Bias evaluation and reduction for sample-path optimization

Fabian Bastin[1]

[1] Department of Computing Science and Operational Research
Université de Montréal; CIRRELT
Québec, Canada

Eigth IMACS Seminar on Monte Carlo Methods
August 29 – September 2, 2011, Borovets BULGARIA

Context

We consider the general stochastic program (SP)

$$\min_{x \in \mathcal{X}} g\left(E[f(x,\xi)]\right),$$

where

- $\mathcal{X}$ is a compact set in $\mathcal{R}^n$;
- $\xi = (\xi_1, \ldots, \xi_m)$ is a random vector of size m;
- $f: \mathcal{R}^n \times \mathcal{R}^m \to \mathcal{R}$;
- $g: \mathcal{R} \to \mathcal{R}$.

We will also denote

$$f(x) := E[f(x,\xi)].$$

Context (cont'd)

g is usually the identity function, so we have

$$\min_{x \in \mathcal{X}} E[f(x, \xi)].$$

This problem has (and is) studied extensively (Bayraksan, Homem-de-Mello, Morton, Robinson, Royset, Pasupathy, Shapiro,...).

Example 1: constraining nonlinear programming

$$\min_x f(x)$$
$$\text{s.t. } E[c_i(x,\xi)] \geq 0,\ i = 1,\dots,s.$$

Log-barrier methods will replace this problem by a sequence of unconstrained problems of the form

$$\min_x f(x) - \mu \sum_{i=1}^{s} \ln E[c_i(x)],$$

which are solved for decreasing values of μ.

Example 2: maximum likelihood

We consider the log-likelihood over I mean probabilities (correspond to I individuals):

$$\min_{\theta} -\ln\left(E\left[\prod_{i=1}^{I} f(i;\boldsymbol{\theta};\boldsymbol{\xi}_i)\right]\right) := -LL(\theta).$$

Here,

$$g = -\ln .$$

If the probabilities are independent, we can rewrite the problem as

$$\min_{\theta} -\frac{1}{I}\sum_{i=1}^{I} \ln\left(E\left[f(i;\boldsymbol{\theta};\boldsymbol{\xi}_i)\right]\right).$$

Such a problem occurs for instance in discrete choice theory (more specifically, for mixed logit models estimation).

Sample average approximation (SAA)

Assume

- $\mathcal{X}$ is deterministic;
- smoothness and regularity assumptions;
- range(f) is compact.

Monte Carlo sample over ξ. With R random draws:

$$\hat{f}_R(x) := \frac{1}{R} \sum_{r=1}^{R} f(x, \xi_r).$$

The SAA problem is

$$\min_{x \in \mathcal{X}} \hat{g}(x) = g\left(\hat{f}_R(x)\right).$$

Consistency for the identity case

Consistency:

$$D\left(S^R, S^*\right) \to 0, \text{ a.s. when } R \to \infty.$$

The distance between approximate solutions and real solutions goes to infinity when the sample size goes to infinity.

Solutions?

- global minimizers;
- first-order critical points.

Not true for second-order critical points. But works well in practice.

Consistency: general case

First-order consistency can sill be proved, using similar arguments to those known in the litterature.

Moreover, if $g \in C^1$, we still can use the Delta theorem: if $\sqrt{R}(\boldsymbol{Y}_R - \boldsymbol{\mu}) \Rightarrow N(0, \Sigma_y)$ when $R \to \infty$, then we have the central limit theorem:

$$\sqrt{R}(g(\boldsymbol{Y}_R) - g(\boldsymbol{\mu}))/\sigma_g \Rightarrow N(0,1) \quad \text{quand } R \to \infty,$$

where ${\sigma_g}^2 = (\nabla g(\boldsymbol{\mu}))^T \Sigma_y \nabla g(\boldsymbol{\mu})$.

Simulation bias

For finite R,

$$E\left[g\left(\frac{1}{R}\sum_{r=1}^{R} f(x,\xi_r)\right)\right] \neq g(f(x)).$$

Let

$$B_R(\theta) = E\left[g\left(\frac{1}{R}\sum_{r=1}^{R} f(x,\xi_r)\right)\right] - g(f(x)).$$

denotes the bias of our estimator, when using R draws.

Simulation bias - statistical Taylor expansion

Assume also the f and g are in C^2. Let's introduce

$$h(x) = \hat{f}_R(x) - f(x),$$

For R large enough, the probability that $\hat{f}_R(x)$ is close to $f(x)$ is high. The (statistical) Taylor expansion gives us

$$g(\hat{f}_R(x)) = g(f(x)) + g'(f(x))h(x) + \frac{1}{2}g''(x)h^2(x) + O(h^3).$$

Since $E[h(x)] = 0$ and $E[h^2(x)] = \frac{1}{R}\mathrm{Var}[f(x,\xi)]$,

$$E[g(\hat{f}_R(x)] - g(f(x)) = \frac{1}{2}g''(f(x))\mathrm{Var}[f(x,\xi)] + O(E[h^3]).$$

Bias correction

This suggests the correction

$$\hat{B}_R(x) = \frac{1}{2} g''\left(\hat{f}_R(x)\right) \hat{\text{Var}}[f(x,\xi)],$$

as long as evaluating $g''(\cdot)$ is not too expensive and we can neglect the higher-order terms.

Idea: solve the modified optimization problem

$$\min_{x \in \mathcal{X}} g\left(\hat{f}_R(x)\right) - \hat{B}_R(x).$$

Issue: $\hat{B}_R(x)$ is itself a statistical estimator.

Application in log-likelihood estimation

In our application, we have

$$\hat{B}_R(\theta) = -E[SLL^R(\theta)] + LL(\theta) \approx \frac{1}{2IR} \sum_{i=1}^{I} \frac{\sigma^2_{ij_i}(\theta)}{\left(P_{ij_i}(\theta)\right)^2} \geq 0.$$

Note: one also has

$$\text{Var}[SLL^R(\theta)] = \frac{1}{I^2} \sum_{i=1}^{I} \sigma^2_{ij_i}(\theta).$$

Therefore

- variance is in $\mathcal{O}(1/(IR))$,
- bias is in $\mathcal{O}(1/R)$.

For large populations, the bias tends to dominate.

Application in log-likelihood estimation (cont'd)

The idea to correct the bias if such log-likehood estimation problems is not new...

- Similar ideas expressed in Gouriéroux and Monfort (1996), but using different arguments.
- Tsagkanos (2007) suggests using bootstrap bias estimate.
- More recently, Kristensen and Salanie (2010) make comparison between bootstrap, Taylor, and a new method base on Newton-Raphson. Practical recommendation: Taylor-based correction.

In practical experiments on mixed-logit models, Bastin and Cirillo (2010) obtain mitigated results. Why?

Evaluation of the bias correction

One therefore aims to solve the modified problem

$$\min_{x \in \mathcal{X}} g(\hat{f}_R(x)) - \hat{B}_R(x).$$

But. . .

1. the variance of the new objective function could increase, since

$$\text{Var}\left[g(\hat{f}_R(x)) - \hat{B}_R(x)\right]$$
$$= \text{Var}\left[g\left(\hat{f}_R(x)\right)\right] + \text{Var}\left[\hat{B}_R(x)\right] - \text{Cov}\left[(g\left(\hat{f}_R(x)), \hat{B}_R(x)\right)\right].$$

2. Usually, $E[\hat{B}_R(x)] \neq B_R(x)$ since
   1. one neglects high-order terms;
   2. most important, typically, $E[g''(f_R(x))] \neq g''(f(x))$.

Gains in terms of MSE?

Bias correction assessment

$$\text{Var}\left[\hat{B}_R(x)\right], \ \text{Cov}\left[g\left(\hat{f}_R(x)\right), \hat{B}_R(x))\right]?$$

No real theoretical clue here. We therefore turn to a more practical approach: bootstrap.

We consider the realisations $\xi_1, \ldots, \xi_R$. From them, we can construct the empirical distribution function $\hat{F}_R$ of ξ.

We then generate R draws from $\hat{F}_R$ of ξ, that is we produce R draws from $\{\xi_1, \ldots, \xi_R\}$ with replacement, in order to obtain the new sample

$$\{\xi_1^b, \ldots, \xi_R^b\},$$

and calculate

$$\hat{B}_{b,R}(x, \xi_1^b, \ldots, \xi_R^b) =: \hat{B}_{b,R}(x).$$

Bias properties

We take q bootstrap samples at x^*, delivering m values

$$\hat{B}_{b_1,R}(x^*), \ldots, \hat{B}_{b_q,R}(x^*)$$

The variance of the bias estimation can be estimated as

$$\hat{\mathrm{Var}}\left[\hat{B}_{b,R}\right],$$

and its own bias, as

$$E_{\hat{F}}\left[\hat{B}_{b,R}(x^*)\right] - \hat{B}_{b,R}(x^*).$$

Note: existence of an improved bootstrap bias estimator.

(Bastin and Cirillo, 2010) 674 individuals, 4089 obs.

Bootstrap analysis at solution with uncorrected log-likelihood.

Nb of draws Corr.	500 std	500 corr.	1000 std.	1000 corr.	2000 std.	2000 corr.
Mean	-3.3186	-3.3078	-3.2964	-3.2903	-3.2830	-3.2787
Std. dev.	0.0066	0.0066	0.0060	0.0061	0.0047	0.0048
Boot. bias	-0.0139	-0.0031	-0.0088	-0.0027	-0.0056	-0.0018
Imp. bias	-0.0134	-0.0026	-0.0088	-0.0026	-0.0054	-0.0017

Bootstrap analysis at solution with corrected log-likelihood.

Nb of draws Corr.	500 std	500 corr.	1000 std.	1000 corr.	2000 std.	2000 corr.
Mean	-3.3173	-3.3060	-3.2968	-3.2905	-3.2886	-3.2849
Std. dev.	0.0079	0.0080	0.0061	0.0062	0.0046	0.0048
Boot. bias	-0.0166	-0.0052	-0.0091	-0.0027	-0.0056	-0.0019
Imp. bias	-0.0160	-0.0045	-0.0090	-0.0027	-0.0054	-0.0017

Residual bias is significantly smaller.

Taylor-based bias estimator properties

Nb of draws Corr.	500 std	500 corr.	1000 std.	1000 corr.	2000 std.	2000 corr.
Mean	-0.01080	-0.01142	-0.00616	-0.00632	-0.00372	-0.00372
Std. dev.	0.00049	0.00052	0.00036	0.00037	0.00032	0.00032
Bp bias	0.00095	0.00126	0.00065	0.00068	0.00058	0.00058
Imp. bias	0.00097	0.00122	0.00066	0.00070	0.00058	0.00058

Observations:

1. the bias estimator has a small variance, so it impacts the total variance only marginally;
2. its own bias is small compared to its magnitude.

It is therefore useful in this context.

Enforcing a positive correlation

One would prefer $\text{Cov}[(g(\hat{f}_R(x)), \hat{B}_R(x))]$ to be positive.

$$\hat{B}_R(x) = \frac{1}{2} g''\left(\hat{f}_R(x)\right) \hat{\text{Var}}[f(x,\xi)],$$

Cheap to evaluate when $g''\left(\hat{f}_R(x)\right)$ is easy to compute. Reuse of already computed elements usually implies use of common random numbers: one generates the draws from the same uniforms U_r, $r = 1, \ldots, R$.

If the correlation is negative, one can try antithetic variates, using $1 - U_r$, $r = 1, \ldots, R$ for $B^R(x)$), but can make the reuse of previously computed elements less direct.

Enforcing a positive correlation (cont'd)

Other dataset: 274 individuals, 2466 observations.

Covariance values (200 evaluations at the solution):

- common random variables: $-1.89e^{-7}$.
- antithetics: $5.28e^{-9}$.

We found the desired sign,... but

1. the computational cost for antithetics is twice that for common random numbers;
2. the covariance is too small to play a significant effect.

Optimization bias

It is well-known that

$$E\left[\min_{x\in\mathcal{X}} \hat{f}_R(x)\right] \leq \min_{x\in\mathcal{X}} E\left[\hat{f}_R(x)\right] = f(x).$$

Similarly

$$E\left[\min_{x\in\mathcal{X}} g(\hat{f}_R(x))\right] \leq \min_{x\in\mathcal{X}} E\left[g(\hat{f}_R(x))\right] = \min_{x\in\mathcal{X}} \left[g(f(x)) + B_R(x)\right],$$

and

$$\begin{aligned} E\left[\min_{x\in\mathcal{X}} \left(g(\hat{f}_R(x)) - B_R(x)\right)\right] &\leq \min_{x\in\mathcal{X}} E\left[\left(g(\hat{f}_R(x)) - B_R(x)\right)\right] \\ &= \min_{x\in\mathcal{X}} g(f(x)). \end{aligned}$$

Observations

Assume that $\hat{B}_R(x) = B_R(x)$, and keeps the same sign in the vicinity of the point of interest (typically a local solution). We assume this solution to be global in $\mathcal{X}$ (or restrict $\mathcal{X}$).

- Removing the simulation bias does not eliminate optimization bias.
- A negative simulation bias will amplify the optimization bias, if not corrected; a positive simulation bias will play against the optimization bias.
- Difficult to estimate the optimization bias.
- Both biases change at different rates with the number of draws;
- Increasing the number of draws typically reduces the bias contribution in the MSE faster than the variance, both of them being in $\mathcal{O}(1/R)$.

How to evaluate the benefit of the correction?

Application in discrete choice theory (again)

1041 observations delivered by 173 individuals.

Draws Method	500 standard	500 corrected	1000 standard	1000 corrected	2000 standard	2000 corrected
Mean	-0.01080	-0.01142	-0.00616	-0.00632	-0.00372	-0.00372
Std. Dev.	0.00049	0.00052	0.00036	0.00037	0.00032	0.00032
Boot. bias	0.00097	0.00122	0.00066	0.00070	0.00058	0.00058

Table: Properties of bias estimator, 500 bootstrap replications.

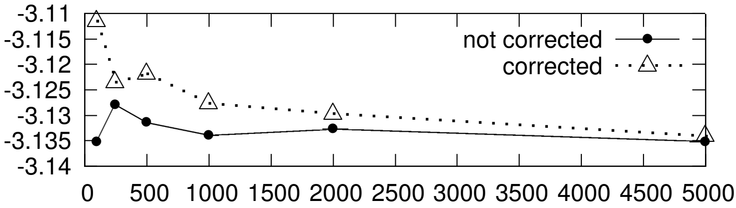

Figure: Evolution of the log-likelihood optimal value

A more general picture

Since two bias sources interact, how to evaluate the real interest of the correction?

Again, bootstrap helps to have a more general picture.

First obtained empirical observations (Bastin and Cirillo, 2011): the optimisation bias is really the key in some applications.

Mixed logit example (Bastin and Cirillo, 2011)

Application to panel data collected at Baltimore Washington International Airport

We also used lattice rules, following Munger, L'Ecuyer, Bastin, Cirillo, and Tuffin (2011).

Par.	LL estimates			Bootstrap mean			Bootstrap bias		
	MC nc	MC wc	Lattice	MC nc	MC wc	Lattice	MC nc	MC wc	Lattice
Wait time 10	-0.619	-0.630	-0.615	-0.614	-0.618	-0.618	-0.004	-0.012	-0.003
Wait time 15	-1.010	-1.022	-1.017	-1.019	-1.025	-1.025	-0.009	-0.003	-0.007
Wait time 20	-1.732	-1.754	-1.740	-1.744	-1.756	-1.753	-0.011	-0.002	-0.013
Cost (m)	-1.993	-1.980	-1.997	-2.008	-1.997	-1.999	-0.015	-0.017	-0.002
Cost (sd)	1.797	1.800	1.807	1.815	1.812	1.810	0.018	0.011	0.004
Pass dropp (m)	1.680	1.702	1.708	1.719	1.734	1.728	0.040	0.032	0.020
Pass dropp (sd)	1.589	1.612	1.607	1.563	1.579	1.575	-0.026	-0.033	-0.032
Auto cyb (m)	-0.226	-0.230	-0.227	-0.213	-0.218	-0.216	0.013	0.012	0.012
Auto cyb (sd)	1.200	1.226	1.176	1.175	1.195	1.190	-0.026	-0.032	0.014
Human cyb (m)	0.129	0.130	0.139	0.155	0.157	0.155	0.026	0.027	0.017
Human cyb (sd)	0.721	0.731	0.744	0.652	0.654	0.659	-0.070	-0.076	-0.086
Guided way (m)	-0.099	-0.099	-0.095	-0.133	-0.134	-0.133	-0.033	-0.032	-0.037
Guided way (sd)	1.029	1.050	1.051	1.035	1.050	1.046	0.007	0.023	-0.005
LL	-4.418	-4.414	-4.409	-4.368	-4.366	-4.367	0.050	0.048	0.042

Using RQMC or correcting the simulation bias do not give a big improvement.

Conclusions

Nonlinear stochastic programming:

- Delta theorem ensures consistency under some regularity conditions;
- for finite sample sizes, there is often a simulation bias (different than optimization bias);
- statistical Taylor expansion allows to estimate this bias;
- not without drawbacks:
  - this estimator is typically biased too;
  - it can result in an increase of the variance;
- optimization bias and simulation can counteract.

Conclusions: maximum likelihood

- Lot of efforts put to correct inner simulation bias.
- In our experiments, Taylor-based bias estimator worked well, and has very limited impact on variance.
- It seems that we lost the big picture: bias involved by population sampling.
- Data are costly to obtain, and efforts in the literature to justify small populations. Really a good idea?
- More efforts needed on this level.

References

Fabian Bastin and Cinzia Cirillo.
Reducing simulation bias in mixed logit model estimation.
Journal of Choice Modelling, 3(2):71–88, 2010.

Fabian Bastin, Cinzia Cirillo, and Philippe L. Toint.
Formulation and solution strategies for nonparametric nonlinear stochastic programs, with an application in finance.
Optimization, 59(3):355–376, 2010.

Bradley Efron and Robert J. Tibshirani.
An Introduction to the Bootstrap.
Number 57 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, Boca Raton, Florida, USA, 1993.

Christian Gouriéroux and Alain Monfort.
Simulation-based Econometric Methods.
Oxford University Press, Oxford, United Kingdom, 1996.

Athanasios G. Tsagkanos.
A bootstrap-based minimum bias maximum simulated likelihood estimator of mixed logit.
Economic Letters, 96(2):282–286, 2007.