Bayesian Elegance in Resolving Semiconductor Manufacturing Challenges

Zhengqi Gao

Department of EECS, MIT
zhengqi@mit.edu

Presented at Lam Research Webminar, Jan 25, 2024

Semiconductor Manufacturing

Problems: yield estimation, recipe optimization, process control, variation analysis,…

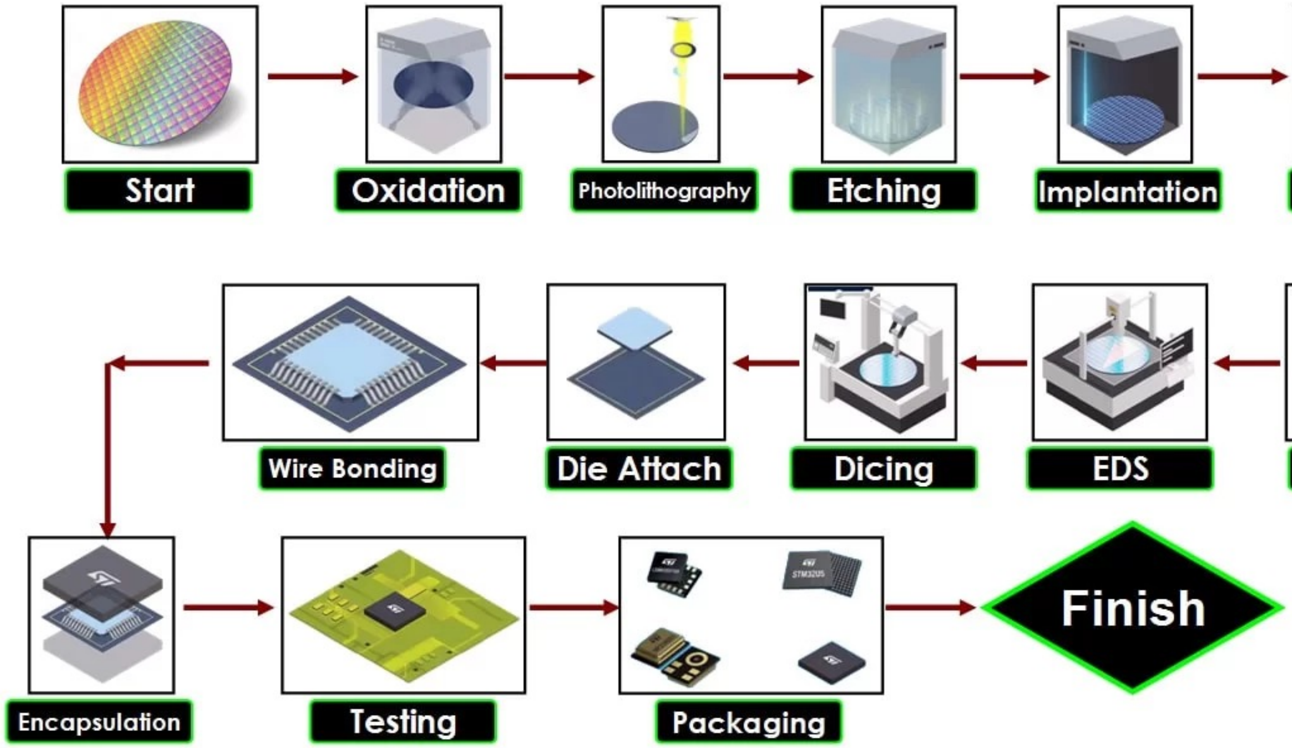


Figure Credit: https://shorturl.at/uBFH1

Three Characteristics of Semiconductor Manufacturing Problems

- **Fact 1: Costly function evaluation (usually black-box)**
  - Simulators (COMSOL, Coventor Products, SPICE, etc.,) usually run slowly
  - Lab measurement/testing takes even longer time

Three Characteristics of Semiconductor Manufacturing Problems

- **Fact 1: Costly function evaluation (usually black-box)**
  - Simulators (COMSOL, Coventor Products, SPICE, etc.,) usually run slowly
  - Lab measurement/testing takes even longer time
- **Fact 2: A restricted amount of data**
  - Algorithms asymptotic performances rely on the amount of data
  - Fewer data, less accurate (e.g., regression accuracy)

Three Characteristics of Semiconductor Manufacturing Problems

- **Fact 1: Costly function evaluation (usually black-box)**
  - Simulators (COMSOL, Coventor Products, SPICE, etc.,) usually run slowly
  - Lab measurement/testing takes even longer time
- **Fact 2: A restricted amount of data**
  - Algorithms asymptotic performances rely on the amount of data
  - Fewer data, less accurate (e.g., regression accuracy)
- **Fact 3: Intricate correlations among various scenarios**
  - Early-stage and late-stage correlations (e.g., front-end and back-end)
  - Multiple-corner correlations (e.g., {SS, TT, FF, SF, FS} process corners)

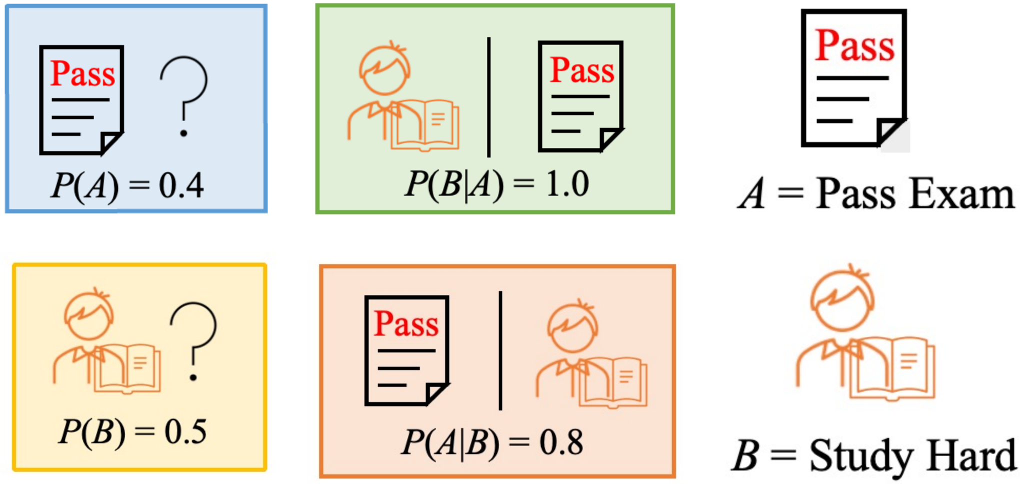
Three Characteristics of Semiconductor Manufacturing Problems

- Aim for a framework
  - Only assumes black-box function
  - Work with limited data
  - Can easily embed human knowledge

Bayesian Method: $$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \propto P(B|A)\,P(A)$$

Bayesian Method

$$\text{Bayesian Method:} \quad P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \propto P(B|A)\,P(A)$$



P(A): Prior probability

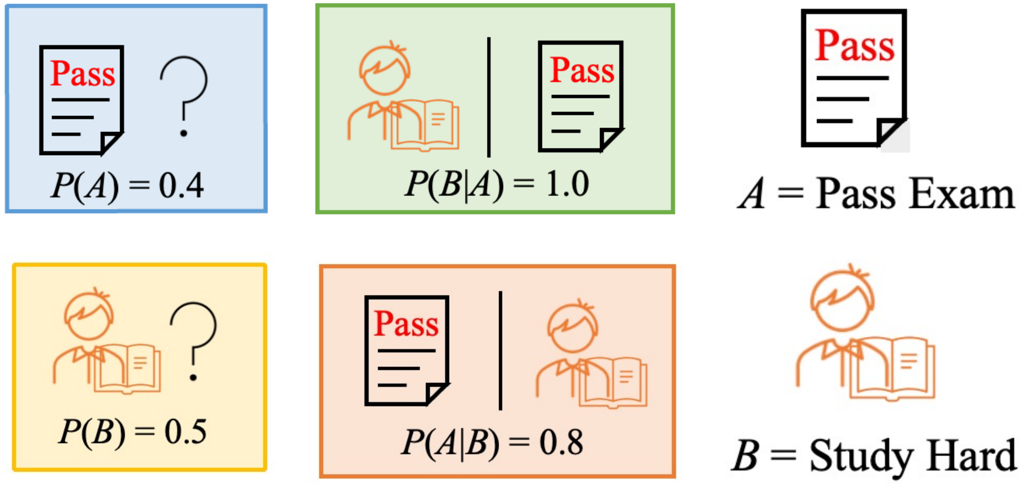
P(B|A): Likelihood of observing B given A

P(B): Model evidence or marginal likelihood

P(A|B): Posterior probability

Bayesian Method

$$\text{Bayesian Method:} \quad P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \propto P(B|A)\,P(A)$$



P(A): Prior probability

P(B|A): Likelihood of observing B given A

P(B): Model evidence or marginal likelihood

P(A|B): Posterior probability

Bayesian Method

Bayesian Method: $$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \propto P(B|A)\,P(A)$$

In essence, Bayes formula says: Posterior $\propto$ Likelihood x Prior

- P(B) is usually not explicitly needed (or sometimes difficult to evaluate)
  - Only P(A) and P(B|A) are needed, as P(A,B)= P(A)P(B|A) and next integration/summation marginalizes 'A' out.
  - If P(A) and P(B|A) are Gaussian, then P(A|B) is Gaussian [1].
  - For arbitrary P(A) and P(B|A), P(A|B) might not have a closed form.

Bayesian Method

Bayesian Method: $P(A|B) = \dfrac{P(B|A)\,P(A)}{P(B)} \propto P(B|A)\,P(A)$

Prior $P(x)$	Likelihood $P(y|x)$	Posterior $P(x|y)$
$\mathcal{N}(x|\mu_0,\sigma_0^2)$	$\mathcal{N}(y|x,\sigma^2)$	$\mathcal{N}(x|\frac{\sigma_0^2}{\sigma^2+\sigma_0^2}y+\frac{\sigma^2}{\sigma^2+\sigma_0^2}\mu,\frac{\sigma_0^2\sigma^2}{\sigma_0^2+\sigma^2})$
$\mathcal{N}(x|\mu_0,\sigma_0^2)$	$\mathcal{N}(y|g(x),\sigma^2)$	No closed form

- For a general case where posterior doesn't have a closed form
  - Variational Inference (e.g., mean-field) uses q_θ to approximate it.
  - Sampling method (e.g., MCMC) is used to draw samples.

Study Case 1: Early- and Late-Stage Performance Estimation

Formulation – Linear regression as an example [11]

Given $D=\{(x_i, y_i) \mid i=1, 2, \ldots, N\}$, find coefficients w such that $\boxed{y \approx \langle w, \phi(x) \rangle}$

Solve the optimization problem: $\min_{\mathbf{w}} \sum_{n=1}^{N} \left(\mathbf{w}^T \phi_n - y_n\right)^2$ linear regression

Least square: $\mathbf{w}_{lr} = \left(\mathbf{\Phi}^T \mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^T \mathbf{y}$

where $\mathbf{\Phi} \in \mathbb{R}^{N \times F}$ is the sample matrix; n-th row is $\phi_n = \phi(\mathbf{x}_n)$

Question: What if N<F?

Limited data regime, matrix not invertible; Use ridge regression

Study Case 1: Early- and Late-Stage Performance Estimation

Bayesian Approach

Assume prior on model coefficient: $\mathbf{w} \sim p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w}|\mathbf{m}_0, \frac{1}{\alpha}\mathbf{I}\right)$

Likelihood function: $p(\mathcal{D}|\mathbf{w}) = \prod_{n=1}^{N} p(y_n \mid \phi_n, \mathbf{w}) = \prod_{n=1}^{N} \mathcal{N}\left(y_n|\mathbf{w}^T\phi_n, \beta^{-1}\right)$

<=> Assume approximation error $y = \mathbf{w}^T\phi(\mathbf{x}) + \epsilon$ and $\epsilon \sim \mathcal{N}(\epsilon|0, \beta^{-1})$

Posterior has closed form: $p(\mathbf{w}|\mathcal{D}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N),$

where $\mathbf{m}_N = \mathbf{S}_N\left(\alpha\mathbf{m}_0 + \beta\mathbf{\Phi}^T\mathbf{y}\right)$ MAP Estimator

$$\mathbf{S}_N^{-1} = \alpha\mathbf{I} + \beta\mathbf{\Phi}^T\mathbf{\Phi}\ .$$

Study Case 1: Early- and Late-Stage Performance Estimation

MAP Estimator

$$\mathbf{m}_N = \left(\alpha \mathbf{I} + \beta \mathbf{\Phi}^T \mathbf{\Phi}\right)^{-1} \left(\alpha \mathbf{m}_0 + \beta \mathbf{\Phi}^T \mathbf{y}\right)$$
$$= \left(\mathbf{I} + \frac{\beta}{\alpha} \mathbf{\Phi}^T \mathbf{\Phi}\right)^{-1} \mathbf{m}_0 + \left(\frac{\alpha}{\beta} \mathbf{I} + \mathbf{\Phi}^T \mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^T \mathbf{y}$$

How to set m_0, alpha, beta?

Key: Set m_0 with early-stage info!

Step1: Construct a low-fidelity (early-stage) data set

Step2: Perform least square on it to obtain w; take it as m_0

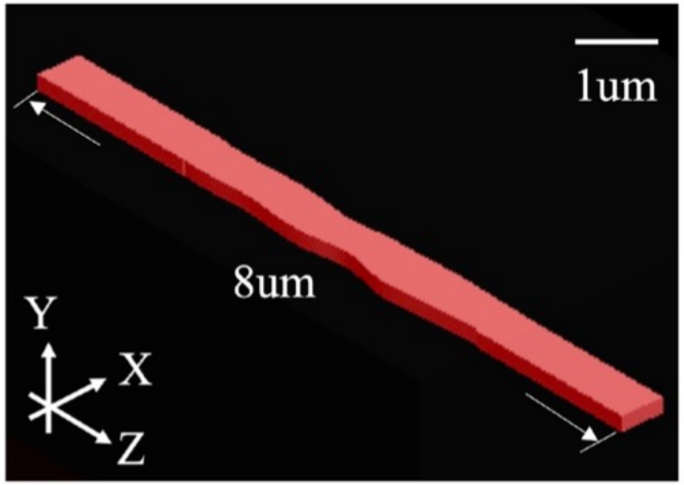
Step3: Combine with only few real data; use the MAP estimator

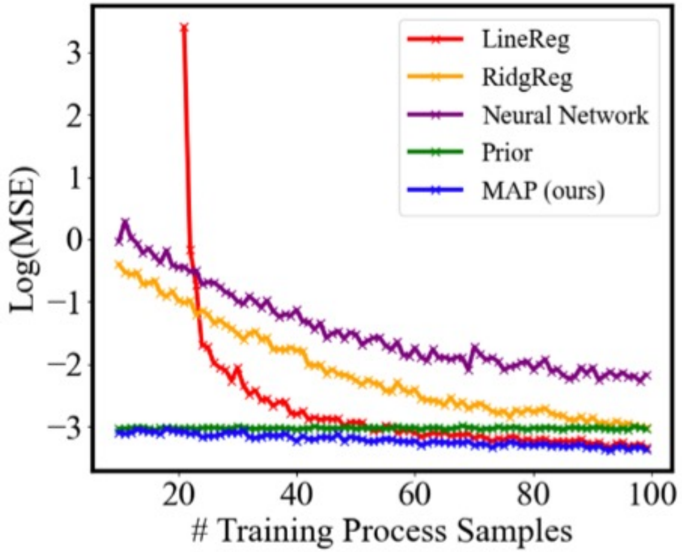
Note: low-fidelity data acquisition is cheap, can be a lot.

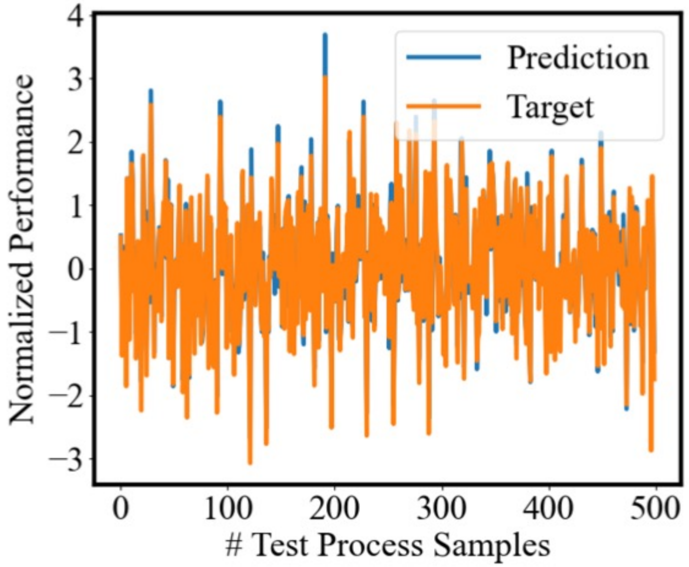
Study Case 1: Early- and Late-Stage Performance Estimation

Numerical Results

Modeling the phase of S parameter under variation





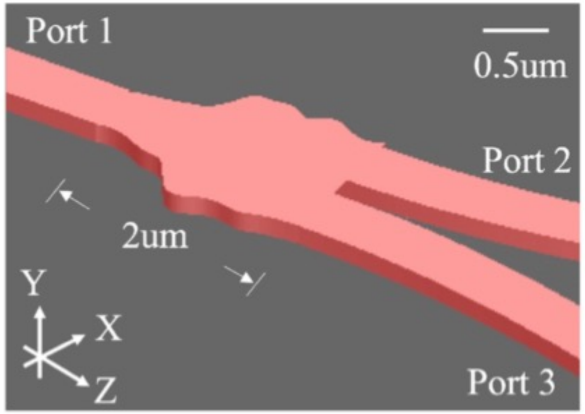


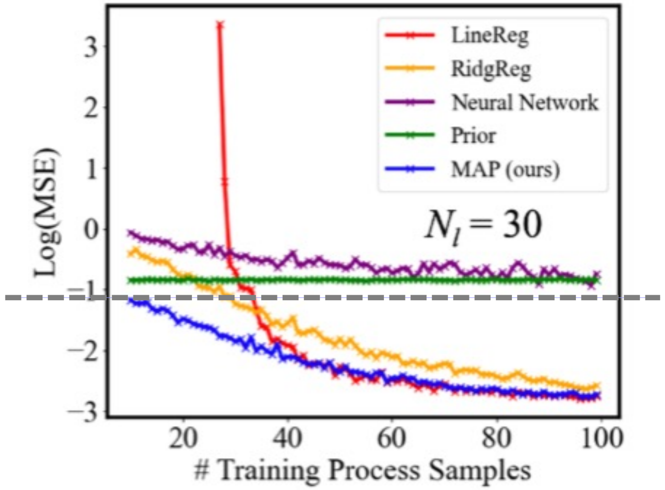
MAP and Prior use 50 low-fidelity samples

Study Case 1: Early- and Late-Stage Performance Estimation

Numerical Results

Modeling the magnitude of S parameter under variation





MAP with N_l = 30 & N= 10

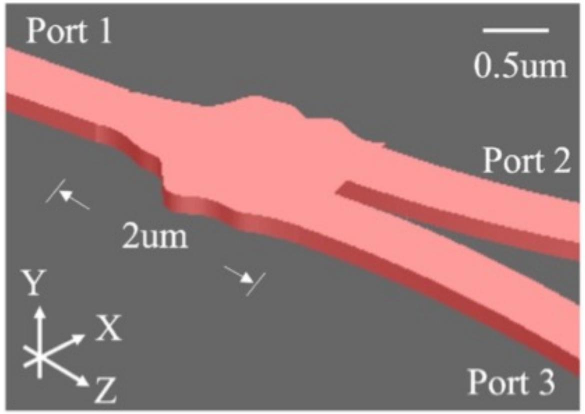
Linear regression with N = 30

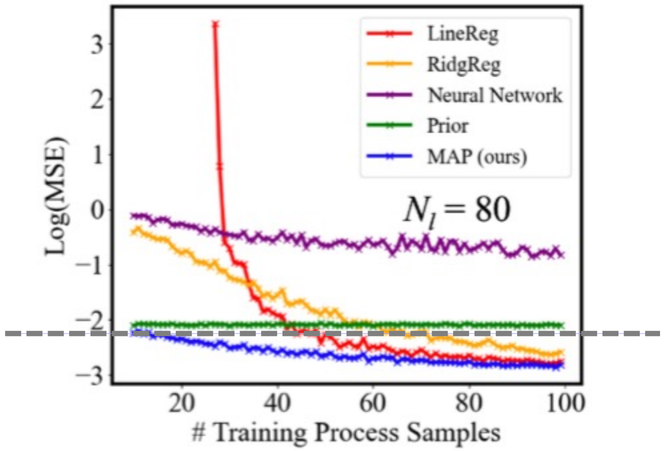
Both gives Log(MSE) = -1

Study Case 1: Early- and Late-Stage Performance Estimation

Numerical Results

Modeling the magnitude of S parameter under variation





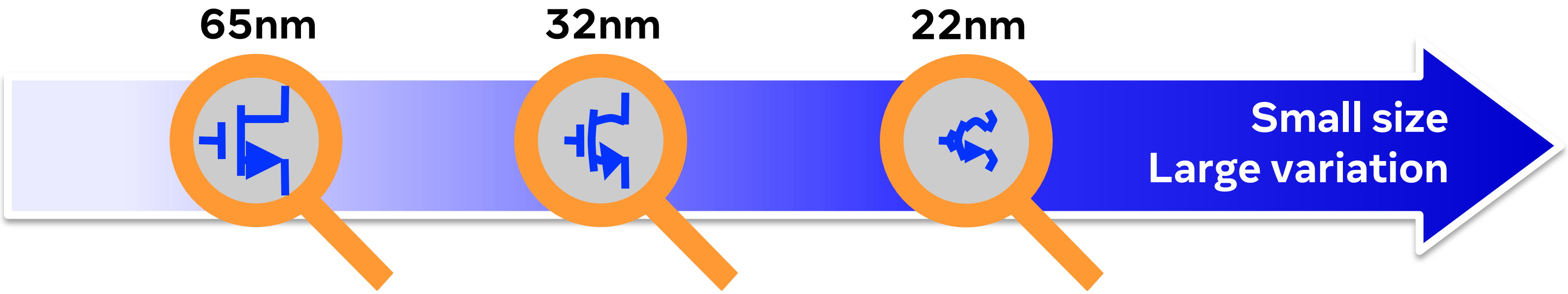
MAP with N_l = 80 & N= 10

Linear regression with N = 40

Both gives Log(MSE) = -2

Remarkable! Only 10 expensive data used in MAP for good accuracy

Study Case 2: Multiple-corner Yield Estimation



Circuit performances uncertainty under process variation

The problem of parametric yield estimation [2]:

Given a desired design, how likely does the fabricated design pass the Spec test?

Math formulation:

The desired design is denoted by **w***.

Process Design Kit (PDK) gives the random variation ε added in manufacturing.

What is the probability that **h=f(w* +** ε**)** locates in a certain “pass” region ?

Study Case 2: Multiple-corner Yield Estimation

Trivial approach --- Monte Carlo

Step 1: Generate N samples $\{\mathbf{w}^* + \varepsilon_1, \mathbf{w}^* + \varepsilon_2, \ldots, \mathbf{w}^* + \varepsilon_N\}$

Step 2: Simulate the corresponding $\{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_N\}$.

Binary variable
1 = 'inside', 0 = 'outside'

Step 3: Examine each sample locate in Ω or not $\{x_1, x_2, \ldots, x_N\}$

Step 3: Calculate the ratio how many of the N samples locate in Ω: $\beta = \frac{1}{N}\sum_{i=1}^{N} x_n$

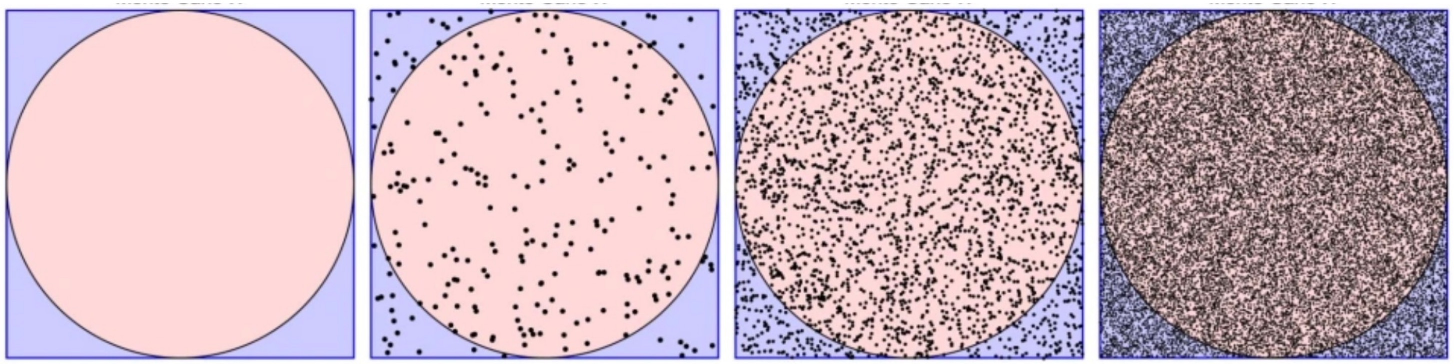


Figure credit: https://shorturl.at/dqsL2

Study Case 2: Multiple-corner Yield Estimation

Reinterpret MC

$$x = \begin{cases} 1 & \text{success} \\ 0 & \text{fail} \end{cases}$$ => $$p(x|\beta) = \begin{cases} \beta & x = 1 \\ 1 - \beta & x = 0 \end{cases}$$ => $$p(x|\beta) = \beta^x (1-\beta)^{1-x}$$

Conditional independence

$$p(D|\beta) = \prod_{n=1}^{N} p(x_n|\beta) = \prod_{n=1}^{N} \beta^{x_n} (1-\beta)^{1-x_n}$$

Maximum Likelihood Estimation (MLE):

$$\max_{\beta} \; p(D|\beta)$$ => $$\beta = \frac{1}{N} \sum_{i=1}^{N} x_n$$

Study Case 2: Multiple-corner Yield Estimation

When multiple corners?

What if we now want to estimate the yield at K corners (e.g., {TT,SS,FF,FS,SF})?

Certainly, we can apply MC independently at each corner: $\beta_k = \frac{1}{N_k} \sum_{n=1}^{N_k} x_{n,k}$

But, if yield at TT is 70%, the yield at other corners should be around 70% as well!

Embed it into the prior distribution!

Study Case 2: Multiple-corner Yield Estimation

Prior distribution: $p(\beta) = \mathcal{N}(\beta|\mu, \Sigma)$

Mu and Sigma are hyper-parameters control the shape of the prior

For example, all K elements in Mu equal 70%, Sigma is an Identity matrix

Likelihood function : $p(D|\beta) = \prod_{n=1}^{N} \prod_{k=1}^{K} \beta_k^{x_{n,k}} (1 - \beta_k)^{1-x_{n,k}}$

Posterior distribution: $p(\beta|D) \propto p(D|\beta)p(\beta)$

Maximum-a-posteriori (MAP) Estimation: $\max_{\beta} \; p(\beta|D)$ How to interpret MAP?

Study Case 2: Multiple-corner Yield Estimation

A few technical details

In essence, the proportional sign means: $p(\beta|D) = \frac{p(D|\beta)p(\beta)}{Z}$

Z is some normalization constant making the expression valid as a distribution

From Bayes theorem Z = p(D) and independent of Beta.

MAP estimation: Maximize the product of the prior and likelihood (Z can be ignored!)

$$\max_{\beta} p(\beta|D) = \min_{\beta} -\ln p(\beta|D) \quad \text{or} \quad \max_{\beta} p(D|\beta)p(\beta) = \min_{\beta} -\ln p(D|\beta) - \ln p(\beta)$$

Equivalently, in logarithm. Advantage: reduce numerical error (Overflow).

Study Case 2: Multiple-corner Yield Estimation

A few technical details

$$\min_{\beta} -\ln p(D|\beta) - \ln p(\beta)$$

where $p(\beta) = \mathcal{N}(\beta|\mu, \Sigma) = \frac{1}{(2\pi)^{K/2}} \frac{1}{|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(\beta-\mu)^T \Sigma^{-1} (\beta-\mu)\}$

$$p(D|\beta) = \prod_{n=1}^{N} \prod_{k=1}^{K} \beta_k^{x_{n,k}} (1-\beta_k)^{1-x_{n,k}}$$

- Observation 1: no closed form for posterior distribution
- Observation 2: Given Mu and Sigma, we can solve the MAP estimator Beta.

How to set hyper-parameters {Mu, Sigma}?

Study Case 2: Multiple-corner Yield Estimation

A simplified algorithm flow; More details refer to [2]

Step1: Given the observation $\{x_{n,k}\}$ and initialize hyper-parameters

All K elements in Mu initialize to the same, and Sigma to a diagonal matrix

Step2: Use IRLS (Newton method) to calculate MAP (i.e., minimize the -log)

An iteration algorithm using gradient and Hessian of the -log

Step3: Use Laplacian approximation to define an approximate posterior

Step4: Use Expectation-Maximization to update hyper-parameters

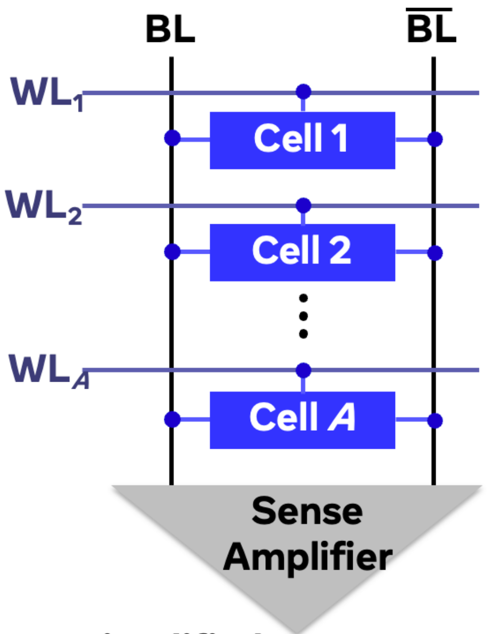
Step5: If convergence not reached, go to Step 2 with the new hyper-parameters

Study Case 2: Multiple-corner Yield Estimation

Numerical examples

Corner	Method	$N = 50$	$N = 100$	$N = 150$	$N = 200$	$N = 250$
TT	MC	4.22	3.15	2.73	2.46	1.75
	BI-BD	3.58	2.88	2.59	2.42	1.73
SS	MC	4.22	3.59	3.18	2.73	2.16
	BI-BD	3.87	3.40	3.08	2.57	2.04
FF	MC	4.73	3.79	3.14	2.28	2.06
	BI-BD	4.51	3.58	2.94	2.04	1.88
FS	MC	7.39	6.01	4.38	3.36	2.57
	BI-BD	5.13	4.09	3.10	2.39	1.85
SF	MC	8.81	6.31	5.86	4.74	3.52
	BI-BD	5.18	4.05	3.88	3.22	2.27

Estimation Error for MC and BI-BD (proposed)



Simplified SRAM array with A cells

Setting: 65nm PDK, five process corners, simulate in Hspice, error from 30 repeated runs

Baseline: independently run MC at each corner (ignore correlations!)

Study Case 2: Multiple-corner Yield Estimation

Some extensions

- Recall we deliberately introduce a Gaussian as prior
  - Not necessarily, this is at our choice.
  - In fact, Gaussian prior + Bernoulli likelihood -> no closed-form posterior
  - How about a prior resulting closed-form posterior? Easier calculation?
  - Indeed, we can do so with the concept of conjugate prior. [3]

Study Case 2: Multiple-corner Yield Estimation

Some extensions

- What if yield is very close to 100%?
  - In literature, usually referred to as rare failure rate estimation [4,5,6,7]
  - Practical usage/example: SRAM cell [4]
  - Challenge: if failure rate = 1e-6, we need (roughly) at least 1e7 MC samples!
  - Problem: Estimate rare failure rate with as few samples as possible
  - Metric: #samples used && logarithm prediction error
  - Past: Multiple-corner failure rate [6]; Recent: single-corner with NF [5]

Where else can Bayes methods be applied?

Minimum Testing [9]

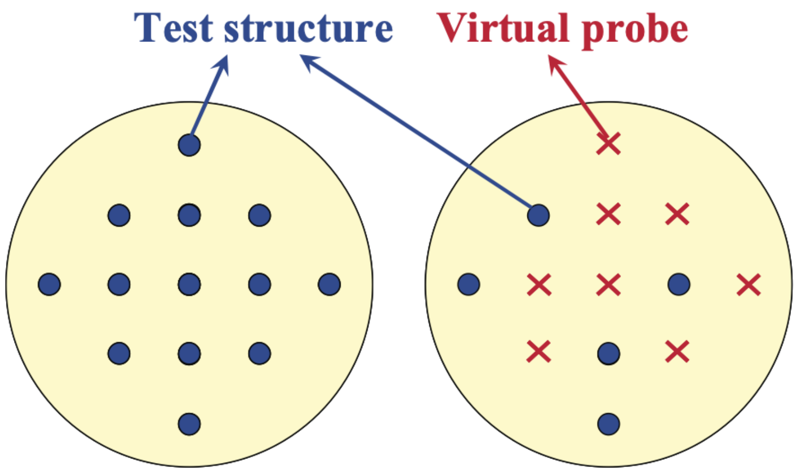


Figure credit: https://shorturl.at/ilGKQ

- Left: Test every location to characterize spatial variation.
- Right: Only few are tested, and others are inferred.
- Prior is introduced to make the system solvable. [?]

Where else can Bayes methods be applied?

Parameter Extraction [10]

- Given limited I-V measurement, extract MOS parameters.
- Prior: Novel transistor relates to existing transistor

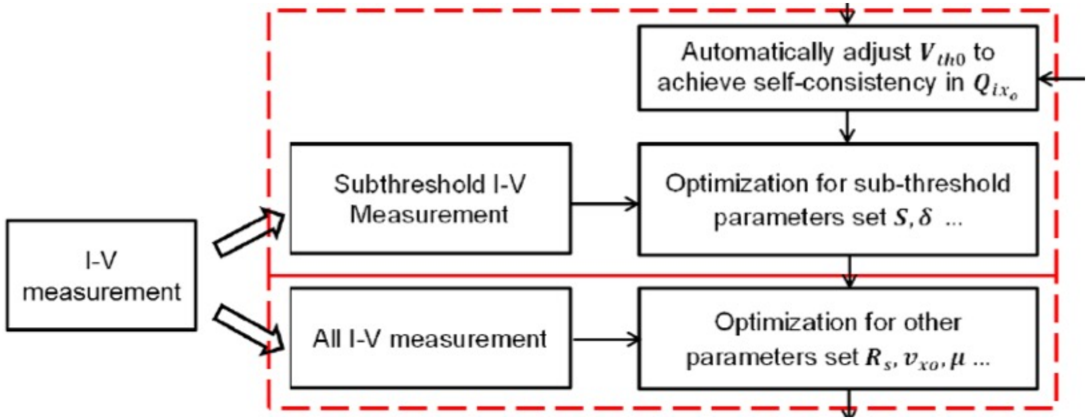
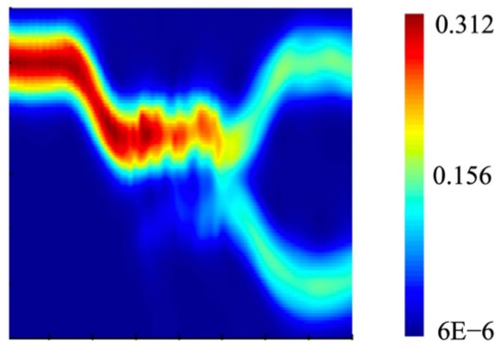


Figure Credit: https://shorturl.at/prNV9

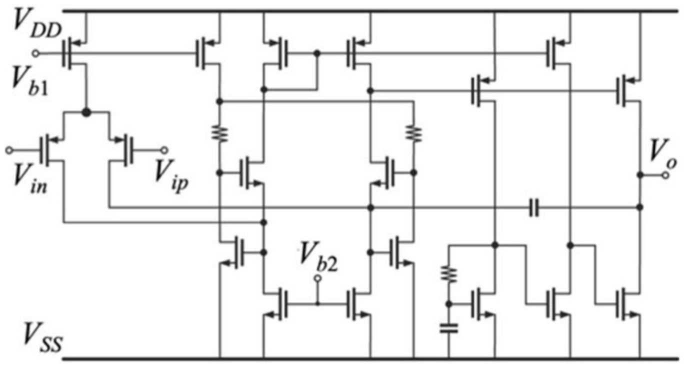
Where else can Bayes methods be applied?

Bayesian Optimization [13,14,15]

- Works with Black-box simulator
- Gaussian process regression (GPR) as surrogate model
- Work well in industrial examples (dim <= 40), e.g., analog sizing, photonic device design.
- Do notice that the definition of GPR requires a mean and a covariance (prior!)



Photonic Y-branch opt



Electronic Opamp

My Recent Focus: Likelihood Free Inference

Classical Parametric MLE: max $p(x|w)$

Where $p(x|w)$ has a parametric form

MAP: Further introduce a prior $p(w)$ and max $p(w|x)$

What is Likelihood Free Inference?

We can only sample an x from $p(x|w)$, but cannot evaluate the distribution value

"simulator-based likelihood"

A lot of approaches, e.g., approximate Bayesian computation (ABC)

Takeaway

- Bayesian methods naturally suit a lot of semiconductor manufacturing problems.
- Need to excavate the knowledge, correlations, and embed them with the prior.
  - Correct embedding improves sample efficiency!
- Usually, the problem is converted to inference a posterior distribution.
  - Posterior might not be analytical --- variational inference, or sampling.
  - EM method for a full Bayesian treatment (inference hyper-parameters)

References

1. C. Bishop, 'Pattern Recognition and Machine Learning', page 93, 2006.
2. Z. Gao et al., 'Efficient Parametric Yield Estimation Over Multiple Process Corners via Bayesian Inference Based on Bernoulli Distribution,' IEEE TCAD, 2019.
3. J. Shi et al., 'Multi-Corner Parametric Yield Estimation via Bayesian Inference on Bernoulli Distribution with Conjugate Prior,' ISCAS, 2020.
4. R. Kanj et al., 'Mixture Importance Sampling and Its Application to The Analysis of SRAM Designs in The Presence of Rare Failure Events', DAC, 2006.
5. Z. Gao et al., 'Rare Event Probability Learning by Normalizing Flows,' Arxiv Preprint, 2023.
6. Z. Gao et al., 'Efficient Rare Failure Analysis over Multiple Corners via Correlated Bayesian Inference,' IEEE TCAD, 2019.
7. SK Au and JL Beck, 'Estimation of Small Failure Probabilities in High Dimensions by Subset Simulation,' Probabilistic Engineering Mechanics, 2001.
8. Z. Gao and D. S. Boning, 'A Review of Bayes Methods in Electronic Design Automation,' Arxiv Preprint, 2022.

References

9. X. Li et al., ‘Virtual Probe: A Statistically Optimal Framework for Minimum- Cost Silicon Characterization of Nanoscale Integrated Circuits,’ ICCAD, 2009.
10. L. Yu et al., ‘Compact Model Parameter Extraction Using Bayesian Inference, Incomplete New Measurements, and Optimal Bias Selection, ’ IEEE TCAD, 2015.
11. Z. Gao et al., ‘Few-Shot Bayesian Performance Modeling for Silicon Photonic Devices Under Process Variation,’ IEEE JLT, 2023.
12. Z. Gao et al., ‘Rare Event Probability Learning by Normalizing Flows,’ Arxiv Preprint, 2023.
13. P. Frazier, ‘A Tutorial on Bayesian Optimization,’ Arxiv Preprint, 2018.
14. Z. Gao et al., ‘Automatic Synthesis of Broadband Silicon Photonic Devices via Bayesian Optimization’, IEEE JLT, 2022.
15. Z. Gao et al., ‘Efficient performance trade-off modeling for analog circuit based on Bayesian neural network,’ ICCAD, 2019.

Disclaimer: This slide deck primarily referenced the presenter’s papers for convenience, not aiming to be comprehensive.