Bayesian Approaches to Data-driven Stochastic Optimization

Enlu Zhou

School of Industrial and Systems Engineering
Georgia Institute of Technology

Erice Workshop on Robustness and Resilience in Stochastic Optimization and Statistical Learning: Mathematical Foundations
May 23, 2022

Georgia Institute of Technology

Data-driven Stochastic Optimization

- Stochastic optimization:

$$\min_{x \in \mathcal{X}} \mathbb{E}_{\xi \sim \mathbb{P}^c}[h(x;\xi)]$$

- The true distribution $\mathbb{P}^c$ is rarely known in practice.

Data	$\rightarrow$	Model	$\rightarrow$	Decision
observations of ξ		distribution of ξ		x

Data-driven Stochastic Optimization

- Stochastic optimization:

$$\min_{x \in \mathcal{X}} \mathbb{E}_{\xi \sim \mathbb{P}^c}[h(x;\xi)]$$

- The true distribution $\mathbb{P}^c$ is rarely known in practice.

Data	$\rightarrow$	Model	$\rightarrow$	Decision
observations of ξ		distribution of ξ		x

- Distributional Uncertainty: distribution of ξ is estimated from data

Data-driven Stochastic Optimization

- Stochastic optimization:

$$\min_{x \in \mathcal{X}} \mathbb{E}_{\xi \sim \mathbb{P}^c}[h(x;\xi)]$$

- The true distribution $\mathbb{P}^c$ is rarely known in practice.

Data	$\rightarrow$	Model	$\rightarrow$	Decision
observations of ξ		distribution of ξ		x

- Distributional Uncertainty: distribution of ξ is estimated from data

How to deal with unknown distributions?

Empirical Optimization:

$$\min_{x \in \mathcal{X}} \mathbb{E}_{\hat{\mathbb{P}}}[h(x, \xi)]$$

- No consideration of distributional uncertainty
- Solution could be far from the optimal when the data set is small

How to deal with unknown distributions?

Distributionally Robust Optimization (DRO):

$$\min_{x \in \mathcal{X}} \max_{\mathbb{P} \in \mathcal{D}} \mathbb{E}_{\mathbb{P}}[h(x, \xi)]$$

- Model distributional uncertainty by an ambiguity set
  – Treats every point in the set equally likely
- Optimization w.r.t. the worst case
  – Worst case usually happens with a small probability

How to deal with unknown distributions?

- Is there a more convenient and natural way to model distributional uncertainty?

  Bayesian posterior distribution
  – Encodes the likelihoods from data
  – Bayesian consistency

- Is there a less conservative risk attitude?

How to deal with unknown distributions?

- Is there a more convenient and natural way to model distributional uncertainty?
  Bayesian posterior distribution
  – Encodes the likelihoods from data
  – Bayesian consistency

- Is there a less conservative risk attitude?
  Risk measure: from risk-neutral to risk-averse

How to deal with unknown distributions?

- Is there a more convenient and natural way to model distributional uncertainty?
  Bayesian posterior distribution
  – Encodes the likelihoods from data
  – Bayesian consistency

- Is there a less conservative risk attitude?
  Risk measure: from risk-neutral to risk-averse

Bayesian Risk Optimization[1]

- Parameterized distributional uncertainty

$$\min_{x \in \mathcal{X}} \mathbb{E}_{\theta^c}[h(x,\xi)]$$

The randomness $\xi \sim \mathbb{P}(\cdot;\theta^c)$, where θ^c is unknown.

- Model the parameter uncertainty by Bayesian posterior distribution

$$\mathbb{P}_n = \mathbb{P}(\theta|\xi_1,\ldots,\xi_n), \quad \xi_i \overset{\text{iid}}{\sim} \mathbb{P}(\cdot;\theta^c)$$

Bayesian Risk Optimization[1]

- Parameterized distributional uncertainty

$$\min_{x\in\mathcal{X}} \mathbb{E}_{\theta^c}[h(x,\xi)]$$

The randomness $\xi \sim \mathbb{P}(\cdot;\theta^c)$, where θ^c is unknown.

- Model the parameter uncertainty by Bayesian posterior distribution

$$\mathbb{P}_n = \mathbb{P}(\theta|\xi_1,\ldots,\xi_n), \quad \xi_i \overset{\text{iid}}{\sim} \mathbb{P}(\cdot;\theta^c)$$

- Bayesian Risk Optimization (BRO):

$$\min_{x\in\mathcal{X}} \rho_{\mathbb{P}_n}(\underbrace{\mathbb{E}_\theta[h(x,\xi)]}_{=H(x,\theta)}),$$

where ρ is a risk functional (e.g., expectation, mean-variance, VaR^α, CVaR^α).

[1] D. Wu, H. Zhu, E. Zhou, "A Bayesian Risk Approach to Data-Driven Stochastic Optimization: Formulations and Asymptotics", SIAM Journal on Optimization, 2018.

Bayesian Risk Optimization[1]

- Parameterized distributional uncertainty

$$\min_{x \in \mathcal{X}} \mathbb{E}_{\theta^c}[h(x, \xi)]$$

The randomness $\xi \sim \mathbb{P}(\cdot; \theta^c)$, where θ^c is unknown.

- Model the parameter uncertainty by Bayesian posterior distribution

$$\mathbb{P}_n = \mathbb{P}(\theta | \xi_1, \ldots, \xi_n), \quad \xi_i \overset{\text{iid}}{\sim} \mathbb{P}(\cdot; \theta^c)$$

- Bayesian Risk Optimization (BRO):

$$\min_{x \in \mathcal{X}} \rho_{\mathbb{P}_n}(\underbrace{\mathbb{E}_\theta[h(x, \xi)]}_{=H(x,\theta)}),$$

where ρ is a risk functional (e.g., expectation, mean-variance, VaR^α, CVaR^α).

[1] D. Wu, H. Zhu, E. Zhou, "A Bayesian Risk Approach to Data-Driven Stochastic Optimization: Formulations and Asymptotics", SIAM Journal on Optimization, 2018.

BRO and DRO: some connections

BRO: $\min_{x\in\mathcal{X}} \rho_{\mathbb{P}_n}[H(x,\theta)]$, DRO: $\min_{x\in\mathcal{X}} \max_{\theta\in\tilde{\Theta}} H(x,\theta)$

1. DRO interpretation of coherent risk measures:

$$\rho(Z) := \sup_{\zeta\in\mathcal{Z}^*} \left\{\langle\zeta, Z\rangle - \rho^*(\zeta)\right\}, \quad \forall Z \in \mathcal{Z}.$$

2. Setting ρ as the worst-case measure, we can reformulate BRO as a DRO problem:

$$\min_{x\in\mathcal{X}} \max_{\theta\in\Theta} H(x,\theta)$$

Why does BRO work?

- When data size $n \to \infty$, does BRO "recover" the true problem?

  Yes, by consistency of BRO.

- What kind of robustness can we gain from BRO?

  Asymptotic normality of BRO reveals

  BRO objective = posterior mean objective + weight $\times$ CI-width of true performance

  where the weight is controlled by the choice of ρ.

Why does BRO work?

- When data size $n \to \infty$, does BRO "recover" the true problem?

  Yes, by consistency of BRO.

- What kind of robustness can we gain from BRO?

  Asymptotic normality of BRO reveals

  BRO objective = posterior mean objective + weight × CI-width of true performance

  where the weight is controlled by the choice of ρ.

Why does BRO work?

- When data size $n \to \infty$, does BRO "recover" the true problem?

  Yes, by consistency of BRO.

- What kind of robustness can we gain from BRO?

  Asymptotic normality of BRO reveals

  BRO objective $=$ posterior mean objective $+$ weight $\times$ CI-width of true performance

  where the weight is controlled by the choice of ρ.

Consistency of BRO

Bayesian consistency (Lorraine Schwartz 1985)

$\mathbb{P}_n \Rightarrow \delta_{\theta^c}$ as $n \to \infty$, under regularity conditions.

Consistency of objective functions

Suppose the posterior converges and H is continuous in θ for every x in $\mathcal{X}$. Then for every $x \in \mathcal{X}$,

$$\underbrace{\rho_{\mathbb{P}_n}[H(x,\theta)]}_{\text{BRO objective}} \to \underbrace{H(x,\theta^c)}_{\text{true objective}} \quad \text{a.s.} \quad \text{as } n \to \infty.$$

Consistency of optimal solutions

Let $S_n := \arg\min_{x\in\mathcal{X}} \rho_{\mathbb{P}_n}[H(x,\theta)], \quad S := \arg\min_{x\in\mathcal{X}} H(x,\theta^c)$. Under stronger assumptions,

$$\mathbb{D}(S_n, S) := \sup_{x\in S_n} \inf_{y\in S} \|x-y\| \to 0 \quad \text{a.s.} \quad \text{as } n \to \infty,$$

Asymptotic Normality at a Fixed x

1. $\rho =$ mean-variance:

$$\sqrt{n}\,\{\rho_{\mathbb{P}_n}[H(x,\theta)] - H(x,\theta^c)\} \Rightarrow \mathcal{N}(0,\sigma_x^2),$$

2. $\rho =$ VaR:

$$\sqrt{n}\,\{\rho_{\mathbb{P}_n}[H(x,\theta)] - H(x,\theta^c)\} \Rightarrow \mathcal{N}\left(\sigma_x \Phi^{-1}(\alpha), \sigma_x^2\right),$$

3. $\rho =$ CVaR:

$$\sqrt{n}\,\{\rho_{\mathbb{P}_n}[H(x,\theta)] - H(x,\theta^c)\} \Rightarrow \mathcal{N}\left(\frac{\sigma_x}{1-\alpha}\phi(\Phi^{-1}(\alpha)), \sigma_x^2\right),$$

where

$$\sigma_x^2 = \nabla_\theta H(x,\theta^c)^\intercal \mathcal{I}_{\theta^c}^{-1} \nabla_\theta H(x,\theta^c).$$

- $\nabla_\theta H(x,\theta^c)$: sensitivity of H w.r.t. θ^c
- $\mathcal{I}_{\theta^c}$: Fisher information

Asymptotic Normality of Optimal Values

Asymptotics of optimal values

$$\sqrt{n}\left(\min_{x\in\mathcal{X}} \rho_{\mathbb{P}_n}[H(x,\theta)] - \min_{x\in\mathcal{X}} H(x,\theta^c)\right) \Rightarrow \min_{x\in S} Y_x,$$

where $S := \arg\min_{x\in\mathcal{X}} H(x,\theta^c)$ and

$$Y_x := \begin{cases} \nabla_\theta H(x,\theta^c)^\intercal Z & \text{if } \rho = \text{mean / mean-variance} \\ \nabla_\theta H(x,\theta^c)^\intercal Z + \sigma_x \Phi^{-1}(\alpha) & \text{if } \rho = \text{VaR} \\ \nabla_\theta H(x,\theta^c)^\intercal Z + \frac{\phi(\Phi^{-1}(\alpha))}{1-\alpha}\sigma_x & \text{if } \rho = \text{CVaR} \end{cases},$$

where $Z \sim \mathcal{N}(0, [I(\theta^c)]^{-1})$.

Robustness of BRO

- $(1-\beta)100\%$ confidence interval for $H(x, \theta^c)$:

$$\left(\mathrm{VaR}^{\alpha}_{\mathbb{P}_n}[H(x,\theta)] - \Phi^{-1}(\alpha)\frac{\sigma_x}{\sqrt{n}}\right) \pm z_{1-\frac{\beta}{2}}\frac{\sigma_x}{\sqrt{n}}.$$

A wider/narrower CI: less/more confidence about actual performance

-

$$\underbrace{\mathrm{VaR}^{\alpha}_{\mathbb{P}_n}[H(x,\theta)]}_{\text{BRO obj.}} \overset{\mathcal{D}}{\approx} \underbrace{\mathbb{E}_{\mathbb{P}_n}[H(x,\theta)]}_{\text{posterior mean obj.}} + \underbrace{\Phi^{-1}(\alpha)}_{\text{weight}} \underbrace{\frac{\sigma_x}{\sqrt{n}}}_{\propto \text{ CI-width}}$$

BRO seeks a trade-off between posterior mean performance and confidence in the actual performance.

Robustness of BRO

- $(1-\beta)100\%$ confidence interval for $H(x, \theta^c)$:

$$\left(\mathrm{VaR}^{\alpha}_{\mathbb{P}_n}[H(x,\theta)] - \Phi^{-1}(\alpha)\frac{\sigma_x}{\sqrt{n}}\right) \pm z_{1-\frac{\beta}{2}}\frac{\sigma_x}{\sqrt{n}}.$$

A wider/narrower CI: less/more confidence about actual performance

- $$\underbrace{\mathrm{VaR}^{\alpha}_{\mathbb{P}_n}[H(x,\theta)]}_{\text{BRO obj.}} \stackrel{\mathcal{D}}{\approx} \underbrace{\mathbb{E}_{\mathbb{P}_n}[H(x,\theta)]}_{\text{posterior mean obj.}} + \underbrace{\Phi^{-1}(\alpha)}_{\text{weight}} \underbrace{\frac{\sigma_x}{\sqrt{n}}}_{\propto \text{ CI-width}}$$

BRO seeks a trade-off between posterior mean performance and confidence in the actual performance.

Compare with Empirical Optimization

Asymptotic Normality of Empirical Optimization

Let $\hat{\theta}_n$ denote the MLE of n i.i.d. data from $f(\cdot;\theta^c)$. Suppose there is a unique optimal solution x^* to $\min_{x\in\mathcal{X}} H(x,\theta^c)$.

$$\sqrt{n}\left\{\min_{x\in\mathcal{X}} H(x,\hat{\theta})] - H(x^*,\theta^c)\right\} \Rightarrow \nabla_\theta H(x^*,\theta^c)^\intercal Z,$$

where $Z \sim \mathcal{N}(0,[I(\theta^c)]^{-1})$.

- Empirical-MLE formulation has the same asymptotics as the BRO-mean formulation. Not surprising since Bayesian posterior has the same "frequentist guarantee" as MLE in the limit.

Georgia Institute of Technology

Newsvendor example: BRO yields more robust solutions

- Newsvendor: assume demand follows an exponential distribution with unknown mean. Draw 1,000 sets of data, each of size 20.

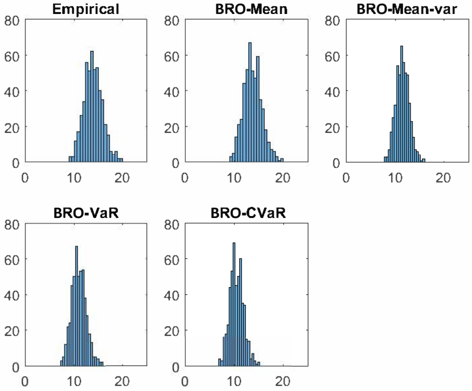


Figure: Empirical and BRO optimal solutions

Newsvendor example: BRO yields more robust solutions

- Newsvendor: assume demand follows an exponential distribution with unknown mean. Draw 1,000 sets of data, each of size 20.

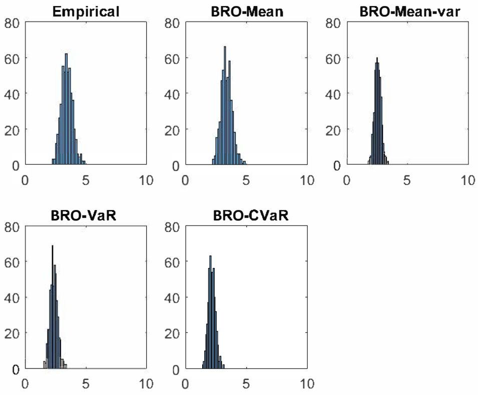


Figure: CI widths for actual performance of Empirical and BRO solutions

How to solve BRO?

$$\text{BRO:} \quad \min_{x \in \mathcal{X}} \rho_{\mathbb{P}_n}\{\mathbb{E}_\theta[h(x,\xi)]\}$$

- If h is convex and ρ is chosen as CVaR, then the problem is convex.
- h is non-convex
    - BRO has a nested objective function to estimate.
    - Gradient of h is available: stochastic gradient descent with (new) nested stochastic gradient estimators [2]
    - h only has black-box evaluation: Bayesian optimization for composition function [3]

[2] S. Cakmak, D. Wu, E. Zhou, "Solving Bayesian Risk Optimization via Nested Stochastic Gradient Estimation", *IISE Transactions*, 2021.

[3] S. Cakmak, R. Astudillo, P. Frazier, E. Zhou, "Bayesian Optimization of Risk Measures", *NeurIPS*, 2020.

How to solve BRO?

$$\text{BRO:} \quad \min_{x \in \mathcal{X}} \rho_{\mathbb{P}_n}\{\mathbb{E}_\theta[h(x,\xi)]\}$$

- If h is convex and ρ is chosen as CVaR, then the problem is convex.
- h is non-convex
  - BRO has a nested objective function to estimate.
  - Gradient of h is available: stochastic gradient descent with (new) nested stochastic gradient estimators [2]
  - h only has black-box evaluation: Bayesian optimization for composition function [3]

---

[2] S. Cakmak, D. Wu, E. Zhou, "Solving Bayesian Risk Optimization via Nested Stochastic Gradient Estimation", *IISE Transactions*, 2021.

[3] S. Cakmak, R. Astudillo, P. Frazier, E. Zhou, "Bayesian Optimization of Risk Measures", *NeurIPS*, 2020.

Bayesian Distributionally Robust Optimization

- The assumed parametric family in BRO sometimes introduces additional model uncertainty.
- Solution: allow more distributions outside the assumed parametric family by constructing an ambiguity set.

[4] A. Shapiro, E. Zhou, and Y. Lin, "Bayesian Distributionally Robust Optimization", arXiv2112.08625.

Bayesian Distributionally Robust Optimization

- The assumed parametric family in BRO sometimes introduces additional model uncertainty.
- Solution: allow more distributions outside the assumed parametric family by constructing an ambiguity set.
- Bayesian Distributionally Robust Optimization (Bayesian-DRO)[4]

$$\min_{x\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_n}\left[\sup_{Q\in\mathcal{M}^\theta} \mathbb{E}_Q[h(x,\xi)]\right],$$

where $\xi \sim Q$, $\theta \sim \mathbb{P}_n$, and $\mathcal{M}^\theta$ is an *ambiguity set* around $f(\cdot;\theta)$ constrained by ϕ-divergence $\leq \epsilon$.

[4] A. Shapiro, E. Zhou, and Y. Lin, "Bayesian Distributionally Robust Optimization", arXiv2112.08625.

Bayesian Distributionally Robust Optimization

- The assumed parametric family in BRO sometimes introduces additional model uncertainty.
- Solution: allow more distributions outside the assumed parametric family by constructing an ambiguity set.
- Bayesian Distributionally Robust Optimization (Bayesian-DRO)[4]

$$\min_{x \in \mathcal{X}} \mathbb{E}_{\mathbb{P}_n} \left[\sup_{Q \in \mathcal{M}^\theta} \mathbb{E}_Q[h(x, \xi)] \right],$$

where $\xi \sim Q$, $\theta \sim \mathbb{P}_n$, and $\mathcal{M}^\theta$ is an *ambiguity set* around $f(\cdot;\theta)$ constrained by ϕ-divergence $\leq \epsilon$.

[4] A. Shapiro, E. Zhou, and Y. Lin, "Bayesian Distributionally Robust Optimization", arXiv2112.08625.

Robustness of Bayesian-DRO

- For small $\epsilon > 0$ and KL divergence,

$$\underbrace{\mathbb{E}_{\mathbb{P}_n}\left[\sup_{Q\in\mathcal{M}^\theta} \mathbb{E}_Q[h(x,\xi)]\right]}_{\text{Bayesian-DRO objective}} \approx \underbrace{\mathbb{E}_{\mathbb{P}_N}\left[\mathbb{E}_\theta[h(x,\xi)]\right]}_{\text{posterior mean}} + \underbrace{\sqrt{2\epsilon}}_{\text{weight}}\, \underbrace{\mathbb{E}_{\mathbb{P}_N}\left[\sigma_\theta[h(x,\xi)]\right]}_{\text{posterior std. dev.}}.$$

- Similar interpretation has also been observed for empirical DRO (see Gotoh et al. (2018), Duchi et al. (2021)).

Consistency of Bayesian-DRO

- True distribution: q_*; Parametric model: f_θ

$$\Theta^* := \arg\min_{\theta \in \Theta} D_{KL}(q_* \| f_\theta)$$

If the model is correct, then $\Theta^* = \{\theta \in \Theta : q_* = f_\theta\}$.

- If $\Theta^* = \{\theta^*\}$ is the singleton, then for almost every data sequence $\{\xi_1, \ldots\}$, $\theta_n \xrightarrow{p} \theta^*$.

Consistency of Bayesian-DRO

- True distribution: q_*; Parametric model: f_θ

$$\Theta^* := \arg\min_{\theta \in \Theta} D_{KL}(q_* \| f_\theta)$$

  If the model is correct, then $\Theta^* = \{\theta \in \Theta : q_* = f_\theta\}$.

- If $\Theta^* = \{\theta^*\}$ is the singleton, then for almost every data sequence $\{\xi_1, \ldots\}$, $\theta_n \xrightarrow{p} \theta^*$.
- Almost sure convergence of Bayesian-DRO objective values and optimal solutions as data size $n \to \infty$.

Consistency of Bayesian-DRO

- True distribution: q_*; Parametric model: f_θ

$$\Theta^* := \arg\min_{\theta \in \Theta} D_{KL}(q_* \| f_\theta)$$

  If the model is correct, then $\Theta^* = \{\theta \in \Theta : q_* = f_\theta\}$.
- If $\Theta^* = \{\theta^*\}$ is the singleton, then for almost every data sequence $\{\xi_1, \ldots\}$, $\theta_n \xrightarrow{p} \theta^*$.
- Almost sure convergence of Bayesian-DRO objective values and optimal solutions as data size $n \to \infty$.

Newsvendor Example

Customer demand: true distribution is truncated Normal, chosen parametric family is Exponential (with mean parameter θ).

Table: Non-contaminated data: all data from true distribution

data size=50	Bayesian-DRO	BRO-mean	empirical	true
solution	18.05	10.68	18.71	19.28
out-of-sample mean	11.50	28.17	11.08	10.95
out-of-sample std	8.37	11.79	8.25	8.33

Table: Contaminated data: 80% from true distr., 20% from another distr.

data size=50	Bayesian-DRO	BRO-mean	empirical	true
solution	19.45	9.45	17.27	19.28
out-of-sample mean	10.96	31.80	12.33	10.95
out-of-sample std	8.39	11.91	8.34	8.33

- Bayesian-DRO is robust against distribution model mis-specification.
- Bayesian-DRO is also robust against contaminated data.

Some extensions

Bayesian approaches are natural for streaming data and dynamic settings.

- Online stochastic optimization under streaming data[5]
- Multi-stage stochastic optimization with Bayesian learning[6]
- Bayesian risk Markov decision processes[7]

---

[5] T. Liu, Y. Lin, and E. Zhou, "Bayesian Stochastic Gradient Descent for Stochastic Optimization with Streaming Input Data", arXiv2202.07581.

[6] Y. Li, T. Liu, E. Zhou, and F. Zhang, "Bayesian Learning Model Predictive Control for Process-aware Source Seeking", *IEEE Control Systems Letters*, 2021

[7] Y. Lin, Y. Ren, and E. Zhou, "Bayesian Risk Markov Decision Processes", submitted.

Online Stochastic Optimization with Streaming Data[5]

$$\min_{x\in\mathcal{X}} \mathbb{E}_{\theta^c}[h(x,\xi)] \quad (\theta^c \text{ unknown})$$

Data of ξ come in batches sequentially over time.

- Exogenous (decision-independent) uncertainty: $\xi \sim f(\cdot;\theta^c), \forall x$.
- Endogenous (decision-dependent) uncertainty: $\xi \sim f(\cdot;x,\theta^c)$.

  Examples: Source seeking (signal depends on location);

  Dynamic pricing (demand depends on price).

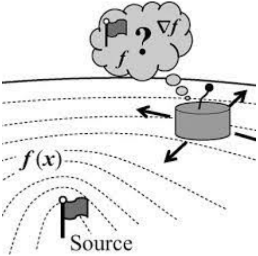
[5] T. Liu, Y. Lin, and E. Zhou, "Bayesian Stochastic Gradient Descent for Stochastic Optimization with Streaming Input Data", arXiv2202.07581.

Online Stochastic Optimization with Streaming Data[5]

$$\min_{x \in \mathcal{X}} \mathbb{E}_{\theta^c}[h(x,\xi)] \quad (\theta^c \text{ unknown})$$

Data of ξ come in batches sequentially over time.

- Exogenous (decision-independent) uncertainty: $\xi \sim f(\cdot;\theta^c), \forall x$.
- Endogenous (decision-dependent) uncertainty: $\xi \sim f(\cdot;x,\theta^c)$.
  Examples: Source seeking (signal depends on location);
  Dynamic pricing (demand depends on price).



[5] T. Liu, Y. Lin, and E. Zhou, "Bayesian Stochastic Gradient Descent for Stochastic Optimization with Streaming Input Data", arXiv2202.07581.

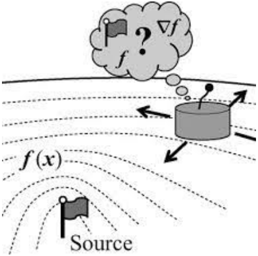
Online Stochastic Optimization with Streaming Data[5]

$$\min_{x \in \mathcal{X}} \mathbb{E}_{\theta^c}[h(x, \xi)] \quad (\theta^c \text{ unknown})$$

Data of ξ come in batches sequentially over time.

- Exogenous (decision-independent) uncertainty: $\xi \sim f(\cdot; \theta^c), \forall x$.
- Endogenous (decision-dependent) uncertainty: $\xi \sim f(\cdot; x, \theta^c)$.

  Examples: Source seeking (signal depends on location);
  Dynamic pricing (demand depends on price).



[5]T. Liu, Y. Lin, and E. Zhou, "Bayesian Stochastic Gradient Descent for Stochastic Optimization with Streaming Input Data", arXiv2202.07581.

Algorithm: decision-dependent case

At time t, given a new batch of data $\xi_t \sim f(\cdot; x_t, \theta^c)$, do the following.

1. Bayesian updating of the posterior:

$$\pi_t(\theta) \propto \pi_{t-1}(\theta) f(\xi_t | x_t, \theta).$$

2. Use one or more iterates of SGD to solve

$$\min_{x \in \mathcal{X}} H(x, \pi_t) = \underbrace{\mathbb{E}_{\theta \sim \pi_t} \mathbb{E}_{\xi \sim f(\cdot; x_t, \theta)}[h(x, \xi)]}_{\text{BRO-mean}}$$

Under mild conditions

$$x_{t+1} = \Pi_{\mathcal{X}}\{x_t - \alpha_t \underbrace{[\nabla_x h(x_t, \xi) + h(x_t, \xi) \nabla_x \ln f(\xi; x_t, \theta)]}_{\text{unbiased gradient estimator}}\}$$

Convergence: even with non-i.i.d. data $\{\xi_t\}_t$, the posterior $\{\pi_t\}_t$ is strongly consistent and $\{x_t\}_t$ converges weakly.

Algorithm: decision-dependent case

At time t, given a new batch of data $\xi_t \sim f(\cdot; x_t, \theta^c)$, do the following.

1. Bayesian updating of the posterior:

$$\pi_t(\theta) \propto \pi_{t-1}(\theta) f(\xi_t | x_t, \theta).$$

2. Use one or more iterates of SGD to solve

$$\min_{x \in \mathcal{X}} H(x, \pi_t) = \underbrace{\mathbb{E}_{\theta \sim \pi_t} \mathbb{E}_{\xi \sim f(\cdot; x_t, \theta)}[h(x, \xi)]}_{\text{BRO-mean}}$$

Under mild conditions

$$x_{t+1} = \Pi_{\mathcal{X}}\{x_t - \alpha_t \underbrace{[\nabla_x h(x_t, \xi) + h(x_t, \xi) \nabla_x \ln f(\xi; x_t, \theta)]}_{\text{unbiased gradient estimator}}\}$$

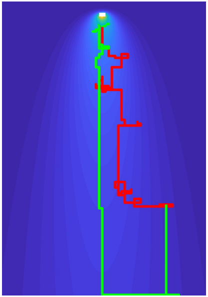
Convergence: even with non-i.i.d. data $\{\xi_t\}_t$, the posterior $\{\pi_t\}_t$ is strongly consistent and $\{x_t\}_t$ converges weakly.

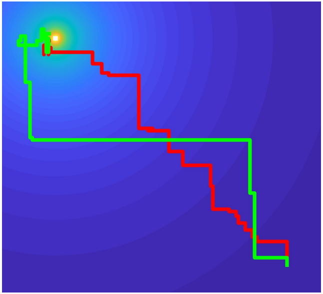
Application: source seeking[6]

Adapt the previous algorithm to robot source seeking

- solve a model predictive control problem instead of the original
- replace SGD by a neighborhood search

Figure: Trajectories of our proposed algorithm (red) and the expected rate algorithm (green) in different scenarios (with and without wind)





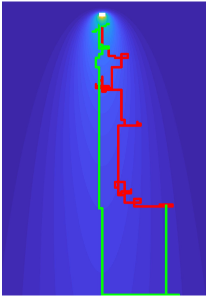
[6] Y. Li, T. Liu, E. Zhou, and F. Zhang, "Bayesian Learning Model Predictive Control for Process-aware Source Seeking", *IEEE Control Systems Letters*, 2021.

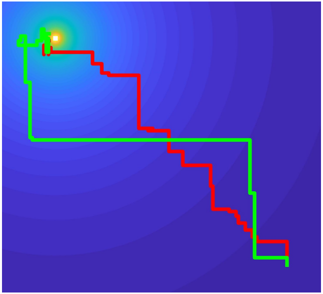
Application: source seeking[6]

Adapt the previous algorithm to robot source seeking

- solve a model predictive control problem instead of the original
- replace SGD by a neighborhood search

Figure: Trajectories of our proposed algorithm (red) and the expected rate algorithm (green) in different scenarios (with and without wind)





[6] Y. Li, T. Liu, E. Zhou, and F. Zhang, "Bayesian Learning Model Predictive Control for Process-aware Source Seeking", *IEEE Control Systems Letters*, 2021.

Application: source seeking

Table: Comparison between our algorithm with expected rate algorithm. Results are averaged over 20 stochastic simulations.

Scenario	Algorithm	Trajectory Length	Measurements	Total Search Time
Without Wind	Our Alg.	254.7	76.4	636.7
	Expected Rate	261.5	261.5	1569
With Wind	Our Alg.	275.7	65.1	601.2
	Expected Rate	252.7	252.7	1516.2

Conclusions

- Bayesian posterior distributions (instead of ambiguity sets) provide a natural and convenient way to model distributional uncertainty.
- Two new formulations: Bayesian Risk Optimization (BRO), Bayesian distributionally robust optimization (Bayesian-DRO). Both have consistency and robustness.
- Bayesian approaches are amenable to streaming data and dynamic settings.

Collaborators:

Alexander Shapiro (ISyE, Georgia Tech), Fumin Zhang (ECE, Georgia Tech), Peter Frazier (ORIE, Cornell)

Di Wu, Helin Zhu, Sait Cakmak, Tianyi Liu, Yifan Lin, Rahul Austillo, Yingke Li

References

1 D. Wu, H. Zhu, E. Zhou, "A Bayesian Risk Approach to Data-Driven Stochastic Optimization: Formulations and Asymptotics", *SIAM Journal on Optimization*, 2018. (INFORMS Outstanding Simulation Publication Award, 2020)

2 S. Cakmak, D. Wu, E. Zhou, "Solving Bayesian Risk Optimization via Nested Stochastic Gradient Estimation", *IISE Transactions*, 2021.

3 S. Cakmak, R. Astudillo, P. Frazier, E. Zhou, "Bayesian Optimization of Risk Measures", *NeurIPS*, 2020.

4 A. Shapiro, E. Zhou, and Y. Lin, "Bayesian Distributionally Robust Optimization", arXiv2112.08625.

5 T. Liu, Y. Lin, and E. Zhou, "Bayesian Stochastic Gradient Descent for Stochastic Optimization with Streaming Input Data", arXiv2202.07581.

6 Y. Li, T. Liu, E. Zhou, and F. Zhang, "Bayesian Learning Model Predictive Control for Process-aware Source Seeking", *IEEE Control Systems Letters*, 2021.

7 Y. Lin, Y. Ren, and E. Zhou, "Bayesian Risk Markov Decision Processes", submitted.

Thank you!