Variable-Number Sample-Path Optimization

Geng Deng · Michael C. Ferris

Presented by Tan Wang

December 13, 2021

Sample-path method

In this paper, we consider the following unconstrained stochastic optimization problem:

$$\min_{x \in \mathbb{R}^n} f(x) = \mathbb{E}[F(x, \xi(\omega))]$$

- The sample response function F takes two inputs, the simulation parameters $x \in \mathbb{R}^n$ and a random sample of $\xi(\omega)$ in $\mathbb{R}^d$.
- $f(x)$ is well defined.
- The solution is x^*.

Sample-path method

- The sample-path method is sometimes called the Monte Carlo sampling approach or the sample average approximation method.
- The basic idea of the method is to approximate the expected value function $f(x)$ by averaging sample response functions.

$$f(x) \approx \hat{f}^N(x) := \frac{1}{N}\sum_{i=1}^{N} F(x, \xi_i)$$

- The averaged sample-path problem

$$\min_{x \in \mathbb{R}^n} \hat{f}^N(x)$$

serves as a substitute for the original problem.

Sample-path method

- An optimal solution $x^{*,N}$ is then treated as an approximation of x^*.
- Under the assumption that the sequence of functions $\{\hat{f}^N\}$ epiconverges to the function f, the optimal solution sequence $\{x^{*,N}\}$ converges to x^* almost surely for all sample paths.

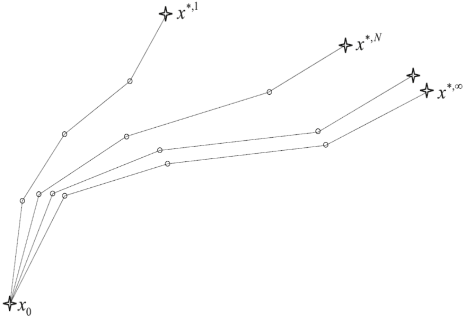

Fig. 1 Mechanism of the sample-path optimization method. Starting from x_0, for a given N, a deterministic algorithm is applied to solve the sample-path problem. The sequence of solutions $\{x^{*,N}\}$ converges to the true solution $x^{*,\infty} = x^*$ almost surely

Variable-number sample-path scheme

- The new variable-number sample-path (VNSP) scheme is designed to generate different numbers of samples (N_k) at each iteration.
- The VNSP scheme integrates Bayesian techniques to determine N_k.

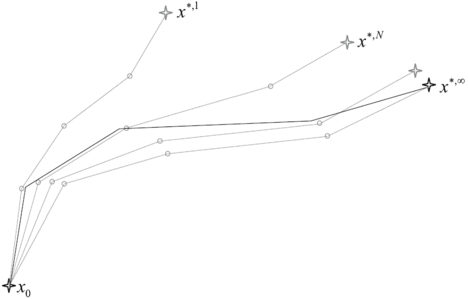

Fig. 2 Mechanism of the new sample-path method with the VNSP scheme. Starting from x_0, the algorithm generates its iterates across different averaged sample functions. In an intermediate iteration k, it first computes a satisfactory N_k which guarantees certain level of accuracy, then an optimization step is taken exactly the same as in problem (3), with $N = N_k$. The algorithm has a globally convergent solution x^{*,N_∞}, where $N_\infty := \lim_{k\to\infty} N_k$. The convergence is almost sure for all the sample paths, which correspond to different runs of the algorithm. The solution, we will prove later, matches the solution $x^{*,\infty}$

Variable-number sample-path scheme

- We require $x_k \to x^*$ almost surely, but we do not impose the convergence condition $\hat{f}^{N_k} \to f$. As a consequence, $\{N_k\}$ is a non-decreasing sequence with the limit value N_∞ being either finite or infinite.

- A toy example:

$$F(x, \xi(\omega)) = \phi(x) + \xi(\omega)$$

- $\phi(x)$ is a deterministic function and $\xi(\omega) \sim N\left(0, \sigma^2\right)$.
- The solutions of $\hat{f}^k$ are: $x^{*,1} = x^{*,2} = \cdots = x^{*,\infty}$.
- In this case, the VNSP scheme turns out to use a constant sequence of sample numbers $N_k : N_1 = N_2 = \cdots = N_\infty < +\infty$.
- We obtain $\lim_{k\to\infty} x_k = x^{*,N_1} = \cdots = x^{*,N_\infty} = x^*$, but obviously $\lim_{k\to\infty} \hat{f}^{N_k} \neq f$.

The UOBYQA algorithm

We apply Powell's Unconstrained Optimization BY Quadratic Approximation (UOBYQA) algorithm as our base sample-path optimization solver.

- The algorithm is a derivative-free approach.
- The general structure of UOBYQA follows a model-based approach, which constructs a chain of local quadratic models that approximate the objective function.
- The method is an iterative algorithm in a trust region framework.
- It differs from a classical trust region method in that it creates quadratic models by interpolating a set of sample points instead of using the gradient and Hessian values of the objective function.

The UOBYQA algorithm

Basic assumptions regarding the nature of the objective function:

Assumption 1

For a fixed $y \in \mathbb{R}^d$ the function $F(\cdot, y)$ is twice continuously differentiable and its gradient and Hessian are uniformly bounded on $\mathbb{R}^n \times \mathbb{R}^d$. There exist constants $\kappa_{Fg} > 0$ and $\kappa_{Fg} > 0$, such that the following inequalities hold:

$$\sup_{x\in\mathbb{R}^n, y\in\mathbb{R}^d} \left\| \frac{\partial F(x,y)}{\partial x} \right\| \leq \kappa_{Fg} \text{ and } \sup_{x\in\mathbb{R}^n, y\in\mathbb{R}^d} \left\| \frac{\partial^2 F(x,y)}{\partial^2 x} \right\| \leq \kappa_{Fh}$$

Assumption 2

For a fixed $y \in \mathbb{R}^d$, the function $F(\cdot, y)$ and the underlying function $f(\cdot)$ are bounded below on $\mathbb{R}^n$.

Interpolating quadratic model properties

At every iteration of the algorithm, a quadratic model

$$Q_k^N(x) = c_k^N + \left(g_k^N\right)^T (x - x_k) + \frac{1}{2}(x - x_k)^T G_k^N (x - x_k),$$

is constructed by interpolating a set of adequate points $\mathcal{I}_k = \left\{y^1, y^2, \ldots, y^L\right\}$,

$$Q_k^N\left(y^i\right) = \hat{f}^N\left(y^i\right), \quad i = 1, 2, \ldots, L$$

- The coefficient c_k^N is a scalar, g_k^N is a vector in $\mathbb{R}^n$, and G_k^N is an $n \times n$ real symmetric matrix.
- To ensure a unique quadratic interpolator, the number of interpolating points should satisfy

$$L = \frac{1}{2}(n+1)(n+2)$$

Interpolating quadratic model properties

For each quadratic interpolation model, we require that the Hessian matrix is uniformly bounded.

Assumption 3

The Hessian of the quadratic function Q_k^N is uniformly bounded for all x in the trust region, i.e., there exists a constant $\kappa_{Qh} > 0$ such that

$$\left\| G_k^N \right\| \leq \kappa_{Qh}, \quad \text{for all } x \in \{x \in \mathbb{R}^n \mid \|x - x_k\| \leq \Delta_k\}$$

Interpolating quadratic model properties

The error of the approximation:

Lemma 1

Suppose Assumptions 1-3 hold and $\mathcal{I}_k$ is adequate in the trust region $\mathcal{B}_k(\Delta_k)$. Furthermore, if at iteration k, Q_k^N is the interpolative approximation model for the function $\hat{f}^N$, then assume the bias of the function value and the gradient are bounded within the trust region. Then there exist constants κ_{em} and κ_{eg} , for each $x \in \mathcal{B}_k(\Delta_k)$, the following inequalities hold

$$\left|\hat{f}^N(x) - Q_k^N(x)\right| \leq \kappa_{em} \max\left[\Delta_k^2, \Delta_k^3\right]$$

$$\left\|\nabla \hat{f}^N(x) - g_k^N\right\| \leq \kappa_{eg} \max\left[\Delta_k, \Delta_k^2\right]$$

Within a small trust region, the model Q_k^N is also a decent approximation model.

Interpolating quadratic model properties

We have seen that Q_k^N interpolates the function f^N at the points in $\mathcal{I}_k$. Let Q_k^∞ be the 'expected' quadratic model interpolating the function f at the same points. The following lemma provides convergence of Q_k^N to Q_k^∞.

Lemma 2

$Q_k^N(x)$ converges pointwise to $Q_k^\infty(x)$ with probability 1(w.p.1) as $N \to \infty$

- The law of large numbers (LLN) guarantees the pointwise convergence of $\hat{f}^N(x)$ to $f(x)$ w.p.1.
- By solving the system of linear equations, each component of the coefficients of Q_k^N, c_k^N, $g_k^N(i)$, $G_k^N(i,j)$, $i,j = 1,2,\ldots,n$, is uniquely expressed as a linear combination of $\hat{f}^N\left(y^i\right)$, $\hat{f}^N\left(y^i\right)\hat{f}^N\left(y^j\right)$, $i,j = 1,2,\ldots,L$.
- Therefore, as $N \to \infty$ the coefficients c_k^N, g_k^N, G_k^N converge to c_k^∞, g_k^∞, G_k^∞ w.p.1.
- Finally, for a fixed value $x \in \mathbb{R}^n$, $Q_k^N(x)$ converges to $Q_k^\infty(x)$ w.p.1.

The core algorithm

Starting the algorithm requires an initial trial point x_0 and an initial trust region radius Δ_0. As in a classical trust region method, a new promising point is determined from a subproblem:

$$\min_{s \in \mathbb{R}^n} Q_k^N(x_k + s), \quad \text{subject to } \|s\| \leq \Delta_k$$

The new solution $s^{*,N}$ is accepted (or not) by evaluating the 'degree of agreement' between $\hat{f}^N$ and Q_k^N:

$$\rho_k^N = \frac{\hat{f}^N(x_k) - \hat{f}^N\left(x_k + s^{*,N}\right)}{Q_k^N(x_k) - Q_k^N\left(x_k + s^{*,N}\right)}$$

If the ratio ρ_k^N is large enough, which indicates a good agreement between the quadratic model Q_k^N and the function $\hat{f}^N$, the point $x_k + s^{*,N}$ is accepted into the set $\mathcal{I}_k$.

The core algorithm

We introduce the following lemma concerning the ‘sufficient reduction’ within a trust region step.

Lemma 3

The solution $s_k^{*,N}$ of the subproblem satisfies

$$Q_k^N(x_k) - Q_k^N\left(x_k + s^{*,N}\right) \geq \kappa_{mdc} \left\| g_k^N \right\| \min\left[\frac{\left\| g_k^N \right\|}{\kappa_{Qh}}, \Delta_k \right]$$

for some constant $\kappa_{mdc} \in (0,1)$ independent of k.

This is an important but standard result in the trust region literature.

The core algorithm

Choose a starting point x_0, an initial trust region radius Δ_0 and a termination trust region radius Δ_{end}.

- Generate initial trial points in the interpolation set $\mathcal{I}_k$. Determine the first iterate $x_1 \in \mathcal{I}_k$ as the best point in $\mathcal{I}_k$.
- For iterations $k = 1, 2, \ldots$
    - Determine N_k via the VNSP scheme in Sect. 2.3.
    - Construct a quadratic model $Q_k^{N_k}$ of the form (4) which interpolates points in $\mathcal{I}_k$. If $\left\| g_k^{N_k} \right\| \leq \epsilon_1$ and $\mathcal{I}_k$ is inadequate in $\mathcal{B}_k \left(\epsilon_2 \left\| g_k^{N_k} \right\| \right)$, then improve the quality of $\mathcal{I}_k$.
    - Solve the trust region subproblem (20). Evaluate $\hat{f}^{N_k}$ at the new point $x_k + s^{*,N_k}$ and compute the agreement ratio $\rho_k^{N_k}$ in (21).
    - If $\rho_k^{N_k} \geq \eta_1$, then insert $x_k + s^{*,N_k}$ into $\mathcal{I}_k$. If a point is added to the set $\mathcal{I}_k$, another element in $\mathcal{I}_k$ should be removed to maintain the cardinality $|\mathcal{I}_k| = L$. If $\rho_k^{N_k} < \eta_1$ and $\mathcal{I}_k$ is inadequate in $\mathcal{B}_k$, improve the quality of $\mathcal{I}_k$.

The core algorithm

- For iterations $k = 1, 2, \ldots$
  - Update the trust region radius Δ_k :

$$\Delta_{k+1} \begin{cases} \in [\Delta_k, \gamma_2 \Delta_k], & \text{if } \rho_k^{N_k} \geq \eta_1 \\ \in [\gamma_0 \Delta_k, \gamma_1 \Delta_k], & \text{if } \rho_k^{N_k} < \eta_1 \text{ and } \mathcal{I}_k \text{ is adequate in } \mathcal{B}_k(\Delta_k) \\ = \Delta_k, & \text{otherwise.} \end{cases}$$

  - When a new point $x^+$is added into $\mathcal{I}_k$, if

$$\hat{\rho}_k^{N_k} = \frac{\hat{f}^{N_k}(x_k) - \hat{f}^{N_k}(x^+)}{Q_k^{N_k}(x_k) - Q_k^{N_k}(x_k + s^{*,N_k})} \geq \eta_0$$

    then $x_{k+1} = x^+$, otherwise, $x_{k+1} = x_k$.

  - Check whether any of the termination criteria is satisfied, otherwise repeat the loop. The termination criteria include $\Delta_k \leq \Delta_{\text{end}}$ and hitting the maximum limit of function evaluations.

Bayesian VNSP scheme

- The goal of a VNSP scheme is to determine the suitable sample number N_k to be applied at iteration k.
- In our algorithm, $Q_k^N(x_k) - Q_k^N\left(x_k + s^{*,N}\right)$ is the observed model reduction.
- The key idea for the global convergence of algorithm is that, by replacing g_k^N with g_k^∞ in (22), we force the model reduction $Q_k^N(x_k) - Q_k^N\left(x_k + s^{*,N}\right)$ to regulate the size of $\|g_k^\infty\|$, and so drive $\|g_k^\infty\|$ to zero.
- We present the modified 'sufficient reduction' criterion:

$$Q_k^N(x_k) - Q_k^N\left(x_k + s^{*,N}\right) \geq \kappa_{mdc} \|g_k^\infty\| \min\left[\frac{\|g_k^\infty\|}{\kappa_{Qh}}, \Delta_k\right]$$

Bayesian VNSP scheme

To ensure the 'sufficient reduction' criterion (27) is satisfied accurately, we require

$$\Pr\left(E_k^N\right) = \Pr\left(Q_k^N\left(x_k\right) - Q_k^N\left(x_k + s^{*,N}\right) < \kappa_{mdc}\left\|g_k^\infty\right\| \min\left[\frac{\left\|g_k^\infty\right\|}{\kappa_{Qh}}, \Delta_k\right]\right)$$
$$\leq \alpha_k$$

- The event E_k^N is defined as the failure of (27) for the current N
- α_k is the significance level.
- In practice, the risk $\Pr\left(E_k^N\right)$ is difficult to evaluate.
- By adapting knowledge from Bayesian inference, we approximate the risk value by a Bayesian posterior estimation based on the current observations X^N

$$\Pr\left(E_k^N\right) \approx \Pr\left(E_k^N \mid X^N\right)$$

Bayesian VNSP scheme

- The value $\Pr\left(E_k^N \mid X^N\right)$ is thus called Bayes risk.
- α_k is the significance level.
- Given the observations X^N, we have

$$
\begin{aligned}
&\Pr\left(E_k^N \mid X^N\right) \\
&\quad = \Pr\left(Q_k^N\left(x_k\right) - Q_k^N\left(x_k + s^{*,N}\right) < \kappa_{mdc}\left\|g_k^\infty\right\| \min\left[\frac{\left\|g_k^\infty\right\|}{\kappa Qh}, \Delta_k\right] \mid X^N\right) \\
&\quad = \Pr\left(Q_k^N\left(x_k\right) - Q_k^N\left(x_k + s^{*,N}\right) < \kappa_{mdc}\left\|g_k^\infty \mid X^N\right\| \min\left[\frac{\left\|g_k^\infty \mid X^N\right\|}{\kappa_{Qh}}, \Delta_k\right]\right)
\end{aligned}
$$

- The left-hand side $Q_k^N\left(x_k\right) - Q_k^N\left(x_k + s^{*,N}\right)$ of the inequality becomes a fixed quantity given X^N. The probability evaluation is computed with respect to the posterior distribution $g_k^\infty \mid X^N$.

Bayesian VNSP scheme

Lemma 4

The Bayes risk $\Pr\left(E_k^N \mid X^N\right)$ converges to zero as $N \to \infty$.

Lemma 4 guarantees that $\Pr\left(E_k^N \mid X^N\right) \leq \alpha_k$ will eventually be satisfied when N is large enough.

Bayesian VNSP scheme

- The exact evaluation of the probability is hard to compute, especially involving the component $\kappa_{mdc} \left\| g_k^\infty \mid X^N \right\| \min \left[\frac{\| g_k^\infty | X^N \|}{\kappa Q_h}, \Delta_k \right]$.
- Instead we use the Monte Carlo method to approximate the probability value:
- We generate M random samples from the posterior distribution of $g_k^\infty \mid X^N$. Based on the samples, we check the event of 'sufficient reduction' and make a count on the failed cases: M_{fail} .
- The probability value is then approximated by

$$\Pr\left(E_k^N \mid X^N\right) \approx \frac{M_{\text{fail}}}{M}$$

- The approximation becomes accurate as M increases.

The VNSP scheme

At the k th iteration of the algorithm, start with $N = N_{k-1}$. Loop

- Evaluate N replications at each point y^j in the interpolation set $\mathcal{I}_k$, to construct the data matrix X^N. Note: data from previous iterations can be included.
- Construct the quadratic model Q_k^N and solve the subproblem for $x_k + s^{*,N}$.
- Update the value of κ_{Qh} by (24).
- Compute the Bayesian posterior distributions for the parameters of Q_k^∞ as described above.
- Validate the Monte Carlo estimate (32). If the criterion is satisfied, then stop with $N_k = N$; otherwise increase N, and repeat the loop.

The VNSP scheme

Two approximation steps (29) and (31) are employed in the computation. The following assumptions formally guarantee that risk $\Pr\left(E_k^N\right)$ is eventually approximated by the Monte Carlo fraction value M_{fail}/M.

Assumption 4

The difference between the risk $\Pr\left(E_k^N\right)$ and the Monte Carlo estimation value is bounded by $\frac{\alpha_k}{2}$

$$\left|\Pr\left(E_k^N\right) - \frac{M_{\text{fail}}}{M}\right| \leq \frac{\alpha_k}{2}$$

Under this assumption and the criterion (32), it implies

$$\left|\Pr\left(E_k^N\right)\right| \leq \left|\Pr\left(E_k^N\right) - \frac{M_{fail}}{M}\right| + \left|\frac{M_{fail}}{M}\right| \leq \frac{\alpha_k}{2} + \frac{\alpha_k}{2} = \alpha_k,$$

which guarantees the accuracy of the ‘sufficient reduction’ criterion (28).

The VNSP scheme

Assumption 5

The sequence of significance level values $\{\alpha_k\}$ satisfy the property:

$$\sum_{k=1}^{\infty} \alpha_k < \infty$$

This allows the use of the Borel-Cantelli Lemma in probability theory.

Lemma 5 (First Borel-Cantelli Lemma)

Let $\left\{E_k^N\right\}$ be a sequence of events, and the sum of the probabilities of E_k^N is finite, then the probability of infinitely many E_k^N occur is 0.

The Borel-Cantelli Lemma provides that the events E_k^N only happen finitely many times w.p.1. Therefore, if we define K as the first successful index after all failed instances, then (27) is satisfied w.p.1 for all iterations $k \geq K$.

Numerical results

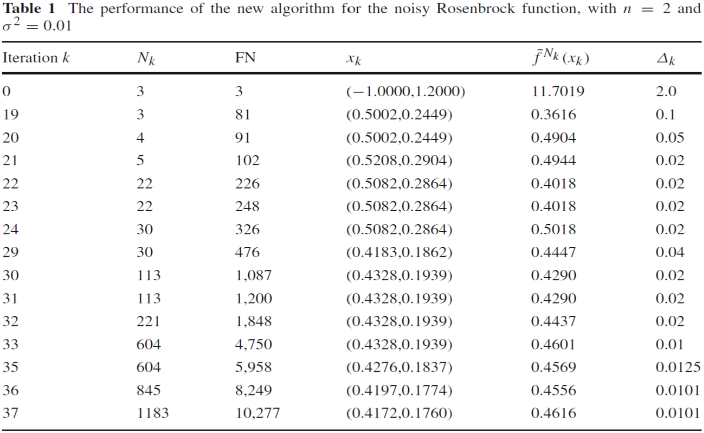

Table 1 The performance of the new algorithm for the noisy Rosenbrock function, with $n = 2$ and $\sigma^2 = 0.01$

Iteration k	N_k	FN	x_k	$\bar{f}^{N_k}(x_k)$	Δ_k
0	3	3	(−1.0000,1.2000)	11.7019	2.0
19	3	81	(0.5002,0.2449)	0.3616	0.1
20	4	91	(0.5002,0.2449)	0.4904	0.05
21	5	102	(0.5208,0.2904)	0.4944	0.02
22	22	226	(0.5082,0.2864)	0.4018	0.02
23	22	248	(0.5082,0.2864)	0.4018	0.02
24	30	326	(0.5082,0.2864)	0.5018	0.02
29	30	476	(0.4183,0.1862)	0.4447	0.04
30	113	1,087	(0.4328,0.1939)	0.4290	0.02
31	113	1,200	(0.4328,0.1939)	0.4290	0.02
32	221	1,848	(0.4328,0.1939)	0.4437	0.02
33	604	4,750	(0.4328,0.1939)	0.4601	0.01
35	604	5,958	(0.4276,0.1837)	0.4569	0.0125
36	845	8,249	(0.4197,0.1774)	0.4556	0.0101
37	1183	10,277	(0.4172,0.1760)	0.4616	0.0101

Numerical results

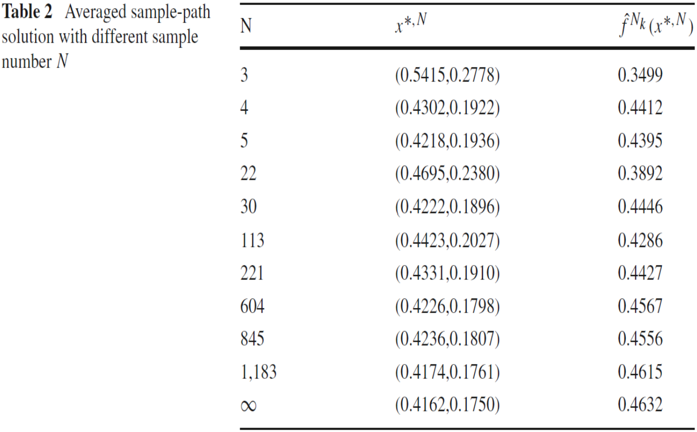

Table 2 Averaged sample-path solution with different sample number N

N	$x^{*,N}$	$\hat{f}^{N_k}(x^{*,N})$
3	(0.5415,0.2778)	0.3499
4	(0.4302,0.1922)	0.4412
5	(0.4218,0.1936)	0.4395
22	(0.4695,0.2380)	0.3892
30	(0.4222,0.1896)	0.4446
113	(0.4423,0.2027)	0.4286
221	(0.4331,0.1910)	0.4427
604	(0.4226,0.1798)	0.4567
845	(0.4236,0.1807)	0.4556
1,183	(0.4174,0.1761)	0.4615
∞	(0.4162,0.1750)	0.4632

Numerical results

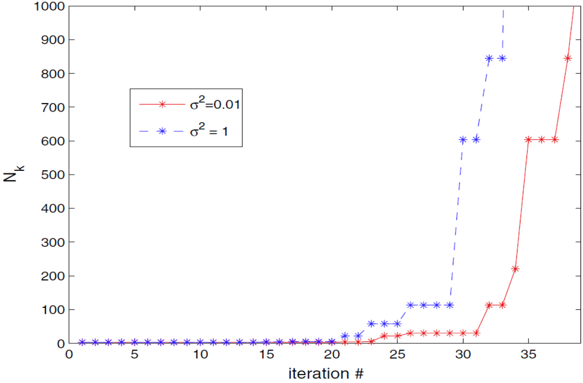

Fig. 5 Compare changes of N_k with different levels of noise

Numerical results

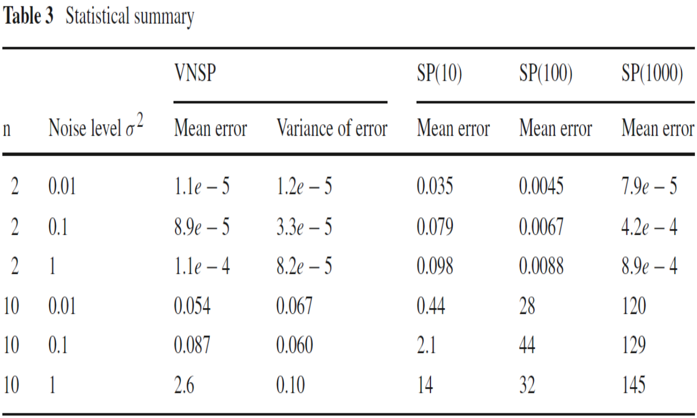

Table 3 Statistical summary

		VNSP		SP(10)	SP(100)	SP(1000)
n	Noise level σ^2	Mean error	Variance of error	Mean error	Mean error	Mean error
2	0.01	$1.1e-5$	$1.2e-5$	0.035	0.0045	$7.9e-5$
2	0.1	$8.9e-5$	$3.3e-5$	0.079	0.0067	$4.2e-4$
2	1	$1.1e-4$	$8.2e-5$	0.098	0.0088	$8.9e-4$
10	0.01	0.054	0.067	0.44	28	120
10	0.1	0.087	0.060	2.1	44	129
10	1	2.6	0.10	14	32	145

Thanks!