, so is discontinuous in nature.
- Computational cost: $\mathcal{O}(Tmn)$ (Pratola et al. 2014).

Disadvantage:

- Unlike GPs, BART posteriors are intractable, so needs to be approximated (e.g. using MCMC).
- BART-Int requires probabilities of the form $\Pi[\mathbb{1}_{\chi^j_{t,k}}]$, which are also intractable.
  - Can be approximated, e.g., by using another sample from Π.
  - Corresponds to the issue of intractable kernel means $\Pi[k(\cdot, X)]$ for BQ.

Theoretical Results

Theorem (Concentration Bound for BPNI; informal)

Suppose f is in some normed space $\mathcal{H} \subseteq L^2(\Pi)$, and the BPNI prior g satisfies some regularity conditions. If $\exists N \in \mathbb{N}_+$ such that:

A1. **(Concentration bounds)** $\exists \{\varepsilon_n\}_{n \geq N}$ *such that* $\lim_{n\to\infty} \mathbb{P}[\|f - g\|_n > A_n \varepsilon_n | X^n, y^n] = 0$ *for any* $A_n \to \infty$ *as* $n \to \infty$.

A2. **(Quadrature rates)** $\exists \{\gamma_n\}_{n \geq N}$ *with* $\gamma_n \to 0$ *as* $n \to \infty$ *such that* $\sup_{\|h\|_{\mathcal{H}} \leq 1} |\frac{1}{n} \sum_{i=1}^n h(x_i) - \Pi[h]| = O(\gamma_n)$.

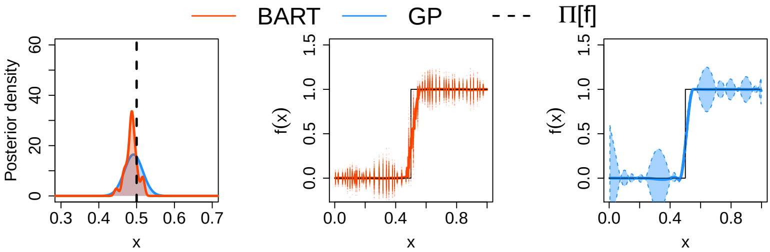
then, we have

$$\lim_{n\to\infty} \mathbb{P}[|\Pi[f] - \Pi[g]| > C_n \max(\varepsilon_n, \gamma_n) | X^n, y^n] = 0$$

for any $C_n \to \infty$ as $n \to \infty$.

Plug in existing results for A1 and A2! (Rockova and Saha 2019, van der Vaart and van Zanten 2011)

$f(x) = \mathbb{1}_{(0.5,1]}(x)$ over $[0,1]$ with BART-Int and BQ with 20 design points with uniform measure.



Experiment II: Portfolio Management (Chan et al. 2012)

Suppose we have d loans to obligators, each with value c_i for $i = 1, \dots, d$. Let x_i denote the **financial strain** on loan i, and suppose p_i is a thresholds after which default occurs. We assume $x_i \sim \text{Exp}(1)$, and define the **portfolio loss** as

$$\ell(x) = \sum_{i=1}^{d} c_i \mathbb{1}_{\{x_i > p_i\}}(x).$$

Probability of making a loss greater than γ:

$$p_\gamma = \int_{\mathcal{X}} \mathbb{1}_{\{\ell(x) > \gamma\}}(x) \Pi(dx).$$

	Method	MAPE	Std. Err.
$d = 5$, $n = 2500$	BART-Int	1.71e-01	2.56e-02
	MI	1.95e-01	2.29e-02
	GP-BQ	**1.68e-01**	2.09e-02
$d = 10$, $n = 5000$	BART-Int	**1.56e-02**	2.35e-03
	MI	9.98e-01	4.47e-04
	GP-BQ	2.72e-02	5.20e-03
$d = 20$, $n = 10000$	BART-Int	**8.40e-03**	1.60e-03
	MI	9.94e-01	6.34e-04
	GP-BQ	2.92e-02	4.90e-03

Summary

- BQ works well for smooth integrands, but is less desirable for discontinuous f.
- We proposed a novel BPNI algorithm, **BART-Int**, using BART instead of a GP.
- Empirically, BART-Int complements, rather than replaces, BQ for discontinuous integrands.

References

[1] H. Zhu, X. Liu, R. Kang, Z. Shen, S. Flaxman, and F.-X. Briol. (2020). *NeurIPS*. Bayesian Probabilistic Numerical Integration with Tree-based Models.

[2] J. Cockayne, C. Oates, T. Sullivan, and M. Girolami. *SIAM Review*. Bayesian probabilistic numerical methods.

[3] Briol, F-X., Oates, C. J., Girolami, M., Osborne, M. A. & Sejdinovic, D. (2019). *Statistical Science*. Probabilistic integration: a role in statistical computation?

[4] H. A. Chipman, E. I. George, and R. E. McCulloch. (2010). *Annals of Applied Statistics*. Bayesian Additive Regression Trees.

[5] Rockova and Saha (2019). *AISTATS*. On Theory for BART.

[6] Linero and Yang (2018). *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*. Bayesian Regression Tree Ensembles that Adapt to Smoothness and Sparsity.

[7] Rockova and van der Pas (2017). *The Annals of Statistics.* Posterior Concentration for Bayesian Regression Trees and Forests.