Global stochastic optimization with low-dispersion point sets

Author: Sidney Yakowitz, Pierre L’Ecuyer, Felisa Vazquez-Abad.
Presenter: Li Jinzhi

Problem statement

- Goal: $\min_{\theta \in \Theta} J(\theta) \triangleq E_{P_\theta}(L(\theta, \varpi))$

  where ∵ is a **compact** region in the **s-dimensional** real space

- Discretization: $S_m \triangleq \{\theta_{m,1}, \theta_{m,2}, \ldots, \theta_{m,m}\} \subset \Theta$

- Budget: total N; for each discrete point $r = \left\lfloor \frac{N}{m} \right\rfloor$.

- Methods: $\widehat{\theta^*} = \arg\min_{\theta_{m,i} \in S_m} \widehat{J}(\theta_{m,i})$

$$\widehat{J}(\theta_{m,i}) = \frac{1}{r} \sum_{j=1}^{r} L_{i,j}$$

Contents

01 **How to balance between m and r**

How to choose S_m

How to allocate budgets adaptively

PART 1

How to balance between *m* and *r*

So, this choice of ε makes the selection error decrease at the same rate as the discretization error, and therefore gives a good tradeoff in rates for the sum of these two errors.

1.1 Basic idea

- Error: $$\Delta_N = \widehat{J}(\widehat{\theta_N^*}) - J^* = (\widehat{J}(\widehat{\theta_N^*}) - J_N^*) + (J_N^* - J^*)$$ $$J_N^* = \min_{1 \le i \le m} J(\theta_{m,i})$$

- The first term: from random sampling

  ——control the tail of $L(\theta, \omega)$! (Assumption A1)

- The second term: from discretization

  ——avoid S_m being too sparse (low-dispersion) (Assumption A3)

  ——avoid J_{θ} being too bumpy locally around θ^* (Assumption A2)

1.2 Discretization error

- Dispersion: $d_p(S_m, \because) = \sup_{\theta \in \because} \min_{1 \le i \le m} \left\| \theta - \theta_{m,i} \right\|_p$

- Smoothness locally around θ^* :

$$H_p(t) = \sup_{\theta \in B_p(\theta^*, t) \cap \because} (J(\theta) - J(\theta^*))$$

$$B_p(\theta, t) = \{x \in \mathbb{R}^s \mid \left\| x - \theta \right\|_p \le t\}$$

- Proposition 1:

$$J_N^\star - J^\star \le H_p(d_p(S_m, \because))$$

1.2 Discretization error

- Assumption A2:

$$H_p(t) \le K_1 t^q \text{ for } t \le t_0$$

- Assumption A3:

$$d_p(S_m, \because) \le \frac{K_2}{\lfloor m^{1/s} \rfloor}$$

1.3 Estimation error

- Assumption A1:

$$P[\left|\widehat{J}(\theta)-J(\theta)\right|>\varepsilon]\leq e^{-r\kappa\varepsilon^2}\ \text{for}\ r\geq R, 0<\varepsilon<\varepsilon_1$$

- Ellis (1998.p.247): if the moment generating function is finite for all real values, then:

$$P[\left|\widehat{J}(\theta)-J(\theta)\right|>\varepsilon]\leq e^{-\frac{r\varepsilon^2}{2\sigma_\theta^2}+O(r\varepsilon^3)}$$

1.3 Estimation error

- Assumption A1:

$$P[\left|\widehat{J}(\theta) - J(\theta)\right| > \varepsilon] \le e^{-r\kappa\varepsilon^2} \text{ for } r \ge R, 0 < \varepsilon < \varepsilon_1$$

- Proposition 2:

$$P[\left|\widehat{J}(\widehat{\theta_N^*}) - J_N^*\right| > 2\varepsilon] \le me^{-r\kappa\varepsilon^2} \text{ for } N \ge N_0, 0 < \varepsilon \le \frac{\varepsilon_1}{2}$$

1.3 Estimation error

- Assumption A1:

$$P[\left|\widehat{J}(\theta) - J(\theta)\right| > \varepsilon] \leq e^{-r\kappa\varepsilon^2} \text{ for } r \geq R, 0 < \varepsilon < \varepsilon_1$$

- Proposition 2:

$$P[\left|\widehat{J}(\widehat{\theta_N^*}) - J_N^*\right| > 2\varepsilon] \leq me^{-r\kappa\varepsilon^2} \text{ for } N \geq N_0, 0 < \varepsilon \leq \frac{\varepsilon_1}{2}$$

- Corollary 1:

$$P[\Delta_N > 2\varepsilon + H_p(d_p(S_m, \because))] \leq me^{-r\kappa\varepsilon^2} \text{ for } N \geq N_0, 0 < \varepsilon \leq \frac{\varepsilon_1}{2}$$

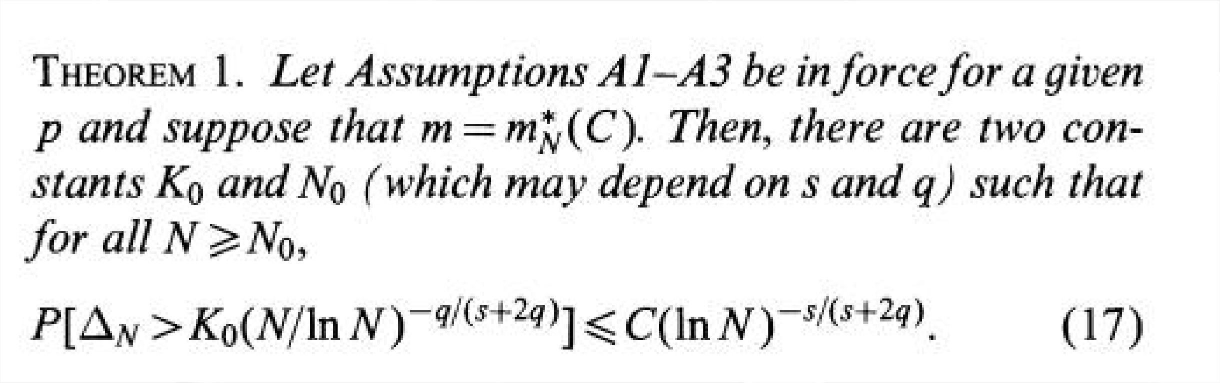
Proposition 1:

$$J_N^* - J^* \leq H_p(d_p(S_m, \because))$$

1.4 Balance discretization and estimation error

- Key idea: make them decrease at the same rate

$$m_N^*(C) \sim C \cdot \left(\frac{N}{\ln N} \right)^{\frac{s}{s+2q}}$$



THEOREM 1. *Let Assumptions A1–A3 be in force for a given p and suppose that $m = m_N^*(C)$. Then, there are two constants K_0 and N_0 (which may depend on s and q) such that for all $N \geqslant N_0$,*

$$P[\Delta_N > K_0(N/\ln N)^{-q/(s+2q)}] \leqslant C(\ln N)^{-s/(s+2q)}. \tag{17}$$

PART 2

How to choose S_m

Niederreiter (1992, Theorem 6.9) gives the following low-dispersion sequence for the sup norm.

2.1 Low dispersion sequence

- Niederreiter (1992, Theorem 6.9): for the sup norm, for $\because = [0,1]^s$:

$$x_m = \begin{cases} 1 & m = 1 \\ (\log_2(2m-3)) \bmod 1 & m \geq 2 \end{cases}$$

$$\lim_{m\to\infty} m^{1/s} d_\infty(S_m, [0,1]^s) = \frac{1}{2\ln 2}$$

  - s=1: asymptotically optimal
  - s>1: asymptotically $\frac{1}{2ln2}$, while the smallest possible value cannot be smaller than 1/2

    “one cannot achieve much better with the sup norm”

PART 3

How to allocate budgets adaptively

At promising points, one should collect more observations because it is with nearly optimal points that sampling noise is more likely to lead to selection error.

3.1 Allocate budgets adaptively

- Key idea: majorization minimization

——control the upper bound of the estimation error

3.2 Majorization

- Key idea: majorization minimization

——control the upper bound of the estimation error

- Update the proposition 2:

Proposition 2:

$$P[\left|\widehat{J}(\widehat{\theta_N^*}) - J_N^*\right| > 2\varepsilon] \le m e^{-r\kappa\varepsilon^2} \text{ for } N \ge N_0, 0 < \varepsilon \le \frac{\varepsilon_1}{2}$$

- Proposition 4:

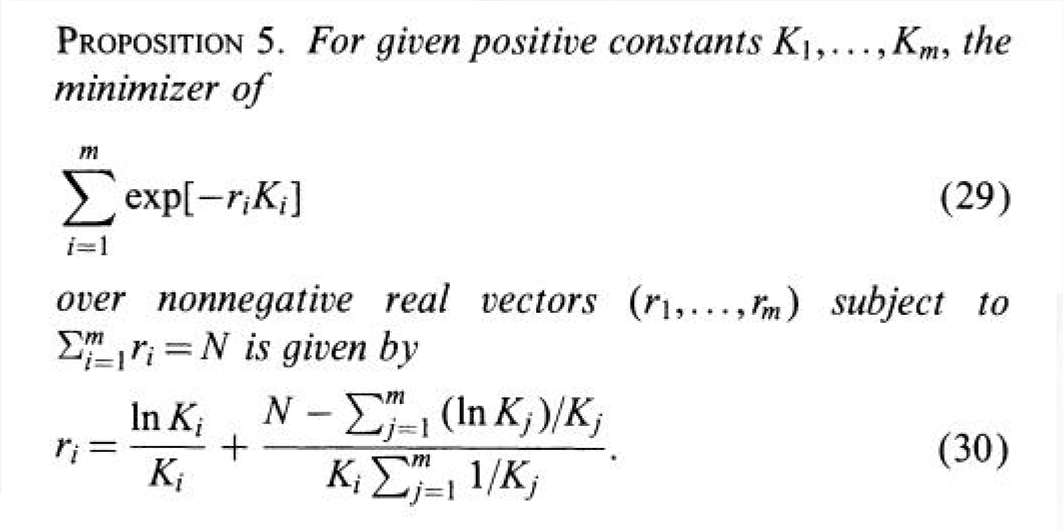
$$P[\widehat{J}(\widehat{\theta_N^*}) - J_N^* > \varepsilon] \le \sum_{i=1}^{m} e^{-\frac{r\kappa(\delta_i+\varepsilon)^2}{16}} \text{ for } N \ge N_0, 0 < \varepsilon \le \frac{\varepsilon_1}{2}$$

3.3 Minimization

- Key idea: majorization minimization

  ——control the upper bound of the estimation error

- Proposition 5:



PROPOSITION 5. *For given positive constants $K_1, \ldots, K_m$, the minimizer of*

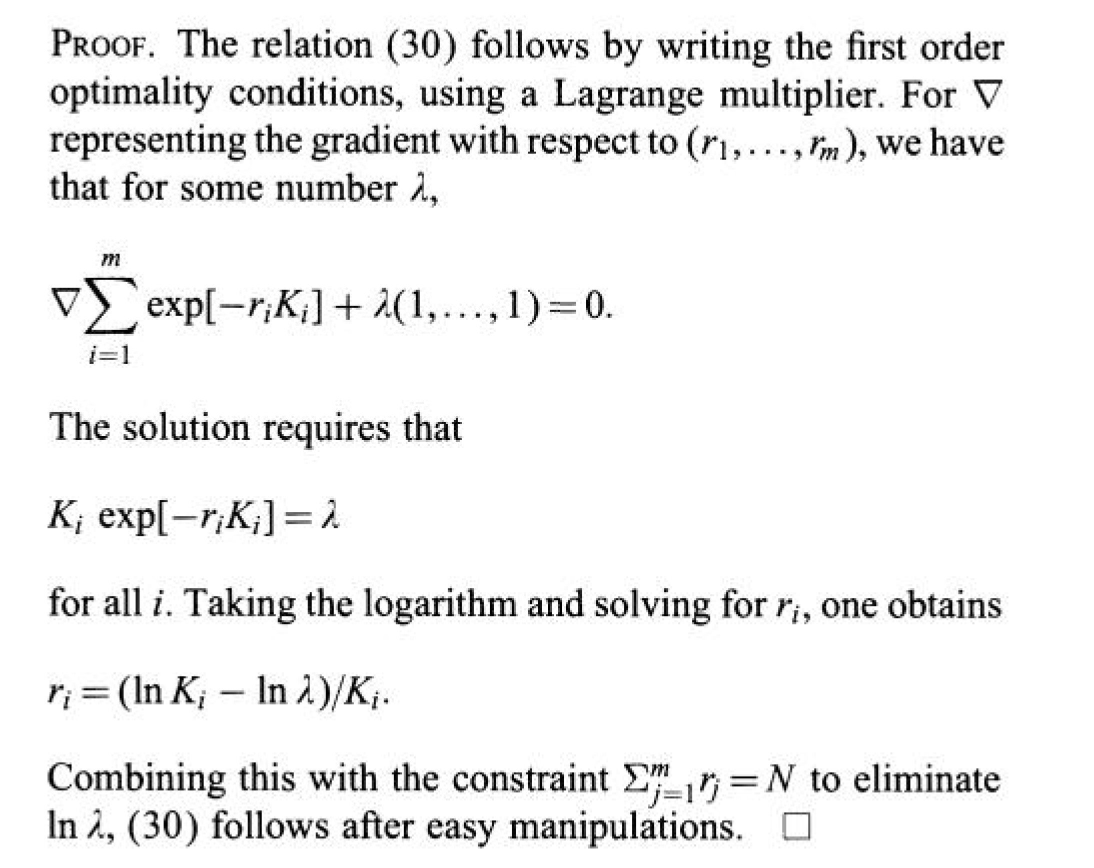
$$\sum_{i=1}^{m} \exp[-r_i K_i] \tag{29}$$

over nonnegative real vectors $(r_1, \ldots, r_m)$ subject to $\Sigma_{i=1}^{m} r_i = N$ is given by

$$r_i = \frac{\ln K_i}{K_i} + \frac{N - \sum_{j=1}^{m} (\ln K_j)/K_j}{K_i \sum_{j=1}^{m} 1/K_j}. \tag{30}$$

3.3 Minimization

- Proof of Proposition 5:



PROOF. The relation (30) follows by writing the first order optimality conditions, using a Lagrange multiplier. For ∇ representing the gradient with respect to $(r_1, \ldots, r_m)$, we have that for some number λ,

$$\nabla \sum_{i=1}^{m} \exp[-r_i K_i] + \lambda(1, \ldots, 1) = 0.$$

The solution requires that

$$K_i \ \exp[-r_i K_i] = \lambda$$

for all i. Taking the logarithm and solving for r_i, one obtains

$$r_i = (\ln K_i - \ln \lambda)/K_i.$$

Combining this with the constraint $\Sigma_{j=1}^{m} r_j = N$ to eliminate $\ln \lambda$, (30) follows after easy manipulations. □

3.4 Allocation strategy

- Choose the minimal:

$$r_i\widehat{K_i}-\ln\widehat{K_i}=r_i\frac{(\widehat{\delta_i}+\varepsilon)^2\kappa}{16}-\ln\frac{(\widehat{\delta_i}+\varepsilon)^2\kappa}{16}$$

- Implementation:

$$\arg\min_i r_i(\widehat{\delta_i}+\varepsilon)^2$$

THANKS!