Sample Average Approximation

Jin Fulong

University of Science and Technology of China

9th May 2025

What's SAA

- Consider that ξ a random variable on a Carathéodory function F

$$\mathrm{Min}_{x\in\mathcal{X}}\{f(x) := \mathbb{E}[F(x,\xi)]\},$$

- Under distribution of ξ unkown, we have to deal with

$$\mathrm{Min}_{x\in\mathcal{X}}\left\{\hat{f}_N(x) := \frac{1}{N}\sum_{j=1}^{N} F\left(x,\xi^j\right)\right\}$$

How we make sure such two solution converge when $N \to \infty$?

Theory guarantee

- **Law of Large Numbers**: Under regularity conditions, the sample average function $\hat{f}_N(x)$ converges pointwise almost surely (w.p.1) to the true objective function $f(x)$ as $N \to \infty$.
- **Unbiase**: $\hat{f}_N(x)$ is an unbiased estimator of $f(x)$, meaning

$$\mathbb{E}\left[\hat{f}_N(x)\right] = f(x)$$

- **Consistency**: As N increases, the optimal value $\hat{\vartheta}_N$ and optimal solution set $\hat{\mathcal{S}}_N$ of the SAA problem converge to their true counterparts (ϑ^* and $\mathcal{S}$) in the original problem.

Equality

- **Pointwise Sequence Convergence**: For any $\bar{x} \in \mathcal{X}$ and any sequence $\{x_N\} \subset \mathcal{X}$ converging to $\bar{x}$, it holds that

$$f_N\left(x_N\right) \to f(\bar{x})$$

- **Continuity & Local Uniform Convergence**: if $f(\cdot)$ is continuous on $\mathcal{X}$, then

$$f_N(\cdot) \to f(\cdot) \text{ uniformly on every compact subset of } \mathcal{X}$$

convergence of optimal set $\mathcal{S}_N$

Theorem (convergence of optimal set $\mathcal{S}_N$)

Suppose that there exists a compact set $C \subset \mathbb{R}^n$ such that

1. *the set $\mathcal{S}$ of the true problem is nonempty and $\mathcal{S} \subset C$,*
2. *the function $f(x)$ is finite valued and continuous on C,*
3. *$\hat{f}_N(x)$ converges to $f(x)$ w.p.l, as $N \to \infty$, uniformly in $x \in C$,*
4. *w.p.1 for N large enough the set $\hat{\mathcal{S}}_N$ is nonempty and $\hat{\mathcal{S}}_N \subset C$.*

Then $\hat{\vartheta}_N \to \vartheta^*$ and $\mathbb{D}\left(\hat{\mathcal{S}}_N, \mathcal{S}\right) \to 0$ w.p. 1 as $N \to \infty$.

Remark

- Condition 2 and 3 yields **Pointwise Sequence Convergence**

$$f_N(x_N) \to f(\bar{x})$$

- The distance $\mathbb{D}\left(\hat{\mathcal{S}}_N, \mathcal{S}\right) \to 0$ w.p. 1 means any measurable selection $\hat{x}_N \in \hat{\mathcal{S}}_N$ satisfies $\operatorname{dist}(\hat{x}_N, \mathcal{S}) \to 0$ w.p.1.

Convexity construction

In most time, we will adapt **contraint** optimal problem in such a **uncontraint** optimal problem by

$$\min f_N(x) + \mathbb{I}_{\mathcal{X}}(x)$$

where,

$$\mathbb{I}_{\mathcal{X}}(x) = \begin{cases} 0 & \text{if } x \in \mathcal{X}, \\ +\infty & \text{else } x \notin \mathcal{X}. \end{cases}$$

The above "penalization" operation preserves convexity of respective functions

Weaken version

Theorem (convergence of optimal set $\mathcal{S}_N$)

Suppose that

1. F *is random lower semi-continuous,*
2. *For almost every* $\xi \in \Xi$, $F(\cdot, \xi)$ *is convex,*
3. $\mathcal{X}$ *is closed and convex*
4. f *is lower semi-continuous and there exists a point* $\bar{x} \in \mathcal{X}$ *such that* $f(x) < +\infty$ *for all* x *in a neighborhood of* $\bar{x}$*,*
5. *the set* $\mathcal{S}$ *of optimal solutions of the true problem is nonempty and bounded,*
6. *the LLN holds pointwise.*

Then $\hat{\vartheta}_N \to \vartheta^*$ *and* $\mathbb{D}\left(\hat{\mathcal{S}}_N, \mathcal{S}\right) \to 0$ *w.p. 1 as* $N \to \infty$.

Remark

1. **Global infinity** $\underbrace{\rightarrow}_{\text{reduction}}$ $\forall x \in B(\bar{x}, \delta), \quad f(x) < +\infty$
2. **Global Continuity** $\underbrace{\rightarrow}_{\text{reduction}}$ $F(\cdot, \xi)$ is lower semi-continuous.

Central Limit Theorem (CLT)

- If fix a point $x \in \mathcal{X}$, by CLT

$$N^{1/2}\left[\hat{f}_N(x) - f(x)\right] \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma^2(x)\right)$$

- Consider $x \in S_N$, under what condition satiating ?

$$N^{1/2}\left(\hat{\vartheta}_N - \vartheta^*\right) \xrightarrow{\mathcal{D}} \inf_{x \in \mathcal{S}} \mathcal{N}\left(0, \sigma^2(x)\right)$$

Assumptions

A1 For some point $\tilde{x} \in \mathcal{X}$ the expectation $\mathbb{E}\left[F(\tilde{x}, \xi)^2\right]$ is finite.

A2 There exists a measurable function $G : \Xi \to \mathbb{R}_+$such that $\mathbb{E}\left[G(\xi)^2\right]$ is finite and

$$\left|F(x, \xi) - F\left(x', \xi\right)\right| \leq G(\xi) \left\|x - x'\right\|,$$

for all $x, x' \in \mathcal{X}$ and a.e. $\xi \in \Xi$.

First Order Asymptotics

Theorem (First Order Asymptotics)

Let $\hat{\vartheta}_N$ be the optimal value of the SAA problem. Suppose that the sample is i.i.d., the set $\mathcal{X}$ is compact, and assumptions (A1) and (A2) are satisfied. Then the following holds

$$\hat{\vartheta}_N = \inf_{x \in \mathcal{S}} \hat{f}_N(x) + o_p\left(N^{-1/2}\right)$$

$$N^{1/2}\left(\hat{\vartheta}_N - \vartheta^*\right) \xrightarrow{\mathcal{D}} \inf_{x \in \mathcal{S}} \mathcal{N}\left(0, \sigma^2(x)\right)$$

If, moreover, $\mathcal{S} = \{\bar{x}\}$ is a singleton, then

$$N^{1/2}\left(\hat{\vartheta}_N - \vartheta^*\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma^2(\bar{x})\right)$$

Danskin Theorem

Lemma (Danskin Theorem)

Consider a continuous f

$$f(x) = \max_{u \in U} \phi(x, u)$$

- U *is non-empty compact set*
- $\phi(x, u)$ *is differential on* x.

Then it can satisfy

$$\lim_{\epsilon \to 0^+} \frac{f(x + \epsilon d) - f(x)}{\epsilon} = \max_{u \in U(x)} \langle \nabla_x \phi(x, u), d \rangle$$

Proof sketch

- $C(\mathcal{X})$ defined on $\mathcal{X}$ functinal space, define a norm $\|\psi\| := \sup_{x\in\mathcal{X}} |\psi(x)|$, by Danskin Theorem,

$$V(\psi) := \inf_{x\in\mathcal{X}} \psi(x),$$

- By A2, we have

$$|V(\psi_1) - V(\psi_2)| \leq \|\psi_1 - \psi_2\|,$$

- $V(\cdot)$ is directionally differentiable at any $\mu \in C(\mathcal{X})$,

$$V'_\mu(\delta) = \inf_{x\in\overline{\mathcal{X}}(\mu)} \delta(x), \quad \overline{\mathcal{X}}(\mu) := \operatorname*{argmin}_{x\in\mathcal{X}} \mu(x)$$

Then by delta method, the result is explicit

Convex form (weaken form)

The compactness of $\mathcal{X}$ can also be weaken by

1. $\mathcal{X}$ is close and convex
2. $F(x, \xi)$ is convex on x
3. $\mathcal{S}$ is non-empty and bound

We can choose a compact set $\mathcal{V} \supset \mathcal{S}$, construct

$$\tilde{\vartheta}_N := \inf_{x \in \mathcal{V}} \hat{f}_N(x)$$

- By theorem 3, $N^{1/2}\left(\tilde{\vartheta}_N - \hat{\vartheta}_N\right) \xrightarrow{p} 0$
- A2 togather with $\mathcal{X}$'s closeness and convexity shows that $\mathcal{S}$ contains a compact subset.

Second order Delta method

Theorem (Second order Delta method)

Denote $\{Y_N\}$ a random sequence, G be secondly order differentiable and $\{\tau_N\}$ be a sequence of positive numbers tending to ∞, then

$$\tau_N^2\left[G\left(Y_N\right)-G(\mu)-G'_\mu\left(Y_N-\mu\right)\right] \xrightarrow{\mathcal{D}} \frac{1}{2}G''_\mu(Y)$$

assumptions

S1 The function $f(x)$ is Lipschitz continuous on U, has unique minimizer $\bar{x}$ over $x \in \mathcal{X}$, and is twice continuously differentiable at $\bar{x}$.

S2 The set $\mathcal{X}$ is second order regular at $\bar{x}$.

S3 The quadratic growth condition holds at $\bar{x}$.

S4 Function $F(\cdot, \xi)$ is Lipschitz continuous on U and differentiable at $\bar{x}$ for a.e. $\xi \in \Xi$.

Second optimal

Theorem

Suppose that the assumptions $(\mathrm{S1})-(\mathrm{S4})$ *hold and* $N^{1/2}\left(\hat{f}_N - f\right)$ *converges in distribution to a random element* Y *of* $W^{1,\infty}(U)$. *Then*

$$\hat{\vartheta}_N = \hat{f}_N(\bar{x}) + \frac{1}{2}V_f''\left(\hat{f}_N - f\right) + o_p\left(N^{-1}\right)$$

and

$$N\left[\hat{\vartheta}_N - \hat{f}_N(\bar{x})\right] \xrightarrow{\mathcal{D}} \frac{1}{2}V_f''(Y)$$

Monte Carlo problem

- Assume we can generate sample $\xi^1, \dots, \xi^N$, if we use the **same sample**, instead get

$$\mathrm{Var}\left[\hat{f}_N(x_1) - \hat{f}_N(x_2)\right] = \mathrm{Var}\left[\hat{f}_N(x_1)\right] + \mathrm{Var}\left[\hat{f}_N(x_2)\right].$$

but

$$\mathrm{Var}\left[\hat{f}_N(x_1) - \hat{f}_N(x_2)\right] = \mathrm{Var}\left[\hat{f}_N(x_1)\right] + \mathrm{Var}\left[\hat{f}_N(x_2)\right] - 2\,\mathrm{Cov}\left(\hat{f}_N(x_1), \hat{f}_N(x_2)\right)$$

- How large N guarantee enough accuracy for convergence by LLN?

Before we start

- ε-optimal solutions of the true and the SAA problems:

$$\mathcal{S}^{\varepsilon} := \{x \in \mathcal{X} : f(x) \leq \vartheta^* + \varepsilon\}$$

and

$$\hat{\mathcal{S}}_N^{\varepsilon} := \left\{x \in \mathcal{X} : \hat{f}_N(x) \leq \hat{\vartheta}_N + \varepsilon\right\}$$

- For parameters $\varepsilon \geq 0$ and $\delta \in [0, \varepsilon]$, consider the event $\left\{\hat{\mathcal{S}}_N^{\delta} \subset \mathcal{S}^{\varepsilon}\right\}$. This event means that any δ-optimal solution of the SAA problem is an ε-optimal solution of the true problem.

Inequality construction

- We estimate now the probability of that event

$$\left\{\hat{\mathcal{S}}_N^\delta \not\subset \mathcal{S}^\varepsilon\right\} = \bigcup_{x \in \mathcal{X} \setminus \mathcal{S}^\varepsilon} \bigcap_{y \in \mathcal{X}} \left\{\hat{f}_N(x) \leq \hat{f}_N(y) + \delta\right\}$$

hence,

$$\Pr\left(\hat{\mathcal{S}}_N^\delta \not\subset \mathcal{S}^\varepsilon\right) \leq \sum_{x \in \mathcal{X} \setminus \mathcal{S}^\varepsilon} \Pr\left(\bigcap_{y \in \mathcal{X}} \left\{\hat{f}_N(x) \leq \hat{f}_N(y) + \delta\right\}\right)$$

Inequality construction

- Consider a mapping $u : \mathcal{X} \backslash \mathcal{S}^\varepsilon \to \mathcal{X}$. it follows

$$\Pr\left(\hat{\mathcal{S}}_N^\delta \not\subset \mathcal{S}^\varepsilon\right) \leq \sum_{x \in \mathcal{X} \backslash \mathcal{S}^\varepsilon} \Pr\left\{\hat{f}_N(x) - \hat{f}_N(u(x)) \leq \delta\right\}$$

- Assume that the mapping $u(\cdot)$ is chosen in such a way that

$$f(u(x)) \leq f(x) - \varepsilon^* \quad \text{for all } x \in \mathcal{X} \backslash \mathcal{S}^\varepsilon$$

- For some $\varepsilon^* \geq \varepsilon$, note that such a mapping always exists with

$$\varepsilon^* := \min_{x \in \mathcal{X} \backslash \mathcal{S}^\varepsilon} f(x) - \vartheta^*$$

Inequality construction

- Note that $\mathbb{E}[Y(x,\xi)] = f(u(x)) - f(x)$, and hence $\mathbb{E}[Y(x,\xi)] \leq -\varepsilon^*$ for all $x \in \mathcal{X} \backslash S^{\varepsilon}$. The corresponding sample average is

$$\hat{Y}_N(x) := \frac{1}{N}\sum_{j=1}^{N} Y\left(x, \xi^j\right) = \hat{f}_N(u(x)) - \hat{f}_N(x)$$

- The inequality together with the LD upper bound implies

$$\Pr\left(\hat{\mathcal{S}}_N^{\delta} \not\subset \mathcal{S}^{\varepsilon}\right) \leq \sum_{x \in \mathcal{X} \backslash \mathcal{S}^{\varepsilon}} \Pr\left\{\hat{Y}_N(x) \geq -\delta\right\} \leq \sum_{x \in \mathcal{X} \backslash \mathcal{S}^{\varepsilon}} e^{-N I_x(-\delta)}$$

where $I(\delta) = \sup_{t\in\mathbb{R}}\{t\delta - \ln \mathbb{E}(e^{tY})\}$

Well-defined sequence

M2 For every $x \in \mathcal{X} \backslash \mathcal{S}^{\varepsilon}$ the moment $\mathbb{E}\left[e^{tY(x,\xi)}\right]$, of the random variable $Y(x,\xi) = F(u(x),\xi) - F(x,\xi)$, is finite valued in a neighborhood of $t = 0$.

Theorem (Well-defined sequence)

Let ε and δ be nonnegative numbers. Then

$$1 - \Pr\left(\hat{\mathcal{S}}_N^{\delta} \subset \mathcal{S}^{\varepsilon}\right) \leq |\mathcal{X}| e^{-N\eta(\delta,\varepsilon)}$$

where

$$\eta(\delta,\varepsilon) := \min_{x \in \mathcal{X} \backslash \mathcal{S}^{\varepsilon}} I_x(-\delta)$$

Moreover, if $\delta < \varepsilon^*$ and assumption (M2) holds, then $\eta(\delta,\varepsilon) > 0$.

sub-gaussian

M3 There is a constant $\sigma > 0$ such that for every $x \in \mathcal{X} \backslash \mathcal{S}^{\varepsilon}$ the random variable $Y(x, \xi) - \mathbb{E}[Y(x, \xi)]$ is σ-subgaussian, i.e., its moment generating function $M_x(t)$ satisfies

$$M_x(t) \leq \exp\left(\sigma^2 t^2 / 2\right), \quad \forall t \in \mathbb{R}$$

- From M3,

$$\ln \mathbb{E}\left[e^{tY(x,\xi)}\right] - t\mathbb{E}[Y(x, \xi)] = \ln M_x(t) \leq \sigma^2 t^2 / 2$$

- hence the rate function

$$I_x(z) \geq \sup_{t \in \mathbb{R}} \left\{t(z - \mathbb{E}[Y(x, \xi)]) - \sigma^2 t^2 / 2\right\} = \frac{(z - \mathbb{E}[Y(x, \xi)])^2}{2\sigma^2}$$

Remark

- Actually,

$$I_x(-\delta) \geq \frac{(-\delta - \mathbb{E}[Y(x,\xi)])^2}{2\sigma^2} \geq \frac{(\varepsilon^* - \delta)^2}{2\sigma^2} \geq \frac{(\varepsilon - \delta)^2}{2\sigma^2}$$

- Theorem takes the form

$$1 - \Pr\left(\hat{\mathcal{S}}_N^\delta \subset \mathcal{S}^\varepsilon\right) \leq |\mathcal{X}| e^{-N(\varepsilon-\delta)^2/(2\sigma^2)}$$

Convergence accuracy

Theorem (Convergence accuracy)

Suppose that assumptions (M1) and (M3) hold. Then for $\varepsilon > 0, 0 \leq \delta < \varepsilon$, and $\alpha \in (0,1)$, and for the sample size N satisfying

$$N \geq \frac{2\sigma^2}{(\varepsilon - \delta)^2} \ln\left(\frac{|\mathcal{X}|}{\alpha}\right)$$

it follows that

$$\Pr\left(\hat{\mathcal{S}}_N^{\delta} \subset \mathcal{S}^{\varepsilon}\right) \geq 1 - \alpha$$

Consistency of SAA

Asymptotics of the Optimal Value x^*

- second asymptotics

Monte Carlo

Variance Reduction Techniques

- Latin Hypercube Sampling

Latin Hypercube Sampling

- Assume $\xi \sim H$

$$\mathbb{E}[F(x,\xi)] = \int_{-\infty}^{+\infty} F(x,\xi)dH(\xi)$$

- Generate $\xi^j = H^{-1}(U^j)$

$$U^j \sim U[(j-1)/N, j/N], \quad j = 1, \ldots, N$$

- Sample permutation $\xi^{j_1}, \ldots, \xi^{j_N}$, by conditinal variance we have

$$\mathrm{Var}\left[\hat{f}_N(x)\right] = N^{-1}\sigma^2(x) + 2N^{-2}\sum_{s<t} \mathrm{Cov}\left(F\left(x,\xi^{j_s}\right), F\left(x,\xi^{j_t}\right)\right)$$